Emerging Named Entity Recognition on Retrieval Features in an Affective Computing Corpus

1st Christian Nawroth
FernUniversität in Hagen
Hagen, Germany
christian.nawroth@fernuni-hagen.de

2nd Felix Engel
FTK e.V.
Dortmund, Germany
fengel@ftk.de

3rd Paul Mc Kevitt
Ulster University
Derry, Northern Ireland
p.mckevitt@ulster.ac.uk

4th Matthias L. Hemmje
GLOBIT GmbH
Barsbüttel, Germany
matthias.hemmje@globit.com

***Abstract*—Affective Computing (AC) is a relatively new, dynamic and interdisciplinary research field. Numerous contributions from fields like computer science, psychology, cognitive science, sociology, physiology and medical science have been made. Consequently, it is difficult to track all recently published trends for early insight utilisation in practise or as basis for innovative research. Even if this fact holds true for many other research fields, AC in this respect is stimulating, due to its dynamic and interdisciplinary characteristics. However, *Emergent* Entities Recognition is a new concept introduced for early detection and prediction of developing professional terminology. Initial software developments have been completed and briefly analysed in general databases (e.g. MEDLINE). Here, we are interested in its evaluation for AC. In this respect, we have created and used a new Benchmark for Emergent Entities recognition specifially for the field of AC and show evaluation results in comparison to state of the art trained named entity recognition models and to a generic corpus (MEDLINE).**

***Index Terms*—Emerging Named Entity Recognition, Machine Learning, Affective Computing**

I. Introduction

Recommendation Rationalisation (RecomRatio [1]), is a research project funded by Deutsche Forchungsmeinschaft (DFG) that aims to support health professionals during decision-making processes (e.g. for or against a certain therapy) through providing evidence based on clinical literature in specific domains, e.g. Affective Computing (AC). To find relevant trials, or medical publications that support or decline a therapy, specialised retrieval mechanisms on medical databases must be available. At its core, the retrieval process must understand the information need of the user, to calculate on this basis a set of applicable documents, sorted on their relevance. To express an information need users apply a professional terminology, formalized in taxonomies or vocabularies like Medical Subject Headings [2]. However, terminology changes over time and to identify new and emerging elements of a terminology is a challenging task. Because AC is a relatively new and dynamic field of research, we have tested our approach, the identification of emerging entities, in an AC document corpus. In general, new entities that arise in a domain are known as Emerging Entities [3], [4].

This work has been funded by the Deutsche Forschungsgemeinschaft (DFG) within the project Empfehlungsrationalisierung, Grant Number 643018, as part of the Priority Program "Robust Argumentation Machines (RATIO)" (SPP-1999).

II. Motivation

Understanding the emotional state of people that suffer from degenerative diseases and gathering insights on how people perceive emotion through mimics of others will heavily contribute to improved insights into disease causes and treatment. Hence, a plethora of recent AC research undertakings exist, that conduct studies based on effective instruments to capture, recognize and represent the emotional condition of affected persons to obtain a deep understanding of the modes of action involved. AC is an interdisciplinary research field (today it spans computer science, psychology, cognitive science, sociology, physiology and medical science) that is concerned with the development of computational systems, capable of detecting, responding to and simulating human emotional states. Health professionals in general and those working with AC solutions are using Information Retrieval Systems daily (e.g. PubMed[1]), e.g. when searching for medical literature in a decision-making process. Previous query log analysis shows that on PubMed more than 50% of the user queries contain terms that refer to entities which are represented by a domain-specific vocabulary, which in this case is MeSH. [5]. Hence, the medical domain, or a subdomain like AC, is predestined for Entity Retrieval (ER) [6] to fulfill medical user information needs on domain-specific entities like e.g. diseases or drugs. In clinical contexts Entity Retrieval faces a major challenge as, like numerous other domains, the medical domain also faces the global trends of Information Explosion [7] and Information Overload [8]. For example, from 1980 the number of citations added to PubMed per year grew from 279.692 citations added in 1980 to 1.178.360 citations in 2016, which means that the yearly growth rate increased by a factor of 3.6 within 35 years [9]. Besides a growth in medical literature there is also a growth in medical vocabularies like MeSH, which grew by 12,226 entries within 10 years from 2007 to 2016 (on "descriptor" level). Typically, each of these new entries is a name for a new medical entity or represents at least a new name for an existing entity. Thus, for Entity

[1]https://www.ncbi.nlm.nih.gov/pubmed

978-1-7281-1867-3/19/$31.00 ©2019 IEEE

Retrieval in the medical domain the identification of new entities and their names (as a textual representation), that arise through research and scientific discourse, represents an ever-increasing challenge within a fast growing document corpus. Hoffart et al. [3] use an approach built on knowledge bases to define emerging entities (EE) as entities that have been previously out-of-knowledge-base (OOKB). Brambilla et al. [4] define emerging entities as entities which are not included in a knowledge graph of a domain but are present in social media. Derczynski et al. [10] define the task of emerging recognition in a generic setup and report a max. F_1[2] of 0.420. Our approach for addressing emerging entities which has been introduced in [12] is different to the EE definitions shown before as it focuses on the textual representation (the name) of an entity instead of a knowledge object and hence we refer to it as emerging Named Entity (eNE). It is based on the generic idea of Named Entities (NE), e.g. for persons, locations, organizations [13] and the task of Named Entity Recognition (NER) [14]. Based on statistical observations we have chosen a temporal definition of eNER [12]:

Definition 1 (Emerging Named Entity (eNE)): A term, that is in use in domain-specific literature since the time t_{USE} and which is afterward acknowledged as a Named Entity by a 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 $[t_{USE}, t_{ACK}]$ (see Figure 1).

