MLT-DFKI at CLEF eHealth 2019:
Multi-label Classification of ICD-10 Codes with BERT

Saadullah Amin, Günter Neumann, Katherine Dunfield, Anna Vechkaeva, Kathryn Chapman, and Morgan Wixted *
DFKI GmbH, Campus D3 2, 66123, Germany
{saadullah.amin,guenter.neumann,katherine.dunfield, anna.vechkaeva,kathryn_annette.chapman, morgan.wixted}@dfki.de
https://www.dfki.de

Abstract. With the adoption of electronic health record (EHR) systems, hospitals and clinical institutes have access to large amounts of heterogeneous patient data. Such data consists of structured (insurance details, billing data, lab results etc.) and unstructured (doctor notes, admission/discharge details, medication steps etc.) documents, of which, the latter is of great significance to apply language processing tools. In parallel, recent advancements in transfer learning for natural language processing (NLP) have made researchers push the state-of-the-art to new limits on many language understanding tasks. In this paper, we fine-tune pre-trained BERT (Bidirectional Encoder Representations from Transformers), a transfer learning language representation model, and its recent variant BioBERT, to the task of automatically assigning ICD-10 codes to non-technical summaries (NTSs) of animal experiments presented as Task 1 at CLEF eHealth 20191. The task documents are in German, where we achieve an F1 score of 76.68% with multilingual BERT. However, we also demonstrate that working with the translated English documents gives best results, achieving an F1 score of 82.04% and 82.90% with BERTBase and BioBERT respectively2.

Keywords: Semantic Indexing, Transfer Learning, Multi-label Classification, ICD-10 Codes

1 Introduction

EHR systems offer a rich source of data that can be utilized to improve health care systems by information extraction, representation learning and predictive modeling [34]. Among many other applications, one such task is the automatic assignment of International Statistical Classification of Diseases (ICD) codes

*On behalf of the PRECISE4Q consortium
1 https://clefhealth.imag.fr/?page_id=26
2 https://github.com/suamin/multilabel-classification-bert-icd10
to clinical notes, otherwise called semantic indexing of clinical documents. The problem is to learn a mapping from natural language free-texts to medical concepts such that, given a new document, the system can assign one or more codes to it. Approximating the mapping in this setting is called multi-label classification and can be one way to solve the problem, besides hierarchical classification, learning to rank and unsupervised methods.

In this study, we describe our work on CLEF eHealth 2019 Task 1, which is about multilingual information extraction from German non-technical summaries (NTSs) of animal experiments collected from AnimalTestInfo database to classify according to ICD-10 codes, German modification version 2016. The AnimalTestInfo database was developed in Germany to make the nontechnical summaries (NTSs) of animal research studies available in a searchable and easily accessible web-based format. Each NTS was manually assigned an ICD-10 code with the goal of advancing the integrity and reporting of responsible animal research. This task requires an automated approach to classify the NTSs, whereby the data exhibits challenging attributes of multilingualism, domain specificity, a relatively small dataset, codes skewness and hierarchical nature.

We explore various models, starting with traditional bag-of-words support vector machines (SVM) to standard deep learning architectures of convolutional neural networks (CNN) and recurrent neural networks (RNN) with three types of attention mechanisms; namely, hierarchical attention Gated Recurrent Unit (GRU), self-attention Long-Short Term Memory (LSTM), and codes attentive LSTM. Finally, we show the effectiveness of fine-tuning state-of-the-art pre-trained BERT models, which requires minimal task specific changes and works well for small datasets. However, the significant performance boost comes from translating the German NTSs to English and then applying the same models, yielding an absolute gain of 6.22% on f-score, from best German model to English model. This can be attributed to the fact that each language has its own linguistic and cultural characteristics that may contain different signals to effectively classify a specific class of documents. Given translated texts, we also find that domain specific embeddings have more effect when considering static word embeddings, giving an avg. gain of 2.77% over contextual embeddings, where the gain is 0.86%.

## 2 Related Work

Automatic assignment of ICD codes to health related documents has been well studied, both in previous CLEF shared tasks and in general. Traditional approaches range from rule based and dictionary look ups to machine learning models. However, more recently the focus has been on applying deep learning.

