Classifying German Animal Experiment Summaries with Multi-lingual BERT at CLEF eHealth 2019 Task 1

Mario Sänger[1⋆], Leon Weber[1⋆], Madeleine Kittner[1⋆], and Ulf Leser[1]

Humboldt Universität zu Berlin, Knowledge management in Bioinformatics, Berlin, Germany
{saengema,weberple,kittner,leser}@informatik.hu-berlin.de

Abstract. In this paper we present our contribution to the CLEF eHealth challenge 2019, Task 1. The task involves the automatic annotation of German non-technical summaries of animal experiments with ICD-10 codes. We approach the task as multi-label classification problem and leverage the multi-lingual version of the BERT text encoding model [6] to represent the summaries. The model is extended by a single output layer to produce probabilities for individual ICD-10 codes. In addition, we make use of extra training data from the German Clinical Trials Register and ensemble several model instances to improve the overall performance of our approach. We compare our model with five baseline systems including a dictionary matching approach and single-label SVM and BERT classification models. Experiments on the development set highlight the advantage of our approach compared to the baselines with an improvement of 3.6%. Our model achieves the overall best performance in the challenge reaching an F_1 score of 0.80 in the final evaluation.

Keywords: ICD-10 Classification · German Animal Experiments · Multi-label Classification · Multi-lingual BERT Encodings

1 Introduction

Biomedical natural language processing (NLP) aims to support biomedical researchers, health professionals in their daily clinical routine as well as patients and the public searching for disease-related information. A large part of Biomedical NLP focuses on extraction of biomedical concepts from scientific publications or classification of such documents to biomedical concepts. In the past biomedical NLP has strongly advanced for biomedical or clinical documents in English [7]. Non-English biomedical NLP lags behind since the availability of annotated

⋆ These authors contributed equally.

corpora and other resources (e.g. dictionaries and ontologies for biomedical concepts) in non-English languages is limited.

Since 2015 the CLEF eHealth community addresses this issue by organising shared tasks on non-English or multilingual information extraction. The subject of CLEF eHealth shared tasks since 2016 [13–15] include the classification of clinical documents according to the International Classification of Diseases and Related Health Problems (ICD-10) [17]. More precisely, the task has been the assignment of ICD-10 codes to death certificates in Frensh, English, Hungarian and Italian. Among the best performing teams in 2018, the task has been treated as a multi-label classification problem or as sequence-to-sequence prediction leveraging neural networks [2]. Other well performing systems were based on a supervised learning system using multi-layer perceptrons and an One-vs-Rest (OVR) strategy supplemented with IR methods [1] or an ensemble model for ICD 10 coding prediction utilising word embeddings created on the training data as well as on language-specific Wikipedia articles [9].

In 2019, the CLEF eHealth Evaluation Task 1 focuses on the assignment of ICD-10 codes to health-related, non-technical summaries of animal experiments in German [10, 16]. According to the laws of the European Union each member state has to publish a comprehensible, nontechnical summary (NTS) of each authorised research project involving laboratory animals to provide greater transparency and increase the protection of animal welfare. In Germany the web-based database AnimalTestInfo[1] houses and publishes planned animal studies to inform researchers and the public. To improve analysis of the database, summaries submitted in 2014 and 2015 (roughly 5.300) were labelled by human experts according to the German version of the ICD-10 classification system[2] in [4]. Based on this pilot study further documents added to the database have been labelled and used to conduct this year's CLEF eHealth challenge. The task is to explore the automatic assignment of ICD-10 codes to the animal experiments, i.e. given the non-technical summary predicting the ICD-10 codes that are investigated in the study.

We treat the task as a multi-label classification problem and apply the multilingual BERT model [6] which recently achieved state-of-the-art results in eleven different NLP tasks [12]. The model is extended by a single output layer to produce probabilities for individual ICD-10 codes. Since training data in this task is sparse, we also use summaries of clinical trails conducted in Germany published by the German Clinical Trials Register (GCTR). We compare our model with five baseline systems including a dictionary matching approach and single-label SVM and BERT classification models. The implementation of our models is available as open source software at github[3].

[1] https://www.animaltestinfo.de/

[2] https://www.dimdi.de/static/de/klassifikationen/icd/icd-10-gm/kode-suche/htmlgm2016/

[3] https://github.com/mariosaenger/wbi-clef19x

2 Method

Here we describe the corpora, used terminologies and classification models we use in the task.

2.1 Corpora and Terminologies

The lab organisers provided a corpus of 8,385 German non-technical summaries of animal experiments (NTS) originating from the AnimalTestInfo database. For each experiment a short title is given followed by a description of expected benefits as well as pressures and damages of the animals. Furthermore, strategies to prevent unnecessary harm to the animals and to improve animal welfare are described. Each summary was labeled by experts using the German version of the ICD-10 classification system. Depending on the level of detail of the summary different levels (e.g. chapter, group) of the ICD-10 ontology are used to annotated the experiment. About two-thirds of the experiments are labeled with exactly one disease and 10% with multiple diseases; the remainder have no annotated disease. For each disease the complete path in the ICD-10 ontology, i.e. up to two parent groups and the chapter of the annotated disease, is given. About two third of the summaries are annotated with 2-level paths (e.g. *I | B50-B64*), 20 % with 3- or 4-level paths (eg. *IV | E70-E90 | E10-E14* or *II | C00-C97 | C00-C75 | C15-C26*) and less than 1 % of the summaries are only annotated with chapters (e.g. *VI*). The data set is divided into a stratified train and development split (7,543 / 842) at document level. For the final evaluation an hold-out set of 407 experiments are used by the organisers.