For our AC research we extended the definition as follows:

Definition 2 (AC Emerging Named Entity (AC-eNE)): An eNE in the domain of Affective Computing is a AC-eNE.

Based on definitions 1 and 2 we define emerging Entity Chunks to model eNEs that are used in actual texts:

Definition 3 (Emerging Named Entity Chunk (eNP)): A noun phrase (NP) in a text that contains an emerging Named Entity (eNE) is an emerging noun chunk (eNP).

Definition 4 (AC - Emerging Named Entity Chunk (AC-eNP)): A noun phrase (NP) in a text that contains an AC emerging Named Entity (AC-eNE) is an emerging noun chunk (AC-eNP).

As far as we are aware, there exist no formal definitions of AC-eNEs or AC-eNPs which follow a similar temporal approach. Here, we show that this particularly temporal approach supports the detection of AC-eNEs and eNEs in indexed medical text corpora through Machine Learning (ML) methods. Initial statistical results shown in [12] lead to our hypothesis (H1) that eNEs in general are actually used in a domain-specific literature corpus before the NE has been acknowledged by the domain-specific community and added to the domain vocabulary. In [12] we show additional statistics that support this hypothesis also for AC-eNEs.

[2]Precision, Recall and F_1 are commonly used to evaluate classification systems, using True Positives (TP), False Positives (FP) and False Negatives (FN) [11]: $Precision = \frac{TP}{TP+FP}, Recall = \frac{TP}{TP+FN}, F_1 = 2\frac{Precision \cdot Recall}{Precision+Recall}$

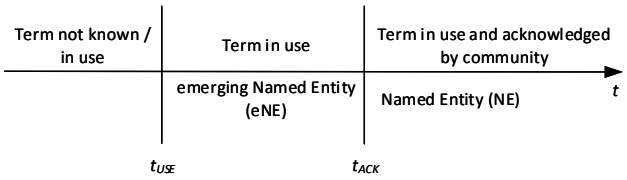

Fig. 1. Definition of emerging Named Entities (eNE).

A practical use use case within RecomRatio is intended to make AC-eNEs available for clinical argumentation processes, e.g. by visualizing them in trial literature evidence or using AC-eNEs as a ranking / filtering criterion for arguments. Here we show how AC-eNEs and eNEs can be recognized in a domain-specific text corpus through a combination of Information Retrieval (IR) and ML methods to provide AC-eNEs for Entity Retrieval use cases.

III. State of the Art and Related Work

This State of the Art review covers selected publications from the fields of Information Retrieval, Entity Retrieval, Named Entity Recognition and their combination for IR / ER use cases. Section B focuses on recent ML techniques and their utilization for NER.

A. Named Entity Recognition, Information Retrieval and Entity Retrieval

For 30 years NER has been a well known sub-task of Natural Language Processing (NLP) [14]. One major task of NER is the recognition of unique names of persons, organizations and locations, which are also referred to as ENAMEX [13]. Traditional approaches for NER are based on local text features (e.g. Part of Speech, characters, upper and lower case, token nearby) and use regular expressions [14] or sequenced-based learning models such as Hidden Markov Models (HMM) [15] or Conditional Random Fields (CRF) [16], [17]. Combining IR and NER has been introduced as Named Entity Retrieval by Petcova and Croft [18] amongst others, who propose an IR approach based on proximity between the text of a document and the entities. A common use case for NER in IR is the task of entity linking and retrieval as introduced before, which aims at satisfying users' information needs by providing actual entities instead of documents that mention them [6]. To support ER, several approaches aim at detecting NE in queries, which is referred to as Named Entity Recognition in Query (NERQ) [19], [20]. In SIGIR's 2014 ERD challenge, multiple approaches for entity recognition and disambiguation have been presented that utilize external knowledge sources (e.g. Wikipedia, Freebase) [21]–[23]. In the clinical sector the MeSH on Demand tool [24] represents a practical implementation of a system that combines IR and NER for medical Entity Retrieval use cases.

B. Machine Learning (for Named Entity Recognition)

Recent ML techniques include Support Vector Machines (SVM) [25], Random Forests (RF) [26], Gradient Boosting

[27] Classification (GBC) and variants of Deep Neural Networks (DNNs) such as Recurrent Neural Networks (RNNs) [28], Long Short-Term Memory (LSTMs) [29] and Convolutional Neural Networks (CNNs) [30]. The traditional NER approaches shown before are based on supervised learning. They typically require domain knowledge and hand-crafted training corpora. In contrast, more recent NER approaches utilize DNNs including their variations: RNNs are used for a variety of language understanding tasks [31]. A recent hybrid approach combines LSTMs and CNNs [32] for NER and does not require hand-crafted training material. Another successful NER approach uses LSTMs and CRFs [33]. Besides deep-learning-based methods, SVMs are known for years as a robust ML technique for a variety of classification tasks such as text categorization [34] or in the medical domain [35]. Recently, unsupervised methods based on non-labeled training models as Bidirectional Encoder Representations from Transformers (BERT) [36] for general language processing tasks and BioBERT [37] for specific clinical language tasks have shown impressive performance.