---

3 https://clefehealth.imag.fr/?page_id=26
4 https://www.dimdi.de/static/de/klassifikationen/icd/icd-10-gm/kode-suche/htmlgm2016/
Many techniques have been proposed using CNNs, RNNs and hybrid systems. [12] uses shallow CNN and improves its predictions for rare labels by dictionary-based lexical matching. [4] addresses the challenges of long documents and high cardinality of label space in MIMIC-III [18] by modifying Hierarchical Attention Network [41] with labels attention. More recent focus has been on using sequence-to-sequence (seq2seq) [37] based encoder-decoder based architectures. [33] first builds a multilingual death cause extraction model using LSTMs encoder-decoder, with concatenated French, Hungarian and Italian fastText embeddings as inputs and causes extracted from ICD-10 dictionaries as outputs. The output representations are then passed to an attention based biLSTM classifier which predicts the codes. [16] uses character level CNN [43] encoders for French and Italian, which are genealogically related languages and similar on a character level, with a biRNN decoder. [17] enriches word embeddings with language-specific Wikipedia and creates an ensemble model from a CNN classifier and GRU encoder-decoder. Few other techniques have also been proposed to use sequence-to-sequence framework and obtained good results [2,27].

While successful, these approaches make an auto-regressive assumption on output codes, which may hold true only when there is one distinct path from parent to child code for a given document. However, in the ICD codes assignment, a document can have multiple disjoint paths in a directed acyclic graph (DAG), formed by concepts hierarchy [35]. Also, for a smaller dataset, the decoder may suffer from low variance vocabulary and data sparsity issues. In [7], a novel Hierarchical Multi-label Classification Network (HMCN) with feed-forward and recurrent variations is proposed that jointly optimizes local and global loss functions for discovering local hierarchical class-relationships in addition to global information from the entire class hierarchy while penalizing hierarchical violations (a child node getting a higher score than parent). However, they only consider tree based hierarchies where a node strictly has one parent.

Contextualized word embeddings, such as ELMo [31] and BERT [11], derived from pre-trained bidirectional language models (biLMs) and trained on large texts have shown to substantially improve performance on many NLP tasks; question answering, entailment and sentiment classification, constituency parsing, named entity recognition, and text classification. Such transfer learning involves fine-tuning of these pre-trained models on a down-stream supervised task to get good results with minimal effort. In this sense, they are simple, efficient and performant. Motivated by this, and recent work of [23], we use BERT models for this task and achieve better results than CNN and RNN based methods. We also show great improvements with translated English texts.

3 Data

The dataset contains 8,385 training documents (including dev set) and 407 test documents, all in German. Each document has six text fields:

- Title of the document
<table>
<thead>
<tr>
<th>ICD-10 Code</th>
<th>No. of documents (train + dev)</th>
</tr>
</thead>
<tbody>
<tr>
<td>II</td>
<td>1515</td>
</tr>
<tr>
<td>C00-C97</td>
<td>1479</td>
</tr>
<tr>
<td>IX</td>
<td>930</td>
</tr>
<tr>
<td>VI</td>
<td>799</td>
</tr>
<tr>
<td>C00-C75</td>
<td>732</td>
</tr>
</tbody>
</table>

Table 1: Top-5 most frequent codes

- Uses (goals) of the experiment
- Possible harms caused to the animals
- Comments about replacement (in the scope of the 3R principles)
- Comments about reduction (in the scope of the 3R principles)
- Comments about refinement (in the scope of the 3R principles)

The documents are assigned one or more codes from ICD-10-GM (German Modification version 2016). The codes exhibit a hierarchy forming a DAG [35], where the highest-level nodes are called chapters and their direct child nodes are called groups. The depth of most chapters is one but in some cases there is a second-level (e.g. M00-M25, T20-T32), and in one case a third-level (C00-C97), of depth. Documents are assigned mixed codes such that, a parent and child node can both be present at the same time and a child node can have multiple parents. Moreover, 91 documents are missing one or more of six text fields and only 6,472 have labels (5,820 in train set and 652 in dev set), while 52 of them have only chapter level codes. Table 1 shows top-5 most frequent codes. These classes account for more than 90% of the dataset leading to a high imbalance. Due to a shallow hierarchy, we consider the problem as multi-label classification instead of hierarchical classification.

4 Methods

Since the documents are domain specific and in German, we argue that it might be difficult for open-domain and multilingual pre-trained models to do effective transfer learning. Furthermore, [1] suggests that each language has its own linguistic and cultural characteristics that may contain different signals to effectively classify a specific class. Based on this, and the fact that translations are always available as domain-free parallel corpora, we use them in our system and show absolute gains of 6% over best German model. Since English has readily more accessible biomedical literature available as free texts [26], we use English translations for our documents. To perform a thorough case study, we tested several models and pre-trained embeddings. Below we describe each of them.

