MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220148

Does Enrichment of Clinical Texts by Ontology Concepts Increases Classification Accuracy?

Kerstin Denecke[a]

[a] *Institute for Medical Informatics, Bern University of Applied Sciences, Bern, Switzerland*

Abstract

In the medical domain, multiple ontologies and terminology systems are available. However, existing classification and prediction algorithms in the clinical domain often ignore or insufficiently utilize semantic information as it is provided in those ontologies. To address this issue, we introduce a concept for augmenting embeddings, the input to deep neural networks, with semantic information retrieved from ontologies. To do this, words and phrases of sentences are mapped to concepts of a medical ontology aggregating synonyms in the same concept. A semantically enriched vector is generated and used for sentence classification. We study our approach on a sentence classification task using a real world dataset which comprises 640 sentences belonging to 22 categories. A deep neural network model is defined with an embedding layer followed by two LSTM layers and two dense layers. Our experiments show, classification accuracy without content enriched embeddings is for some categories higher than without enrichment. We conclude that semantic information from ontologies has potential to provide a useful enrichment of text. Future research will assess to what extent semantic relationships from the ontology can be used for enrichment.

Keywords:

Text classification, machine learning, semantic

Introduction

The availability of digital clinical data (unstructured and structured) brings tremendous opportunities and challenges to health care delivery. In particular, exploring the associations among the different pieces of information captured in clinical documents, but also in structured data is a fundamental problem to make appropriate clinical decisions in time. In this context, predictive systems attempt to support healthcare professionals in correctly interpreting the available data or in optimizing decisions by taking into account several aspects of patient's data from electronic health records (EHRs).

A wealth of information about the clinical history of a patient is locked in free-text clinical narratives, since writing text remains the most natural and expressive method to document clinical events [1]. Development of natural language processing (NLP) methods is essential to automatically transform clinical text into structured clinical data that can be directly processed using machine learning algorithms. NLP has been used in the clinical domain within diverse applications, including identification of biomedical concepts from radiology reports [2], problem lists [3] or discharge summaries [4]. There is evidence that data extracts using NLP improve prediction of clinical outcomes. Marafino et al. [5] applied machine learning (ML) methods and NLP to information routinely collected in EHRs, including lab results, vital signs, and free-text notes. They showed that including free text and NLP applied to it, significantly improves a prediction model for mortality in the intensive care unit, compared with approaches that use only the most abnormal vital signs and laboratory values.

Text embeddings, i.e. vector space representations (e.g. word2vec learned by continuous bag-of-words or skip-gram [6]) able to capture features beyond simple statistical properties, are increasingly used for text classification and prediction tasks. Even though those text embeddings are already successfully used, most approaches largely ignore the semantic information that is often explicitly associated with the input data.

In the medical domain, semantic knowledge is readily available in the form of ontologies (e.g. SNOMED CT, UMLS). They contain semantic concepts, categories and relations among them. Ignoring these semantics can have advantages or disadvantages in learning tasks. If the learned model is not restricted by concepts and relations specified on ontologies and semantic knowledge bases, potentially different representations of the input data can be discovered for a given task [7]. In contrast, ignoring the wealth of existing knowledge means that any useful attribute has to be relearned from scratch. This requires large amounts of training resources, which are often difficult to access in the medical domain due to data protection and data privacy issues.

The goal of this work is to introduce a strategy for using knowledge from medical ontologies to semantically enrich vectors of clinical text embeddings in German. The resulting enriched representation is fed into a deep neural network classifier for solving a classification task. The rationale behind this effort is an optimised classification.

Data set and use case

We consider the following use case related to identifying hints on health status changes in patients based on free text: The postoperative health status of an obese patient indicates the outcome of the surgical treatment. By each postoperative revisit, physicians need to go through the previous patient records to recall the patient status and to evaluate the postoperative risk of readmission. In order to support this process, we want to develop a method to extract indicators and to analyze weight changes using the clinical documentation, so that potential complications and risks of clinical readmission can be recognized in a timely way. This requires identifying sentences dealing with health status changes and categorizing them appropriately. This specific classification task is considered in this work.