C. Discussion

The NER methods above use local textual features and hence require an expert tagged training corpus, large volumes of text or use external knowledge sources. To the best of our knowledge, no appropriate training data for AC-eNEs and generic medical eNEs is available at present. Hence, following our definition of AC-eNEs our approach in contrast uses existing temporal features derived through retrieval from the underlying AC related text corpus and succeeds without expert annotated training material. Our approach follows Chang and Manning [38], who propose to complement statistical or learning-based methods with rule-based approaches especially where there is no appropriate training data available.

IV. General Approach

Our general approach for detecting AC-eNEs and eNEs in a text corpus combines methods from NLP, IR and ML as shown in Figure 2. In stage (1) all documents of the corpus are processed using state of the art NLP technology (e.g. Spacy [39]) to detect candidates for AC-eNEs / AC-eNPs. In stage (2) the AC-eNE candidates are passed to a search engine which has indexed the complete corpus. In stage (3) temporal features from the result set for this query are passed to an ML-based classifier (ML ClF) which has been trained on temporal features. Examples for temporal features are the max. and min. document years returned on a query by SOLR. In stage (4) the detected AC-eNEs are played back to the index for further utilization in the underlying ER use case, e.g. argumentation support in the AC domain. Stage (5+) is the optional use of AC-eNEs for the task of NERQ and not considered here. The experiments presented here cover the stages (1) to (3).

V. Experiments

Our experiments consist of three phases: In the Baseline NLP phase (1), we investigate to which extent rule-based and state of the art learning-based NER are capable of detecting AC-eNEs in medical texts. In the ML phase (2) of the experiments we analyse how AC-eNEs can be detected with ML methods in a domain-specific corpus based on non local features. In both phases our experiments and the evaluations are conducted from the perspective of 2012. By this method, we are able to have a view on AC-eNEs from the perspective of 2012 which have been meanwhile acknowledged and hence can be easily identified in a controlled domain-specific vocabulary today. In the combined phase (3) we investigate how AC-eNP candidates that have been extracted through Baseline NLP can be verified or falsified with the ML-based classification.

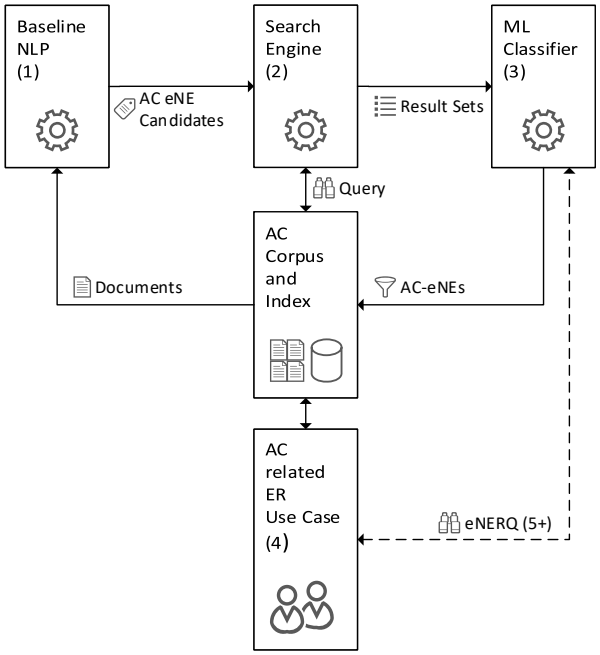

Fig. 2. General Architecture for AC-eNER.

A. Technical Setup

For stage (1) we implemented the Baseline NLP using Spacy [39], the search engine for stage (2) through Solr [40] with standard configuration and the ML classifiers through Python with Scikit-Learn [41].

B. Vocabulary

For our experiments we used MeSH. For each vocabulary entry MeSH provides the year in which an entry has been added to MeSH (e.g. in 2016, 402 new entries). According to our definition the year in which an entry has been added represents t_{ACK}. Hence, it is straightforward to split the vocabulary into a training set and a test set.

C. Corpora

For our experiments we use two different baseline corpora: A handcrafted corpus that is focused on AC (AC_CT) and

MEDLINE Baseline 2018 [42] for comparison. AC_CT is more specific and has a narrower scope during trials and in the field of AC, whilst MEDLINE is a large and generic corpus containing abstracts and titles of articles. We use this combination to ensure that our work is sufficient for AC-related and generic tasks. The ClinicalTrials (CT) register is a database from the National Library of Medicine that holds information about ∼316.200 studies. Studies are not specialised but rather cover a broad range of diseases. To obtain a better understanding on eNEs in AC, we have limited the body of available trials to those that are somehow related to Affective studies. Hence, we extracted a subset of the clinical trials register by filtering all studies related to: Affective Disorder — Symptoms — Psychoses — Personality — Disorder, Psychotic — Psychosis, Bipolar — Disorder Schizo — Disorder, Seasonal. Through this approach, we reduced the available number of studies to 5069 in the years 1999 to 2017 and created the unique AC_CT corpus. For AC_CT the mean number of documents added per year was 253.45 whilst the median was 314.5. On the other hand, the time frame for MEDLINE documents is limited to 50 years dating from 1969 to 2018 to avoid temporal artifacts resulting from historic documents. Within this time frame the overall document count is 24,910,297. The mean number of documents added per year is 508,373.4 and the median is 414,987. The SOLR index Size for AC_CT 10.5 MB and for MEDLINE is ∼25.5 GB. The comparison of size and time frame shows the challenge for AC_CT is that there is significantly less data available for training and evaluation as we show later here.