Baseline: For baseline we use a TF-IDF weighted bag-of-words linear SVM model.
The increasing obesity (obesity) in the population and the associated health problems in the form of increasing cases of type 2 diabetes, hypertension, lipid metabolism disorders and cancers pose a socio-economic challenge and need new therapeutic interventions. A particular problem is the fact that more and more adolescents are morbidly overweight, suffer from type 2 diabetes and suffer the known long-term consequences such as blindness and amputations. The aim of this animal experiment is to investigate the influence of the secreted protein cysteinyI A (Eda) and its intronic microRNA miR-676 on glucose metabolism and the development of diabetes. Eda and miRNA-676 form a so-called "gene microRNA" pair that is parallel and proportionally upregulated in the liver of adipose mice, causing inflammatory processes. The results obtained will contribute to a better understanding of the regulation of glucose metabolism and the development of type 2 diabetes. Finally, the aim is to achieve new therapeutic approaches to obesity and its consequences through the described experiments.

CNN: Convolutional Neural Network (CNN) learns local features of input representation through varying number and sizes of filters performing convolution operation. They have been very successful in many text classification tasks [9,19,43]. While many advanced CNN architectures exist, we use a shallow model of [20].

Attention Models: Attention is a mechanism that was initially proposed in sequence-to-sequence based Neural Machine Translation (NMT) [3] to allow decoder to attend to encoder states while making predictions. More generally, attention generates a probability distribution over features, allowing models to put more weight on relevant features. In our study, we used three attention based models.

- HAN Hierarchical Attention Network (HAN) [41] deals with the problem of long documents classification by modeling attention at each hierarchical level of document i.e. words and sentences. This allows the model to first attend word encoder outputs, in a sentence, followed by attending the sentence encoder outputs to classify a document. Like [41], we also use bidirectional Gated Recurrent Units (GRUs) as word and sentence encoder.

- SLSTM Self-Attention Long-Short Term Memory (SLSTM) network is a simple single layer network based on bidirectional LSTMs encoder. An input sequence is first passed through the encoder and encoded representations are self-attended [24] to produce outputs.

- CLSTM All ICD codes have a textual description, e.g. code A80-A89 is about viral infections of the central nervous system that can help a model while classifying. Figure 1 shows a document containing words related to those found in the descriptions of their labeled codes. Such words may or may not be present but, the intuition is to use this additional meta-data to enrich the encoder representation by attention. [4] use label attention but they consider label as a unit of representation instead of their text descriptions. We call this network, Codes Attentive LSTM (CSLSTM).

Let $X = \{x_1, x_2, ..., x_n\} \in \mathbb{R}^{n \times d}$ be an n-length input document sequence, where $x_i$ is a d-dimensional embedding vector for input word $w_i$ belonging to docu-
ments vocabulary \( V_D \). Let \( T = \{ t_1, t_2, ..., t_m \} \in \mathbb{R}^{m \times l} \) be \( m \)-codes by \( l \)-length titles representation matrix, where each \( t_i = \{ t_{i1}, t_{i2}, ..., t_{il} \} \in \mathbb{R}^{l \times d} \) and \( t_{ij} \) is \( d \)-dimensional embedding vector for code \( i \)'s title word \( j \), belonging to titles vocabulary \( V_T \). The embedding matrices are different for documents and codes titles, this is because the title words can be missing in documents vocab. Similarly, we used different LSTM encoders for document and code words (shared encoder under performed on dev set; not reported). The network then transforms input as \( X_{out} = \text{CLSTM}(X, T) \), with following operations:

\[
X_{enc} = [x_{1enc}, x_{2enc}, ..., x_{nenc}]
\]

\[
x_{ienc} = \text{LSTM}_W(x_i)
\]

\[
T_{enc} = [t_{1enc}, t_{2enc}, ..., t_{menc}]
\]

\[
t_{ienc} = \frac{1}{l} \sum_{j=1}^{l} \text{LSTM}_C(t_{ij})
\]

\[
X_{out} = [X_{enc}, T_{enc}] \in \mathbb{R}^{(n+m) \times h}
\]

\[
A = \text{softmax}(X_{out}X_{out}^T) \in \mathbb{R}^{(n+m) \times (n+m)}
\]

\[
X_{out} = X_{out} + A^TX_{out}
\]

\[
X_{out} = \frac{1}{n} \sum_{j=1}^{n} X_{outj}
\]

Where, \( X_{enc} \) is a sequence of word encoder LSTM \( W \) outputs and \( T_{enc} \) is a sequence of averaged title words encoding by code encoder LSTM \( C \). We concatenate document words sequence with titles sequence and perform self-attention \( A \), followed by residual connection and average over resulting sequence to get final representation.