The available data set comprises de-identified outpatient clinical notes from patients with obesity. It includes 33 postoperative patients with 305 outpatient notes in German collected over a time range of three years. We have defined an annotation schema that distinguishes eight relevant information items to monitor postoperative progress. While some of the aspects are rather clinical (e.g. complaints, or weight progression), others are related to the treatment progress, namely, eating habits, the liquid intake, activities, and multivitamin intake. For each information item between 2 and 5 categories can be distinguished. 604 sentences have been labeled with those categories. They form our dataset (Table 1).

Table 1: Information items and categories per item

Information item and categories	**Number of samples per category (n = 604)**
Multivitamin intake	Regular: 44 Irregular: 15
Drinking behavior: *concerns the intake of fluids, not alcohol.*	Sufficient: 12 Insufficient: 3
Eating behavior	Consistent: 2 Good: 13 Better: 9 Worse: 9 Bad: 7
Activities: *considers the physical activity of a patient.*	Sufficient: 53 Insufficient: 50
Psychological status: *concerns the mood of the patient.*	Stable: 18 Unstable: 8
Complaints: *represents the complaints concerning food consumption, or pain management.*	Food intake: 14 No: 50 Somatic: 53 Psychological: 12
Weight progression	Decreasing: 101 Increasing: 51 Constant: 49
General status: *describes the general progression of the patient status. It shows either the reflection of the objective patient status after the previous treatment or subjective feedback or feeling observed by the physician or expressed by the patient.* (good, bad):	Good: 29 Bad: 2

There are several challenges coming along with our dataset. The total number of available sentences is very small for applying machine learning or training deep neural networks. As can be seen in Table 1, the number of training examples per category is low and imbalanced. A reason for this is that in our dataset, same sentences are used in multiple documents and were therefore removed from the data set as duplicates. We are considering the classification task at the sentence level, which reduces the amount of information captured. The sentences comprise 11.3 words on average. A sentence can also belong to multiple classes. Because of the shortness of the sentences, we believe that content enrichment of the sentences or their vector representation is essential for successful classification.

Related Work

Embeddings are distributed vector representations that map text to dense fixed length vectors and capture prior knowledge that can be used for downstream tasks. Several word representation models have been introduced such as Word2Vec [6], ELMo [8] and BERT [9] trained and tested mainly on corpora from the general domain (e.g. Wikipedia). While Word2Vec learns context independent word representations, ELMo and BERT learn context dependent word representations.

Word distributions might differ between general corpora and biomedical corpora. Clinical sublanguage provides features that differ from general domain language and they complicate data processing [10,11]. Features include an extensive use of Greek and Latin-rooted terminology, complex syntactic embeddings and reduction, i.e. ellipsis of auxiliary and copula verbs, and complex compound words, often built on the fly. These characteristics occur in particular in German clinical texts, while some of them exist for clinical texts in other languages as well. They complicate syntactic analysis, in particular the identification of semantic relationships and the morphological analysis of single words, which is necessary for mapping text to concepts of a medical ontology.

Beyond linguistic peculiarities, there are other challenges for clinical NLP and prediction, such as data quality, domain complexity and temporality. They have to be considered when performing feature engineering to obtain effective and more robust features from those data, and build prediction or clustering models on top of them [12].

BioBERT is a pre-trained language representation model for the biomedical domain [13]. To create BioBERT representations, weights were initialized from BERT which is a bidirectional encoder representation from transformers pre-trained on general domain corpora. Then, BioBERT was pre-trained on corpora from the biomedical domain (PubMed abstracts, PMC full text articles). ClinicalBERT is an application of the BERT model to clinical corpora to address the challenges of clinical text [14]. The text originates from the Medical Information Mart for Intensive Care III (MIMIC-III) dataset [15]. Such domain-specific embeddings have been proven to provide better results on NLP tasks.

