Healthcare Transformation with Informatics and Artificial Intelligence
J. Mantas et al. (Eds.)

doi:10.3233/SHTI230408

A Hybrid AI-Based Method for ICD Classification of Medical Documents

Daniel BRUNESS [a,1], Matthias BAY [b], Christian SCHULZE [a], Michael GUCKERT [a] and Mirjam MINOR [c]
[a] *KITE, Technische Hochschule Mittelhessen, Friedberg, Germany*
[b] *Synpulse8, Synpulse Deutschland GmbH, Düsseldorf, Germany*
[c] *Department of Business Informatics, Goethe University, Frankfurt, Germany*

Abstract. Automatic document classification is a common problem that has successfully been addressed with machine learning methods. However, these methods require extensive training data, which is not always readily available. Additionally, in privacy-sensitive settings, transfer and reuse of trained machine learning models is not an option because sensitive information could potentially be reconstructed from the model. Therefore, we propose a transfer learning method that uses ontologies to normalize the feature space of text classifiers to create a controlled vocabulary. This ensures that the trained models do not contain personal data, and can be widely reused without violating the GDPR. Furthermore, the ontologies can be enriched so that the classifiers can be transferred to contexts with different terminology without additional training. Applying classifiers trained on medical documents to medical texts written in colloquial language shows promising results and highlights the potential of the approach. The compliance with GDPR by design opens many further application domains for transfer learning based solutions.

Keywords. ICD coding, ontologies, transfer learning, hybrid AI Introduction

1. Introduction

Due to increasing cost pressure in the healthcare sector, digitization is becoming more and more important, especially in hospital administration. As part of medical treatment, hospitals are required to document all medically relevant information for each patient in patient records, including diagnoses, surgical procedures and other treatments. This information must be available in a standardized form to ensure comparability with health insurance companies and accounting. The ICD (International Classification of Diseases) and OPS (Operation and Procedure Code) coding systems [1] used in Germany ensure this standardization. Labelling documentation with these classification systems involves an enormous amount of manual work: Hospitals have to assign expensive medical staff such as doctors and medical controllers to this activity. The lack of suitably qualified staff either leads to high personnel costs or, in the case of unfilled positions, to delays

[1] Corresponding Author: Daniel Bruneß, KITE, Technische Hochschule Mittelhessen, Wilhelm-Leuschner-Straße 13, 61169 Friedberg, Germany; E-mail: daniel.bruness@kite.thm.de

and incorrect invoicing.

This paper deals with the automated labelling of patient records with codes from common medical classification systems. The aim is to drastically reduce the manual work involved in labelling patient records with medical codes. The new solution uses natural language processing and machine learning for classification in medical documentation. The initial learning phase is based on historical data of a hospital. Due to the GDPR required protection of patient data, the models trained in a hospital can not be transferred to other institutions. To avoid de novo training in every new environment, which is often not possible due to the lack of training data, transfer learning based on an ontology is used. The classifier input feature space is defined by an ontology. Our approach, as shown in Fig. 1, enriches existing healthcare ontologies for transfer learning: The new solution automatically adapts to the terminology of a new hospital. The classification is no longer based on the pure textual data of the hospital, but on the enriched ontology. This enables the transferability of the learned classification model, while considering the protection of patient data.

As proof of concept for our approach, we have chosen two health domains in each of which patient-related documents have to be classified. Both domains refer to the same concepts but use different vocabulary and therefore form a typical use case for our method.

2. Methods

Transfer learning typically acquires knowledge specific to a task based on given source data which is manifested in a model, e.g., weights of a Machine or Deep Learning model and uses this pre-trained model for further training on target data in the context of a new task resulting in a new model [2]. In contrast to this traditional transfer learning approach that transfers results of statistical learning, our method performs semantic transfer learning based on an ontology with which we achieve a unification of relevant features for source and target task (cf. Fig. 1). While input data may change when moving to the target domain D_T, the actual classifier model and the input feature space stay identical. This allows classifiers to be applied in both domains without additional training. In a straight-forward implementation (*transfer by copy*), we reuse the ontology in its original form to provide the features for the classifiers, and simply copy the classifiers to be applied in the target task. To achieve a better transfer of knowledge from the source domain to the target domain, we implement an ontology enrichment process. This process includes augmentation of the source domain ontology O_S with knowledge from the target domain. We consider this as *transfer by enrichment*. For technical details of the approach, we refer to the literature [3].

The **enrichment process** utilises the Medical Subject Headings (MeSH) as base ontology (denoted as O_{MeSH}). We enrich O_{MeSH} by adding technical terminology from the hospital domain, and then further by adding colloquialism as used in online forums and social media. First, O_{MeSH} is enriched with terminology defined in the Unified Medical Language System (UMLS). Concepts are matched by the MeSH Descriptor ID and the UMLS Concept Unique Identifier (CUI). Thus, the UMLS terms provide additional

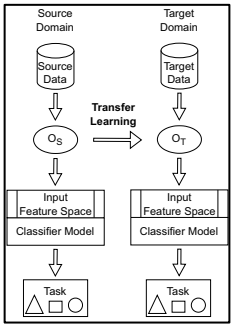

Figure 1. Ontology-based transfer learning process.