D. Statistics on eNEs in AC_CT and MEDLINE

Our statistical analysis focuses on questions on the extent eNEs are actually used in AC_CT and MEDLINE Baseline. The first question is, how long are eNEs in use in the respective corpus before T_{ACK}? To obtain a picture on this we queried the corpora using terms with $T_{ACK} \geq 2012$ and counted the documents with $T_{USE} < T_{ACK}$. From these documents we then calculated mean and median document age. For AC_CT and $T_{ACK} \geq 2012$ the mean age is 7.054 years (median: 7) Whilst for MEDLINE Baseline 2018 and $T_{ACK} \geq 2012$ we observe a mean age of 27.085 years (median: 28). The second question is: in how many documents is an eNE used before T_{ACK}? Our statistical analysis shows that for AC_CT and $T_{ACK} \geq 2012$ the mean number is 7.614 (median: 2) whilst for MEDLINE and $T_{ACK} \geq 2012$ the mean number is 951.4892 (median: 156.5). In both cases the difference between mean and median shows that there are few terms used in a larger number of documents compared to the majority of terms. Although the values for age and document count differ significantly between the two corpora - which is not surprising taking into account the different age ranges and overall document counts of the two corpora - it becomes clear that in both corpora eNEs are in use before T_{ACK} in a significant number of documents and hence should be potentially detectable.

E. Baseline NER Phase

Due to performance scaling reasons for the Baseline NER Phase we created two smaller subsets of AC_CT and MEDLINE that follow the design of the CONLL 2003 corpus for Named Entity Recognition [43]. For AC_CT the sizes for training, validation and evaluation set are 100, 25 and 25 which is ∼0.1 of the size of CONLL due to the overall smaller size of AC_CT, which will negatively affect NLP performance of training AC-eNER compared to the larger MEDLINE corpus. For MEDLINE each of the subsets consists of 1.000, 250 and 250 documents. All documents are taken randomly from the year 2012 and each contains at least one eNE from the perspective of 2012, meaning an entity with $2012 \leq T_{ACK} < 2018$. As today the entities with $2012 \leq T_{ACK} < 2018$ are known through MeSH's metadata it is straightforward to automatically create a gold standard eNE annotation on both subsets using pattern matching on the documents from 2012 with the recent version of the MeSH vocabulary.

F. ML Phase

In the ML phase for each entry of the MeSH vocabulary through SOLR we create an IR result set that contains temporal information (DOC_YEAR of all documents returned by a query). For the query term q, SOLR returns n documents, each with a DOC_YEAR leading to a result vector:

$$\vec{r_q} = \begin{pmatrix} DOC\_YEAR_0 \\ \vdots \\ DOC\_YEAR_{n-1} \end{pmatrix} \quad (1)$$

Based on $\vec{r_q}$ for each query we calculate a feature vector $\vec{f_q}$ (see section V-G below) that we use for training and testing the ML components (see Figure 3). To divide the vocabulary into training and test set we chose the $PIVOT\_YEAR$ $Y_p = 2012$ so that $T_{ACK} \leq T_p$. The training set is split again using the $TRAIN\_PIVOT\_YEAR$ Y_{tp} which is assigned dynamically during the experiments. In the training set, terms with $Y_{tp} \leq T_{ACK} < Y_p$ are labeled as AC-eNE / eNE and terms with $T_{ACK} < Y_{pt}$ as non AC-eNE / eNE. For testing, we use the temporal features of terms with $T_{USE} \leq Y_p$ and evaluate to which extent we are able to predict eNEs with $T_{ACK} > Y_p$. For ML we test several ML pipeline configurations. The generic pipeline consists of a Principal Component Analysis (PCA), an imbalanced sampler (SMOTE [44] or Remote Under Sampling or SMOTEENN [45]), a preprocessing scaler [46] and a classifier (SVM, RF or GBC).

G. Feature Engineering / Exploratory Data Analysis

For feature engineering we used exploratory data analysis (EDA) with R [47] (e.g. density or scatter plots), to identify relevant (temporal) features that are determining AC-eNEs / eNEs and may be used as input features for ML. SOLR result sets for a query of a candidate term are used as raw input for EDA. Figures 4 and 5 show density plots of the document years (DOC_YEAR) for both AC_CT and MEDLINE.

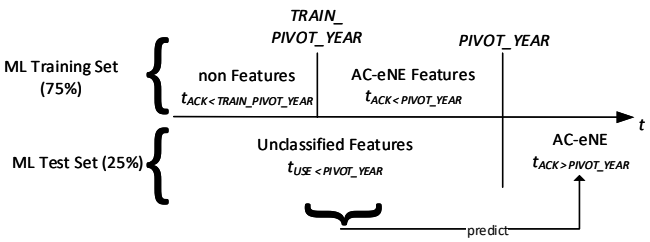

Fig. 3. Training and Test Scenario.

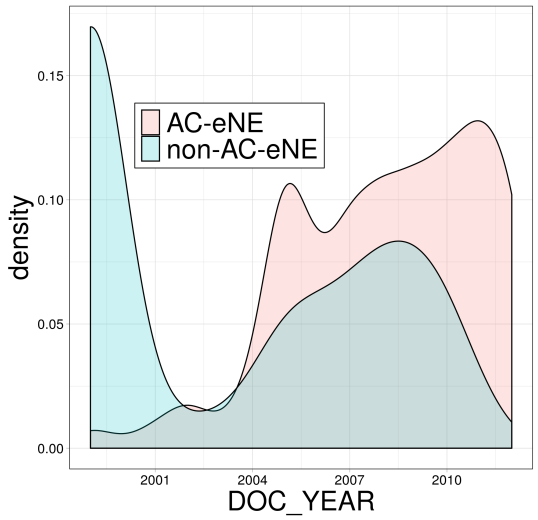

Fig. 4. Density plot of DOC_YEARs for AC_CT.