**BERT**: Pre-training large models on unsupervised corpus with language modeling objective and then, fine-tuning the same model for a downstream supervised task eliminates the need of heavily engineered task-specific architectures [11, 31, 32]. Bidirectional Encoder Representations from Transformers (BERT) is a recently proposed such model, following ELMo and OpenAI GPT. BERT is a multi-layer bidirectional Transformer (feed-forward multi-headed self-attention) [39] encoder that is trained with two objectives, masked language modeling (predicting a missing word in a sentence from the context) and next sentence prediction (predicting whether two sentences are consecutive sentences). BERT has improved the state-of-the-art in many language understanding tasks and recent works show that it sequentially model NLP pipeline, POS tagging, parsing, NER, semantic roles and coreference [38]. Similar works [14, 42] have been performed to understand and interpret BERT’s learning capacity. We therefore use BERT in our task and show that it achieves best results compared to other models and is nearly agnostic to domain specific pre-training (BioBERT; [23]).
5 Experiments

5.1 Pre-processing

We consider each document as one text field i.e. all six fields are considered together to form one document. We then translate all of our documents to English, using Google Translate API v2. For both, German and English, we use language specific sentence and word tokenizer offered by NLTK and spaCy, respectively. Tokens with document frequencies outside 5 and 60% of training corpus were removed and only top-10000 tokens were kept to limit the vocabulary. This applies to all models other than BERT, which uses WordPiece tokenizer and builds its own vocabulary. Lastly, we remove all the classes with frequency less than 15. All the experiments were performed without any cross-validation on dev set to find best parameters.

5.2 Embeddings

We use following pre-trained models for German:

- FT\textsubscript{de}: fastText DE Common Crawl (300d)
- BERT\textsubscript{de}: BERT-Base, Multilingual Cased (768d)

and following pre-trained models for English:

- FT\textsubscript{en}: fastText EN Common Crawl (300d)
- PubMed\textsubscript{en}: PubMed word2vec (400d)
- BERT\textsubscript{en}: BERT-Base, Cased (768d)
- BioBERT\textsubscript{en}: BioBERT (768d)

5.3 Models

**TF-IDF+Linear SVM:** For baseline, we use scikit-learn implementation of LinearSVC with One-vs-All training.

For all models we use batch size of 64, max sequence length of 256, learning rate of 0.001 with Adam and 50 epochs with early stopping. We use binary cross-entropy for each class as our objective function and F\textsubscript{1} score as performance metrics as provided by challenge evaluation script.

**CNN:** We use 64 channels with filter sizes of 3, 4 and 5.

---

5 https://cloud.google.com/translate/docs/translating-text
6 https://spacy.io/usage/models
8 https://archive.org/details/pubmed2018_w2v_400D.tar
<table>
<thead>
<tr>
<th>Model</th>
<th>DE</th>
<th></th>
<th></th>
<th>EN</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
<td>P</td>
<td>R</td>
<td>F1</td>
</tr>
<tr>
<td>Baseline</td>
<td>90.72</td>
<td>58.73</td>
<td>71.30</td>
<td>90.69</td>
<td>65.45</td>
<td>76.63</td>
</tr>
<tr>
<td>CNN+FT_de</td>
<td>86.08</td>
<td>57.37</td>
<td>68.85</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CNN+FT_en</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>85.76</td>
<td>61.59</td>
<td>71.69</td>
</tr>
<tr>
<td>CNN+PubMed_en</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>87.95</td>
<td>65.10</td>
<td>74.82</td>
</tr>
<tr>
<td>HAN+FT_de</td>
<td>78.86</td>
<td>58.79</td>
<td>67.37</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>HAN+FT_en</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>83.52</td>
<td>64.50</td>
<td>72.79</td>
</tr>
<tr>
<td>HAN+PubMed_en</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>85.10</td>
<td>69.61</td>
<td>76.58</td>
</tr>
<tr>
<td>SLSTM+FT_de</td>
<td>85.55</td>
<td>64.86</td>
<td>73.76</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SLSTM+FT_en</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>87.53</td>
<td>67.65</td>
<td>76.32</td>
</tr>
<tr>
<td>SLSTM+PubMed_en</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>87.33</td>
<td>70.09</td>
<td>77.77</td>
</tr>
<tr>
<td>CLSTM+FT_de</td>
<td>83.60</td>
<td>63.97</td>
<td>72.48</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CLSTM+FT_en</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>84.39</td>
<td>69.14</td>
<td>76.01</td>
</tr>
<tr>
<td>CLSTM+PubMed_en</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>87.87</td>
<td>70.21</td>
<td>78.05</td>
</tr>
<tr>
<td>BERT_de</td>
<td>70.96</td>
<td>83.41</td>
<td>76.68</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BERT_en</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>79.63</td>
<td>84.60</td>
<td>82.04</td>
</tr>
<tr>
<td>BioBERT_en</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>80.35</td>
<td>85.61</td>
<td>82.90</td>
</tr>
</tbody>
</table>