BioBERT and ClinicalBERT do not consider expert knowledge, they are only trained on biomedical or clinical text. UmlsBERT is a first attempt to integrate domain knowledge during pre-training process into a contextual embedding [16]. This is realized by connecting "words that have the same underlying 'concept' in UMLS, and by leveraging semantic group knowledge in UMLS to create clinically meaningful input embeddings" [17]. Wang et al. also used the UMLS for semantic enrichment of embeddings [18]. They incorporated semantic association patterns to retrieve associated concepts. The terms identified in this way were appended to the original text and used in this way in classification tasks. In the general domain, attempts to use WordNet for semantic augmentation of embeddings were introduced. Pittaras and Karkaletsis used WordNet hypernym relations to extract term-frequency concept information [19, 20].

A problem with these existing domain-specific embeddings is that they have been trained on English texts. The U.S. MIMIC-III collection is a clinical dataset in English [15] that has been of great value for training language models in organized challenges [21] and representation of biomedical word and sentence embeddings [22]. In German-speaking countries however, there is no freely available anonymous clinical text dataset that can be used for methodical investigations for clinical language processing, even though initiatives arose within the German-speaking community to change this [23]. To overcome this problem, even the generation of synthetic corpora has been suggested [24]. As a substitute of sharing corpora, [25] suggested sharing statistical models trained on access-protected corpora in order to support sentence splitting, tokenization and part-of-speech tagging. They trained such models on a German clinical

corpus that itself was unavailable due to unresolved legal and ethical issues [26,27]. Due to these reasons, not only are clinical data sets for German NLP missing, but also pre-trained language models for German medical language are rare as a BioBERT or ClinicalBERT. To address the problem of limited availability of German clinical text and pre-trained embeddings, we introduce in this work our concept for semantic enrichment of embeddings for German clinical text.

Method

In this section, we describe our content enrichment method that is applied to a given input text. It is realized in several steps.

Content extraction and enrichment

For content extraction, we exploit an NLP pipeline for mapping clinical text to concepts of the Wingert nomenclature [28]. The Wingert nomenclature (WNC) is a derivate of the German translation of SNOMED 2. In the same way as SNOMED 2 evolved to SNOMED CT and became an ontology, the WNC terminology is organized as an ontology. It is fully available (commercial license) in German and all relevant medical domains are covered. In its current version, the WNC contains about 110.000 concepts with about 250.000 descriptions. Such descriptions are typically synonyms and related terms but also translations. All concepts are connected via taxonomic ("is a"), partonomic ("is part of"), and semantic relations like "is contraindication of" [30]. Each concept belongs to one out of 11 semantic categories, e.g. diagnoses, morphologies, treatments, procedures, agents, microbiology, function, materials.

The NLP pipeline comprises stemming, parsing, resolution of abbreviations, disambiguation, and extensive spell-correction algorithms [28]. The output of NLP pipeline and terminology mapping is an xml based structure (conceptual graph) – representing the underlying (linguistic) syntax and semantics. It is worth noting that multiple words can be integrated in one concept (e.g. the phrases *inflammation of the appendix* or *the appendix is inflamed* is represented by one concept). Synonyms are mapped to the same concept.

Sentence representation and classification

The enriched content is used to generate word embeddings. The objective of this step is to generate a single, constant length, semantic vector. Out of the extracted concepts (e.g. the concept terms) together with the words of the original sentences, we generate a vector representation similar to the bag-of-word paradigm. We consider the frequencies of a concept terms or words over each document, resulting in semantic vectors of the form $v^{(i)} = \{v_1, v_2, \ldots, v_d\}$, where $v_j^{(i)}$ denotes the frequency of the j-th concept in the i-th document. After this stage, the generated semantic vector can be used for classification.