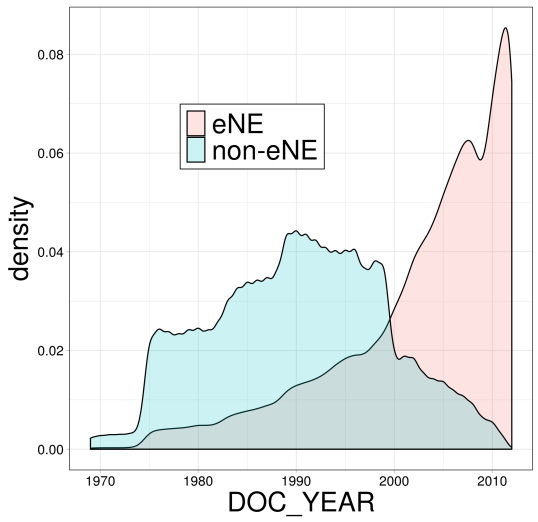

Fig. 5. Density plot of DOC_YEARs for MEDLINE.

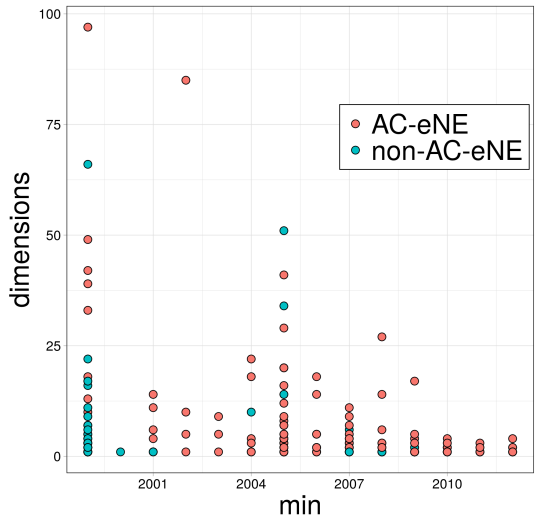

Fig. 6. Scatter Plot of MIN_YEAR and COUNT for AC_CT (all result sets).

In addition we use scatter plots to identify relevant combinations of features. Figures 6 and 7 show examples of a feature combination for MEDLINE and AC_CT. Although differing between the corpora both plots suggest that the classes are separable based on e.g. the features *dimensions* and *min* as shown in the scatter plots. With this method, we identified five features for constructing the feature vector: the number of documents (dimensions) per result vector, minimum, maximum, mean and median of DOC_YEAR per result vector, leading to the following feature vector for a query term q and a result vector $\vec{r_q}$:

$$\vec{f_q} = \begin{pmatrix} dimensions(\vec{r_q}) \\ min(\vec{r_q}) \\ max(\vec{r_q}) \\ mean(\vec{r_q}) \\ median(\vec{r_q}) \end{pmatrix} \quad (2)$$

VI. Evaluation

For evaluation of ML-based classification we use recall, precision and F_1-measure as described in [43]. We decline to use accuracy as an evaluation metric due to its shortcomings in imbalanced setups but use Recall Precision Curves and Area Under the (Recall Precision) Curve (AUC) [48].

A. Baseline NER - Rule-Based and Spacy-Based Approach

Baseline NER in our work is used for two purposes: On the one hand it provides eNE candidates from the corpus for further ML-processing, on the other hand it is used to compare the performance of our ML-learning approach on non local features. Following our general approach of only using already known AC-eNEs for NLP baseline training we annotated the training corpora from 2012 with AC-eNEs from the years 2007 - 2012 and trained the Spacy NER model with them. However, the evaluation is done on eNEs from the years 2013 - 2018

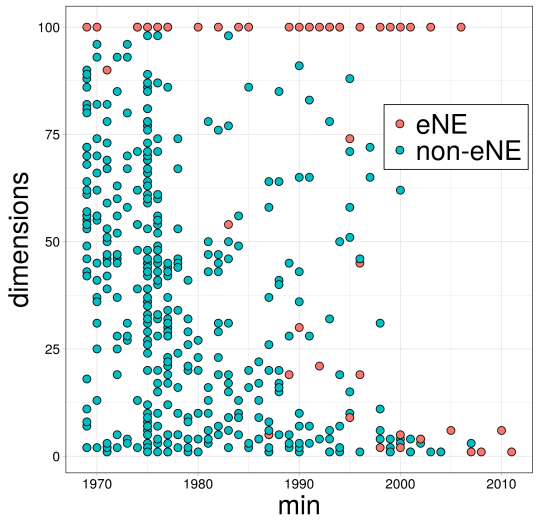

Fig. 7. Scatter Plot of MIN_YEAR and COUNT for MEDLINE (0.02 random sample).