In addition to the provided data set, we use information from the German Clinical Trials Register (GCTR)[4]. The GCTR provides access to basic information (e.g. trial title, short description, studied health condition, inclusion and exclusion criteria) of clinical trials conducted in Germany and is also annotated with ICD-10 codes. We downloaded all trials available through the GCTR website. For each trial we make use of the title as well as the scientific and lay language summary. We use the chapter and all (sub-) groups up to the third level of the ontology of the given ICD-10 codes describing the studied health condition as labels for the trial, similar to ICD-10 coding in the NTS data set. In this way we are able to extend the training set by 7,615 documents having 18,263 ICD-10 codes. ICD-codes of each study in the GCTR data set relate to the ICD-10 version valid at publication of a study. We did not adjust for any differences (e.g. any potentially missing ICD-10 codes) to version 2016 used for the NTS corpus. The two data sets almost fully overlap with regard to the considered health problems. Of the 233 distinct ICD-10 codes occurring in the complete NTS corpus, 226 (97%) are mentioned in GCTR too. Moreover, 27 other ICD-10 codes will be introduced through the additional data set. Table 1 summarises the used corpora.

[4] `https://www.drks.de/drks_web/setLocale_EN.do`

Table 1. Overview about the used data sets. The non-technical animal experiment summaries (NTS) are provided by the task organisers. Furthermore, we build a second data set based on the German Clinical Trails Register (GCTR).

		#Documents	#ICD-10 codes	#ICD-10 codes (distinct)
NTS	Train	7453	15251	230
	Dev	842	1682	156
GCTR		7615	18263	253

2.2 BERT for multi-label classification

Our approach for the task is based on BERT language model [6]. BERT is a text encoding model that recently achieved state-of-the-art results in many different NLP tasks [12]. It is a neural network based on the transformer architecture of [19], which was pretrained using two different language modelling tasks: masked language modeling and next sentence prediction. Specifically, we use the multilingual version of BERT-Base[5] that has been pre-trained on Wikipedia dumps of 104 different languages including German.

Given a sequence of tokens $t_1, \ldots, t_L$, BERT first subdivides the tokens into subword-tokens, yielding a new (usually, longer) sequence $s_1, \ldots, s_N$ using WordPiece [21]. Then, it produces vector representations for each subword-token $e_1, \ldots, e_N \in \mathbb{R}^{768}$ and one vector $c \in \mathbb{R}^{768}$ which is not tied to a specific token. BERT supports sequence lengths up to 512 sub-word tokens. We represent each animal experiment by taking as much as possible sub-word tokens from the title and the description of expected benefits and pressures of the summary text as model input. Following [6], we employ c as a representation for the whole token sequence.

We treat the assignment of ICD-10 codes as a one-versus-rest multi-label classification problem [5], i.e. as $|\mathcal{Y}|$ independent binary classification tasks, where $\mathcal{Y}$ is the set of all ICD-10 codes occurring in the training set. Each example is used as a positive example if it has the respective label, while all other examples are used as negative examples. The only connection between the individual classification tasks is the BERT encoder which is shared between all tasks and which receives parameter updates from all of them. We use a single output layer $W \in \mathbb{R}^{768 \times |\mathcal{Y}|}$ to compute the output probabilities per class with $\sigma(c \cdot W)$, where σ is the element-wise sigmoid function, and use binary cross-entropy as a loss.

We implement our model in PyTorch [18] using the *pytorch-pretrained-BERT*[6] implementation of BERT and use the included modified version of Adam [11] for optimization. We train our model for 60 epochs on a single Nvidia V100 GPU, which takes about nine hours. In principle, it would also be possible to train

[5] `https://storage.googleapis.com/bert_models/2018_11_23/multi_cased_L-12_H-768_A-12.zip`

[6] `https://github.com/huggingface/pytorch-pretrained-BERT`

and evaluate the model using only CPUs but that would take considerably more time.

We train multiple model instances using different random seeds and ensemble their predictions. Ensembling of multiple neural network models has shown to be beneficial in several NLP taks [6]. We ensemble the models in two ways: (1) by averaging the predictions of the different model instances and (2) learning a logistic regression classifier based on the model outputs on the development set. We denote the two ensembling model variants as BERT multi-label Avg and BERT multi-label LogReg. Note, that because BERT multi-label LogReg is trained on the development set, the resulting scores on this data are no longer a reliable estimate for out-of-sample performance and can only be fairly compared to the other approaches on the development set.