Table 2: Results on development set

**HAN**: We use biGRUs as encoders with hidden size of 300, with 10 and 40 as max no. of sentences in a document and words in a sentence respectively.

**SLSTM**: A biLSTM encoder with hidden size of 300.

**CLSTM**: Similar to SLSTM but, with additional $T$ matrix of size total number of titles (230, collected from ICD-10-GM) by max title sequence length of 10.

**BERT**: We use PyTorch implementation of BERT for multi-label classification. Default configurations of each BERT model was used with max sequence length of 256 and batch size of 6.

### 5.4 Results

Table 2 summarizes the results of all models with different pre-trained embeddings. In all of our experiments, working on translated texts (English) improved the score by an avg. of 4.07%. This can be attributed to the fact that there is much more English texts than other languages but it can also be argued that English may have stronger linguistic signals to classify the classes where German models make mistakes [1].

Baseline proved to be a strong one, with highest precision of all and outperforming HAN and CNN models, for both German and English, with common crawl embeddings. HAN performs better when documents are relatively long e.g. [4] reports strong results with HAN based models on MIMIC dataset [18].

---

9 [https://github.com/huggingface/pytorch-pretrained-BERT](https://github.com/huggingface/pytorch-pretrained-BERT)
where average document size exceeds 1900 tokens. After pre-processing, the averaged document length in our case was approximately 340. For CNN, we believe advanced variants may perform better.

SLSTM and CLSTM, both being just one layer, performed comparably and better than baseline. SLSTM is much simpler and relies purely on self-attention, which also compliments higher scores by BERT models, which are stacked multi-headed self-attention networks. For CLSTM, one can argue that since most of the documents were missing the title words (in fact many title words never appeared in corpus) therefore, the model couldn’t formulate strong alignments between documents and this additional meta-data.

BERT performed better than other models, both in German and English with an avg. score of 6% points higher. BioBERT\textsubscript{en} performed slightly (+0.86%) better than BERT\textsubscript{en}, this was also noticeable in Relation Extraction task in [23], where domain specific and general BERT performed comparably. This partly shows BERT’s ability to generalize and being robust to domain shifts (learning from 5k training docs), however, this slightly contradicts the findings of [42], where authors reflect on such issues, and catastrophic forgetting in BERT-like models. On other hand, the effect of using in-domain pre-trained models was more significant for static-embeddings; using pre-trained PubMed\textsubscript{en} vectors outperformed open-domain FT\textsubscript{en} by an avg. of 2.77%. Such analysis wasn’t performed for German due to lack of medical domain German vectors. BERT models had highest recall but relatively poor precision. This is preferable in real-world medical applications, where the recall is of much more importance. On test set, we submitted a single run of our best model and obtained an F\textsubscript{1} score of 73% (with 86% recall and 64% precision), in contrast, on dev set, same model reached F\textsubscript{1} score of 83% with 86% recall and 80% precision. Test results of other systems can be found at challenge website.

6 Discussion

Biomedical text mining is generally a challenging field but recent progresses of transfer learning in NLP can significantly reduce the engineering required to come up with domain sensitive models. Unsupervised data is cheap, and can be obtained in abundance to learn general language patterns [42], however, such data may become harder to obtain when dealing with in-domain and low-resource languages (e.g. Estonian medical documents). Such deficiencies encourage research for better cross-lingual and cross-domain embedding alignment methods that can transferred effectively.

Acknowledgements

We thank Stalin Varanasi for helpful discussions. This work was partially funded by European Union’s Horizon 2020 research and innovation programme under grant agreement No. 777107.
References