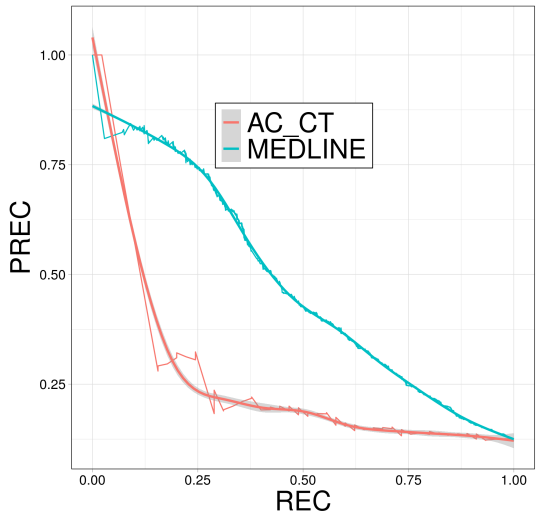

Fig. 8. Smoothed Precision / Recall curve for a selected MEDLINE (AUC 0.485) and AC_CT (AUC 0.26) model.

TABLE I
BASELINE NLP RESULTS

Corpus	Method	Recall	Precision	F_1
AC_CT	Spacy	0.000	0.000	0.000
AC_CT	Rule	1.000	0.036	0.069
MEDLINE	Spacy	0.021	0.477	0.040
MEDLINE	Rule	1.000	0.119	0.212

as the detection of future AC-eNEs is the main task. Besides Spacy's model for Baseline NER, we applied a naive rule-based approach to identify AC-eNPs. Table I shows the results.

B. ML Evaluation

In evaluating ML capability we investigated several combinations of pipeline components and T_{pt} which lead to 936 different combinations for each corpus. For AC_CT the max. F_1 is 0.292 and the max. AUC is 0.26. For AC_CT the best F_1 performance values are achieved with an RF classifier with SMOTE imbalance handling and $2000 \leq Y_{tp} \leq 2002$ whilst best AUC values are achieved using SVM. On the other hand, for MEDLINE we observed a better max. F_1 value of 0.467. For the initial ranks there is $Y_{tp} = 2000$ or $Y_{tp} = 2001$ and a GBC or RF classifier is used. The max. AUC of 0.485 is achieved for $Y_{tp} = 2000$ witch GCC. Figure 8 shows the precision recall curve for the ML models with the highest AUC on MEDLINE and AC_CT. This comparison shows again that for the emerging field of AC with a relatively small document corpus the task of AC-eNER remains challenging due to the smaller training data sets in comparison to MEDLINE.

C. Combining Baseline NER and ML

The final evaluation stage covers the combination of rule-based NER and ML as it will be utilized in most real-world use cases. For this evaluation stage we take the candidate AC-eNPs and candidate eNPs identified in the AC_CT and MEDLINE document subsets (see section V-C) through the rule-based baseline NLP approach and create the retrieval features through SOLR. The retrieval features are used as input for the ML classifiers that have previously been trained on known MeSH terms before. Evaluation results of combining baseline NER and ML for CT_AFF show a max. F_1 of 0.11 and a max. AUC of 0.13. In comparison, for MEDLINE max. F_1 is 0.22 with AUC of 0.17. However, besides the quantitative analysis of the performance we also assessed the qualitative outcome of the classification. Table II shows 10 examples of AC-eNEs / eNEs terms that were classified correctly in AC_CT and MEDLINE.

TABLE II
CORRECT CLASSIFIED TERMS

Term	T_{ACK}	MeSH ID	Corpus
patient health questionnaire	2017	D000073222	AC_CT
the amisulpride	2018	D000077582	AC_CT
apparent sadness	2018	D000078602	AC_CT
perfectionism	2016	D000072639	AC_CT
neurokinin-1 receptor antagonists	2013	D064729	MEDLINE
atazanavir sulfate	2015	D000069446	MEDLINE
bortezomib treatment	2015	D000069286	MEDLINE
single-balloon enteroscopy	2016	D000071087	MEDLINE
pessimism	2015	D000067657	overlap
pioglitazone	2018	D000077205	overlap

D. Discussion of Evaluation Results

Comparing EDA Figures 4 and 5 it becomes clear that the YEAR feature for AC_CT and MEDLINE is distinctive for AC-eNE and non AC-eNE terms, whilst it is more pronounced in MEDLINE for eNEs and non eNEs. The scatter plots (Figures 6 and 7) support the hypothesis that features based on document years can be used for ML training and testing.

Regarding baseline NLP for the Spacy model the performance lags typical values or State of the art training models, which supports our hypothesis that textual features are not sufficient for the detection of AC-eNEs / eNEs. However, the rule-based approaches fit our requirement to provide high recall for further ML-processing for AC_CT and MEDLINE. Coming to ML and assessing the RP curves (Figure 8), the model for AC_CT is less stable than MEDLINE as the decrease of precision for MEDLINE is more linear compared to AC_CT. Whilst the latter provides high precision values for only the cost of very low recall, the plotted MEDLINE model can keep precision values > 0.5 for recall values of $\sim$0.4.This makes it more convenient for real-world use cases in which a high recall is not essential for survival but where the users require an acceptable ratio of true and false positives from 1:1. A reason for the overall lower performance for AC_CT is that the time frame of the underlying corpus significantly influences performance of ML on temporal features. In an additional experiment we limited the time frame of MEDLINE to 1999 - 2017 (similar to AC_CT) and we reached a max. F_1 of 0.33 and an AUC of 0.38 which are significantly lower in comparison to the full MEDLINE corpus and approaching performance values of AC_CT. As MEDLINE 1999 - 2017 still comprises 15,092,159 documents we come to the conclusion that it's essentially the time frame and not the document count that influences ML performance. For recognition of eNPs, evaluation shows that a combination of NLP and ML in our approach is only slightly better compared to the rule-based baseline NER which shows the need for further investigation. Qualitative results show that AC_CT focuses on trials in the relatively narrow field of AC as the terms are related to psychiatry and psychology, whilst MEDLINE terms are from many medical disciplines as the texts cover the full range of medical language.