Experiments and results

Experimental setup

We run two types of experiments. In a first experiment, we aim to find out which feature combinations impact most on the classification accuracy. For this purpose, we use the enriched content to train linear classifier using a simple bag of word representation. FastText supervised, a multinomial logistic regression model, is used for sentence classification. For these varying feature combinations, we calculated Precision at 1 (P@1, see Table 2). First, we used fix hyperparameters: learning rate: 0.05, word ngrams: 5, dimension: 100, epoch 10. Second, for three feature sets: 1) sentences enriched with concept terms, 2) concept terms and 3) raw sentence, we further tested with different varying hyperparameters, but wth fix ngrams = 6. Results are shown in Table 3. In these experiments, we used the entire dataset of 604 sentences and 22 categories.

Second, we run experiments with a deep neural network as a model for sentence classification. Since we are missing a large corpus for learning embeddings for our clinical use case or clinical documents in German, we use pretrained GloVe embeddings from https://deepset.ai/german-word-embeddings. The embeddings have been trained on German Wikipedia data. We define the model with an embedding layer being the first layer, followed by two bidirectional LSTM layers. The bidirectional layers ensure that the model processes the sequence from start to end, as well as backwards. After that we define a dense layer with 6 units (relu activation) and a final output layer (sigmoid activation). We train the model with epoch = 25. We consider a two class classification problem: for each of our 22 categories, we consider the sentences labelled with a specific category as positive examples while all other sentences are considered negative examples for the specific category. This leads to a very imbalanced dataset for each category (e.g. for *multivitamin intake regular* we have 44 positive and 560 negative samples). We only test the model with categories where we have at least 14 sentences as positive examples to have sufficient training material. The number 14 was chosen in order not to lose too many categories. Results are shown in Table 4. Accuracy is determined for enriched and non enriched sentences.

Results

From Table 2 we can see that enrichment by concept terms increases the P@1 value compared to raw sentences. Using semantic categories alone or in combination with other features does not improve classification accuracy. Interestingly, P@1 is similar for the feature sets raw sentences and concept terms. This means that the concept terms represent well the relevant content of a sentence. Results in Table 3 even show, that concept terms as features can lead to better P@1 than for raw sentences. Best P@1 is achieved for concept terms only or together enriching the sentences.

Table 2: Results for varying feature sets (lr: 0.05, wordNgrams: 5, dimension:100, and epoch:10, n(train) = 461, n(test) = 179, 22 categories

Feature set	Words	P@1
Raw sentences	1294	0.311
Concept terms	705	0.311
Semantic cateogries of concepts	11	0.237
Concept terms + semantic categories	718	0.254
Concept terms + sentences	1999	**0.486**
Semantic categories + sentences	1305	0.249
Sentences + Concept terms + semantic categories	2010	0.260

Table 3: Results for varying hyperparamenters, fix ngrams:6, 22 categories

Hyper-parameters	P@1 Raw sentence	P@1 Concept terms	P@1 Concept terms + sentences
lr: 0.05, dim:100, epoch:10	0.299	0.260	0.492
lr: 0.05, dim:1500, epoch:10	0.294	0.299	0.520
lr: 1, dim:1500, epoch:25	0.497	**0.582**	**0.58**

Classification accuracy for two class classification problem is achieved between 73.8% -97.5% for concept terms and between 57.4% - 96.7% for raw sentences. From Table 4, we can see that for some categories, the enriched embedding leads to higher accuracy values than the non-enriched. But this does not hold true for all categories.