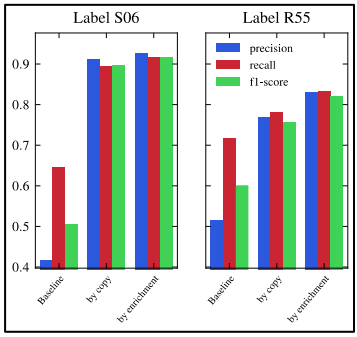

Figure 2. Precision, Recall and F_1 classifier performance in D_T for baseline and using different NCR models.

terminology for concepts in the base ontology. Second, the ontology is further enriched with colloquialism (result: O_{Com}). For this step, terminology from the German synonym database OpenThesaurus[2] and entity labels from Wikidata[3] are matched by using naive string comparison against concept names. Generally, more sophisticated ontology enrichment methods are also applicable [4].

The **feature extraction** from text is accomplished by a modified version of the Neural Concept Recognition (NCR) algorithm [5]. NCR makes use of hierarchical knowledge in ontologies and combines it with a word embedding model for text encoding. Here, we train NCR on the respective ontologies and word embeddings of a German fastText model [6]. The trained NCR model is used as an interface between text and ontological concepts. NCR maps unstructured text terminology them to the most suitable concepts in the ontology used for training. For our experiments, we trained NCR models based on the ontologies O_{MeSH} and O_{Com}.

The **classifier pipeline** is set up as follows. Instance features of a document are provided by NCR over the currently used ontology and are weighted using a TF-IDF vectoriser. The classifiers are trained in a One-vs-All fashion. For each ICD label, the classifier consists of an ensemble of three classifier algorithms, namely a Random Forest, a Logistic Regression algorithm and a Support Vector Machine. A Voting Classifier decides on the classification result using a soft voting procedure.

3. Results

Our method is evaluated in a real-world scenario. The source domain provides patient-related clinical reports from a hospital, written in technical German language, as **training data**. Further, the target domain contains of user-created content from an internet community[4] on health issues composed in everyday language. The underlying task for both domains (ref. Fig. 1) is a multi-class multi-label classification problem, in which medical texts are annotated with ICD codes. The training data consists of documents annotated with the labels *S06.0* (Concussion), *R55* (Syncope and collapse) and *None*

[2] https://www.openthesaurus.de
[3] https://www.wikidata.org
[4] https://www.onmeda.de

label. In detail, we obtained 2,483 unstructured text documents in total, 851 documents for label S06, 548 for label R55 and 1,084 documents without an assigned label (*None*) as negative examples. A held out test data set of 96 documents from the target domain was manually annotated by a domain expert.

We ran three **experiments**, in which the classifiers are trained in D_S and then copied to the target task in D_T. First, we created a simple baseline for our method by evaluating a simple, straightforward bag-of-words-based classification. For this, we used TF-IDF features from raw text as input features for the classifiers. Note, that in the baseline experiment no ontology was involved. Second, we trained the classifier with features extracted from documents in D_S by NCR trained on O_{MeSH}. The third classification experiment relied on features extracted with the NCR model trained on O_{Com}. The features in the test data set were extracted likewise for all experiments. The **results** of the experiments are depicted in Figure 2. The baseline has F_1-scores of 0.51 and 0.60 for the labels S06 and R55. Using O_{MeSH}, the S06-classifier shows a F_1-score of 0.90. The enrichment to O_{Com} results in $F_1 = 0.92$. The R55 classifier shows a F_1 increase of 0.04 points after the enrichment process.

4. Discussion

We showed that a naive switch from a domain D_S to D_T results in poor performance of baseline classifiers in D_T. As D_T has a remarkably sparsely occupied feature space this was an expected result. By applying ontology-based semantic transfer learning good performance of the baseline classifiers can be observed.

5. Conclusion

We show that the use of ontology-based transfer learning is beneficial for the classification of medical documents when moving a classifier to a new but similar domain. Our method is designed for situations in which lack of data or GDPR restrictions prevent the application of deep learning techniques. Remarkably, no training data in the target domain is needed. Using a restricted feature space of a given source domain ontology, we show that ontology enrichment and concept recognition methods mitigate the sparse data situation in the target domain.

References

[1] Stausberg J, Lang H, Obertacke U, Rauhut F. Research Paper: Classifications in Routine Use: Lessons from ICD-9 and ICPM in Surgical Practice. J Am Medical Informatics Assoc. 2001;8(1):92-100.
[2] Yang Q, Zhang Y, Dai W, Pan SJ. Transfer Learning. Cambridge: Cambridge University Press; 2020.
[3] Bruneß D, Bay M, Schulze C, Guckert M, Minor M. An Ontology-based transfer learning method improving classification of medical documents. In: 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE Comput. Soc; 2022. p. 407-12.
[4] Paulheim H. Knowledge graph refinement: A survey of approaches and evaluation methods. Semantic Web. 2017;8(3):489-508.
[5] Arbabi A, Adams DR, Fidler S, Brudno M, et al. Identifying clinical terms in medical text using ontology-guided machine learning. JMIR medical informatics. 2019;7(2):e12596.
[6] Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching Word Vectors with Subword Information. TransAssoc Comput Linguistics. 2017;5:135-46.