VII. Conclusion

The aim of our work here is to provide a classification model that is able to predict AC-eNEs based on already known AC-eNEs, so that no tagging of training material by AC experts is necessary, and that performs well on a specific corpus for AC. As a comparison we used the MEDLINE baseline. Whilst for MEDLINE evaluation showed that our approach is in part capable to meet F_1 values for as shown in [10], for AC_CT the performance lags behind MEDLINE values. We showed that the shorter time frame of an AC corpus in comparison to MEDLINE is the key reason for lower performance of ML-based on temporal features. A secondary reason is that the count of documents in the AC corpus is significantly lower in comparison to MEDLINE leading to smaller training sets for NLP and ML training. Hence, here we show that our approach of ML with non local retrieval features is in principle capable of early detection of eNEs in a corpus. For this task our approach outperforms NER training models on local textual features. However, for future real-world use performance in the field of AC (F_1) should be increased, e.g. through more extensive feature engineering (e.g. on term frequency features) and parameter optimization of the ML models regarding AC corpora and vocabularies as well as improving rule-based baseline NER for eNE chunk candidate detection for AC (e.g. using AC specific gazetteers).

References

[1] *Rationalizing Recommendations (RecomRatio): Project-Homepage*, Bielefeld, 2017. [Online]. Available: http://www.spp-ratio.de/de/projekte/ratiorec/.

[2] C. E. Lipscomb, "Medical subject headings (MeSH)," *Bulletin of the Medical Library Association*, vol. 88, no. 3, p. 265, 2000.

[3] J. Hoffart, Y. Altun, and G. Weikum, "Discovering emerging entities with ambiguous names," in *WWW 2014 - Proceedings of the 23rd International Conference on World Wide Web*, ser. WWW '14, New York, NY, USA: ACM, 2014, pp. 385–395, ISBN: 9781450327442. DOI: 10.1145/2566486.2568003.

[4] E. Della Valle, R. Volonterio, and F. X. A. Salazar, "Extracting emerging knowledge from social media," in *26th International World Wide Web Conference, WWW 2017*, ser. WWW '17, Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2017, pp. 795–804, ISBN: 9781450349130. DOI: 10.1145/3038912.3052697.

[5] J. R. Herskovic, L. Y. Tanaka, W. Hersh, and E. V. Bernstam, "A Day in the Life of PubMed: Analysis of a Typical Day's Query Log," *Journal of the American Medical Informatics Association*, vol. 14, no. 2, pp. 212–220, 2007, ISSN: 10675027. DOI: 10.1197/jamia.M2191.

[6] K. Balog, "Entity Retrieval," in *Encyclopedia of Database Systems*, New York, NY: Springer New York, 2017, pp. 1–6. DOI: 10.1007/978-1-4899-7993-3{_}80724-1.

[7] E. J. Huth, "The information explosion," *Bulletin of the New York Academy of Medicine*, vol. 65, no. 6, p. 647, 1989.

[8] D. Bawden and L. Robinson, "The dark side of information: Overload, anxiety and other paradoxes and pathologies," *Journal of Information Science*, vol. 35, no. 2, pp. 180–191, 2009, ISSN: 01655515. DOI: 10.1177/0165551508095781.

[9] U.S. National Library of Medicine, *Yearly Citation Totals from 2017 MEDLINE/PubMed Baseline*, 2017. [Online]. Available: https://www.nlm.nih.gov/bsd/licensee/2017_stats/2017_Totals.html.

[10] "Results of the WNUT2017 shared task on novel and emerging entity recognition," L. Derczynski, E. Nichols, M. van Erp, and N. Limsopatham, Eds., 2017.

[11] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. New York: Cambridge University Press, 2008, ISBN: 978-0-521-86571-5.

[12] Christian Nawroth, Felix C. Engel, Tobias Eljasik-Swoboda, and Matthias L. Hemmje, "Towards Enabling Emerging Named Entity Recognition as a Clinical Information and Argumentation Support," in *Proceedings of the 7th International Conference on Data Science, Technology and Applications, DATA 2018*, SciTePress, 2018, pp. 47–55, ISBN: 978-989-758-318-6. DOI: 10.5220/0006853200470055.

[13] N. Chinchor, *Message Understanding Conference (MUC) 7 LDC2001T02*, N. Chinchor, Ed., Philadelphia, 1997.

[14] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3–26, 2007, ISSN: 0378-4169.

[15] G. Zhou and J. Su, "Named entity recognition using an HMM-based chunk tagger," G. Zhou and J. Su, Eds., Association for Computational Linguistics, 2001, p. 473. DOI: 10.3115/1073083.1073163.

[16] J. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[17] Andrew McCallum and Wei Li, "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons," in *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, Edmonton, Canada: Association for Computational Linguistics, 2003, pp. 188–191. DOI: 10.3115/1119176.1119206.

[18] D. Petkova and W. B. Croft, "Proximity-based document representation for named entity retrieval," in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, Lisbon, Portugal: ACM, 2007, pp. 731–740, ISBN: 978-1-59593-803-9. DOI: 10.1145/1321440.1321542.

[19] J. Guo, G. Xu, X. Cheng, and H. Li, "Named entity recognition in query," in *Proceedings - 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009*, ser. SIGIR '09, New York, NY, USA: ACM, 2009, pp. 267–274, ISBN: 9781605584836. DOI: 10.1145/1571941.1571989.