Table 4: P@1 values for LSTM, epoch = 25, 2 categories (with enriched sentence = concept terms and raw sentence)

Category	Sentence	Enriched sentence
Multivitamin intake sufficient	0.8443	**0.9050**
Multivitamin intake insufficient	**0.9508**	0.9262
Activities sufficient	**0.9015**	0.8197
Activities insufficient	0.8607	0.8607
Weight constant	0.8607	0.8689
Weight decreasing	0.5738	**0.7377**
Weight increasing	0.7869	**0.8443**
Complaints psychologic	0.9672	0.9754
Complaints somatic	0.8197	**0.8525**
Complaints food intake	0.9426	0.9590
Complaints no	0.8443	**0.9098**

Discussion and Conclusion

In this paper, we assessed how enrichment of text by mapping to concepts of a medical ontology impacts on the classification accuracy. We conclude that enriching sentences with concept terms has the potential to improve the accuracy. The improvement could be due to the increased number of features, but also to the enrichment and aggregation of synonyms to concepts. Additional experiments are needed for confirming this initial impression. A larger, more balanced dataset could be helpful. We used embeddings pretrained from Wikipedia texts since for German language – which is considered here – there is a lack of clinical text corpora and pre-trained embeddings [14]. A clinical word embedding could be useful for increasing the accuracy. Instead of using the GloVe representations for embedding, contextual representation such as BERT, ELMo could be exploited. Experiments on this will be conducted in the future.

With the experiments reported in this work, we still do not exploit the entire semantic power of the ontology. In future work, we assess until which level hypernyms should be included (e.g. only parents or even parents of parents). More specifically, we will follow the edges labeled with an is-a-relation or part-of-relation and include retrieved target concepts into the vector.

The WNC, the ontology we will exploit, does not only comprise is-a-relations, but also semantic relations such as "contraindication-of" or it links a disease concept with an indicated treatment of this disease. We could imagine to check for concepts resulting from the concept mapping of the input text whether there are semantic relations among these concepts and if so, include the relation type in the vector representation as additional semantic information. Such procedure could additionally enrich the vector with implicit knowledge. This would also distinguish our approach with the approach of Wang et al. [18] who used UMLS semantic relations and Pittaras and Karkaletsis who used WordNet hypernym relations [19,20].

We used the WNC since this ontology is already available in German. However, the approach as such would be transferrable to SNOMED CT (which is not yet available in German). Assessing the role of ontologies such as WNC or SNOMED CT for NLP tasks like classification or for even more comprehensive tasks like prediction could increase momentum to the translation of SNOMED CT into German and could help demonstrating the need for well-maintained ontologies.

Acknowledgements

We acknowledge funding of this project from the Hasler Foundation.

References

[1] K. Jensen et al. Analysis of free text in electronic health records for identification of cancer patient trajectories. Scientific reports, 2017, 7(1), p. 1-12.

[2] R.M.V. Flynn, T.M. Macdonald, N. Schembri, G.D. Murray, A.S.F. Doney. Automated data capture from free-text radiology reports to enhance accuracy of hospital inpatient stroke codes. Pharmacoepidemiol Drug Saf 2010;19:843–7.

[3] M. Kreuzthaler, B. Pfeifer, J. Antonio Vera Ramos, D. Kramer, V. Grogger, S. Bredenfeldt, et al. Problem List Clustering for Improved Patient-Based Disease Perception. 2018 IEEE International Conference on Healthcare Informatics Workshop (ICHI-W), ieeexplore.ieee.org; 2018, p. 88–9.

[4] Y. Deng, P. Dolog, J-M. Gass, K. Denecke. Obesity Entity Extraction from Real Outpatient Records: When Learning-Based Methods Meet Small Imbalanced Medical Data Sets. CBMS 2019. https://doi.org/10.1109/cbms.2019.00087.

[5] B.J. Marafino, M. Park, J.M. Davies, R. Thombley, H.S. Luft, D.C. Sing, et al. Validation of Prediction Models for Critical Care Outcomes Using Natural Language Processing of Electronic Health Record Data. JAMA Netw Open 2018;1:e185097.

[6] T. Mikolov et al. Distributed representations of words and phrases and their compositionality. arXiv preprint arXiv:1310.4546 (2013).