[20] J. Du, Z. Zhang, J. Yan, Y. Cui, and Z. Chen, "Using search session context for Named Entity Recognition in Query," in *SIGIR 2010 Proceedings - 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '10, New York, NY, USA: ACM, 2010, pp. 765–766, ISBN: 9781605588964. DOI: 10.1145/1835449.1835605.

[21] F. Piccinno and P. Ferragina, "From Tagme to WAT: A new entity annotator," in *ERD 2014 - Proceedings of the 1st ACM International Workshop on Entity Recognition and Disambiguation, Co-located with SIGIR 2014*, ser. ERD '14, New York, NY, USA: ACM, 2014, pp. 55–61, ISBN: 9781450330237. DOI: 10.1145/2633211.2634350.

[22] M. Cornolti, P. Ferragina, M. Ciaramita, S. Rüd, and H. Schütze, "The SMAPH system for query entity recognition and disambiguation," in *ERD 2014 - Proceedings of the 1st ACM International Workshop on Entity Recognition and Disambiguation, Co-located with SIGIR 2014*, ser. ERD '14, New York, NY, USA: ACM, 2014, pp. 25–30, ISBN: 9781450330237. DOI: 10.1145/2633211.2634348.

[23] S. Cucerzan, "Name entities made obvious: The participation in the ERD 2014 evaluation," in *ERD 2014 - Proceedings of the 1st ACM International Workshop on Entity Recognition and Disambiguation, Co-located with SIGIR 2014*, ser. ERD '14, New York, NY, USA: ACM, 2014, pp. 95–100, ISBN: 9781450330237. DOI: 10.1145/2633211.2634360.

[24] Dan Cho, *MeSH on Demand Tool: An Easy Way to Identify Relevant MeSH Terms*, 2014. [Online]. Available: https://www.nlm.nih.gov/pubs/techbull/mj14/mj14%5C_mesh%5C_on%5C_demand.html.

[25] C. Cortes and V. Vapnik, "Support vector machine," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[26] R. Forests, "Random Forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2013.

[27] J. H. Friedman, "Greedy Function Approximation : A Gradient Boosting Machine 1 Function estimation 2 Numerical optimization in function space," *North*, vol. 1, no. 3, pp. 1–10, 1999, ISSN: 00905364. DOI: 10.2307/2699986. [Online]. Available: http://www.jstor.org/stable/2699986%20http://www.jstor.org/page/info/about/policies/terms.jsp.

[28] J. L. Elman, "Finding structure in time," *Cognitive Science*, vol. 14, no. 2, pp. 179–211, 1990, ISSN: 03640213. DOI: 10.1016/0364-0213(90)90002-E.

[29] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997, ISSN: 0899-7667.

[30] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2323, 1998, ISSN: 00189219. DOI: 10.1109/5.726791.

[31] K. Yao, G. Zweig, M. Y. Hwang, Y. Shi, and D. Yu, "Recurrent neural networks for language understanding," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, K. Yao, G. Zweig, M.-Y. Hwang, Y. Shi, and D. Yu, Eds., 2013, pp. 2524–2528.

[32] J. P. C. Chiu and E. Nichols, "Named entity recognition with bidirectional LSTM-CNNs," *arXiv preprint arXiv:1511.08308*, 2015.

[33] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,*

NAACL HLT 2016 - Proceedings of the Conference, pp. 260–270, 2016.

[34] T. Joachims, *Text categorization with support vector machines: Learning with many relevant features*. Springer, 1998, ISBN: 3540644172.

[35] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, no. 1-3, pp. 389–422, 2002.

[36] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," no. Mlm, 2018.

[37] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," pp. 1–8, 2019.

[38] A. X. Chang and C. D. Manning, *TOKENSREGEX: Defining cascaded regular expressions over tokens*, 2014. [Online]. Available: https://nlp.stanford.edu/pubs/tokensregex-tr-2014.pdf%0Ahttp://nlp.stanford.edu/pubs/tokensregex-tr-2014.pdf.

[39] M. Honnibal and I. Montani, *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*, 2017. [Online]. Available: https://github.com/explosion/spaCy/issues/1555.

[40] T. G. Potter and Timothy, *Solr in Action MEAP v1*, Online-Aus. Shelter Island, NY: Manning, 2012, pp. 1–96, ISBN: 9781617291029.

[41] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg, "Scikit-learn: Machine learning in Python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[42] "MEDLINE®/PubMed® Baseline Repository (MBR) Reference Material," U.S. National Library of Medicine, Bethesda, MD, Tech. Rep., 2017. [Online]. Available: https://mbr.nlm.nih.gov/.

[43] Sang, Erik F. Tjong Kim and F. de Meulder, "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, 2003, pp. 142–147. DOI: 10.3115/1119176.1119195.

[44] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority oversampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[45] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, p. 20, Jun. 2004, ISSN: 19310145. DOI: 10.1145/1007730.1007735.

[46] Scikit-Learn, *Preprocessing data*, 2019. [Online]. Available: https://scikit-learn.org/stable/modules/preprocessing.html#preprocessing-scaler.

[47] H. Grolemund and G. Wickham, *R for Data Science*. O'Reilly Media, Inc, 2016, ISBN: 9781491910382.

[48] J. Keilwagen, I. Grosse, and J. Grau, "Area under precision-recall curves for weighted and unweighted data," *PLoS ONE*, vol. 9, no. 3, Mar. 2014, ISSN: 19326203. DOI: 10.1371/journal.pone.0092209.