[7] Y. Bengio. Learning deep architectures for AI. Now Publishers Inc, 2009.

[8] M. E. Peters et al. Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, LA. 2018: 2227–2237.

[9] J. Devlin et al. Bert: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA. 2019: 4171–4186.

[10] E. Kara, T. Zeen, A. Gabryszak, K. Budde, D. Schmidt, R. Roller. A domain-adapted dependency parser for german clinical text. Proceedings of the 14th Conference on Natural Language Processing (KONVENS 2018). Vienna Austria, 2018.
[11] R. Roller, H. Uszkoreit, F. Xu, L. Seiffe, M. Mikhailov, O. Staeck et al. A fine-grained corpus annotation schema of German nephrology records. Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP), aclweb.org; 2016, p. 69–77.
[12] R. Miotto, F. Wang, S. Wang, X. Jiang, J.T. Dudley. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform 2018;19:1236–46.
[13] J. Lee et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36.4 (2020): 1234-1240.
[14] E. Alsentzer, J.R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann et al. Publicly Available Clinical BERT Embeddings. Proceedings of the 2nd Clinical Natural Language Processing Workshop, 2019:72-78
[15] A.E.W. Johnson, T.J. Pollard, L. Shen, L.-W.H. Lehman, M. Feng, M. Ghassemi, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3:160035.
[16] G. Michalopoulos et al. UmlsBERT: Clinical Domain Knowledge Augmentation of Contextual Embeddings Using the Unified Medical Language System Metathesaurus. arXiv preprint arXiv:2010.10391 (2020).
[17] M. König et al. Knowledge-based best of breed approach for automated detection of clinical events based on German free text digital hospital discharge letters. PloS one 14.11 (2019): e0224916.
[18] H. Wang, Y. Qiu, J. Jiang, J. Zhang, J.Yuan. Leveraging Word Embeddings and Semantic Enrichment for Automatic Clinical Evidence Grading. ICBCB 2018. doi:10.1145/3194480.3194492
[19] N. Pittaras, G. Giannakopoulos, G. Papadakis, V. Karkaletsis. Text classification with semantically enriched word embeddings. Natural Language Engineering.2020:1-35.
[20] N. Pittaras, V. Karkaletsis. A study of semantic augmentation of word embeddings for extractive summarization. Workshop MultiLing 2019: Summarization Across Languages, Genres and Sources, pp. 63-72.
[21] C.-C. Huang, Z. Lu. Community challenges in biomedical text mining over 10 years: success, failure and the future. Brief Bioinform 2016;17:132–44.
[22] Q. Chen, Y. Peng and Z. Lu. BioSentVec: creating sentence embeddings for biomedical texts, 2019 IEEE ICHI, Xi'an, China, 2019, pp. 1-5
[23] U. Hahn, F. Matthies, C. Lohr, M. Löffler M. 3000PA-Towards a National Reference Corpus of German Clinical Language. Stud Health Technol Inform 2018;247:26–30.
[24] L. Lohr, S. Buechel, U. Hahn. Sharing Copies of Synthetic Clinical Corpora without Physical Distribution—A Case Study to Get Around IPRs and Privacy Constraints Featuring the German JSYNCC Corpus. LREC 2018
[25] J. Hellrich, F. Matthies, E. Faessler, U. Hahn. Sharing models and tools for processing German clinical texts. Stud Health Technol Inform 2015;210:734–8.
[26] J. Wermter, U. Hahn. An Annotated German-Language Medical Text Corpus as Language Resource. LREC, 2004.
[27] E. Faessler, J. Hellrich, U. Hahn. Disclose Models, Hide the Data-How to Make Use of Confidential Corpora without Seeing Sensitive Raw Data. LREC, 2014, p. 4230–7.

Address for correspondence

Kerstin Denecke
Bern University of Applied Sciences
Quellgasse 21
2502 Biel / Switzerland
kerstin.denecke@bfh.ch