2.3 Baselines

To gain better insights about the performance level of our approach we compare it with five different baseline methods. First, we implement a dictionary-matching approach. For this we took the concept descriptions of all codes listed in the ICD-10 ontology as well as all given synonyms and search for occurrences of these terms in the title and goals (line 1 and 2) of an animal trial summary. Dictionary matching is performed by indexing all ICD-10 concepts using Apache Solr 7.5.0[7] and applying exact and fuzzy matching. Each ICD-10 concept is linked to its related path up to the chapter-level which is used for annotation. All concepts matched by the dictionary are reported as results. We do not perform any further post-processing like sorting out overlapping ICD-10 paths. For the other baselines we transform the task into (1) a group-level or (2) a sub-group-level classification problem, i.e. we use the label on the second level of the ICD-10 hierarchy (e.g. for *I* | *C00-C97* | *C00-C75* we use *C00-C97*) resp. the deepest label (e.g. for *I* | *C00-C97* | *C00-C75* we use *C00-C75*) for a given trial summary as gold standard. In both cases, for instances with multiple codes originating from different branches of the ICD-10 ontology we use the first label as gold standard. Moreover, we add a special *no-class* label to support documents without any annotated ICD-10 code.

We investigate two different classification methods for the tasks, Support Vector Machines (SVM) and BERT sequence classification model[6]. For the former, we build TF-IDF vectors as input representation for the trial summaries. For the latter, the model architecture is equivalent to our multi-label model except that the final linear layer calculates a soft-max over the classes of the classification task and hence applies a (single-class) cross-entropy loss for training.

For the both classification baselines, we augment the predictions of the models according to the ICD-10 hierarchy, e.g. if a group-level model predicts *C00-C97* we automatically add the parent chapter (in this case *I*) to the prediction.

[7] `https://lucene.apache.org/solr/`

3 Results & Discussion

3.1 Experimental setup

We use the training split of the provided corpus as well as the documents from the GCTR data set to train our multi-label as well as all baseline models. For the BERT multi-label and the SVM classification models we perform hyperparameter optimisation and select the best model of each approach based on the development set performance. Regarding the SVM models, we follow [8] and test $\{2^{-5}, 2^{-3}, \ldots, 2^{15}\}$ as values for the C parameter. The best scores are reached with $C = 2$ / $C = 0.5$ for group level / sub-group-level classification. In case of our BERT multi-label approach we only tune the learning rate parameter. We evaluate the sequence $\{5e-5, 4e-5, 3e-5, 2e-5, 1e-5\}$ and found that $4e-5$ achieves the highest scores. We omit hyperparameter tuning for the BERT classification models due to time constraints. Therefore, we use the default parameter settings of the model, i.e. learning rate of $5e-5$.

As described in Section 2.2 we learn eight model instances of our approach using different random seeds and ensemble them. The two ensemble variants are built (a) by averaging of the two best model instances and (b) learning a logistic regression classifier based on the output of the three models with the highest scores. The latter is trained on the output of the individual model instances on the development set. We opt for this settings based on preliminary experiments on the training and development set.

To gain insights about the effectiveness of the additional data, we evaluate each model (except for the ensemble models) in two data configuration settings: with and without the additional texts from the GCTR data set (see Section 2.1). We use the provided evaluation script and report precision, recall and F1 scores as evaluation metrics.

3.2 Development results

Table 2 highlights the results of all evaluated models on the provided development set, both with and without the additional data from GCTR as training data. The best single model performance is reached by the BERT subgroup baseline model. In this setting the model achieves an F_1 score of 0.778. Almost the same performance can be reached by our BERT multi-label approach (0.776). However, the latter offers a clearly better performance if the provided training set is extended by the GCTR samples (0.81 vs. 0.782). This represents an improvement of 0.028 (+3.6%) in terms of F_1. For both baseline classification methods the sub-group models outperform the group-based variants, as to be expected. In case of the SVM, the performance increases from 0.655 to 0.717 (+ 9.5%) if considering sub-group labels instead group labels. With the BERT model the performance increases by 10.4% from 0.705 to 0.778. Interestingly, the BERT group level model performs nearly on par with the sub-group level SVM model. This is especially noteworthy as we do not perform hyperparameter optimisation for BERT group / sub-group but for the corresponding SVM models.

This highlights the effectiveness and suitability of the BERT model for this task, since in general SVMs offer competitive performance for document classification problems [20].

The dictionary matching can't compete with the machine learning based solutions. Even through the matching of the concept terms with the trail summaries provides the highest recall (0.894) of all evaluated approaches, the precision of the approach is very low (0.416) due to many false positives. In particular, the approach often predicts incorrect chapter annotations, for instance the chapter *XXI* 681 times. This is because of the broad and general topic of the chapters respectively their descriptions, e.g. *XXI* is about "Factors influencing health status and contact with health services".

Comparing the configurations with and without the GCTR documents, it can be seen that the performance increases (at least slightly) for all considered models. Improvements range from 0.5% (SVM Sub-group) to 1.9% (BERT group) for the baseline systems with respect to their variants without the additional data. In contrast, the multi-label model can benefit more greatly from the extended training set (+3.2%).

The overall best performance is achieved by ensembling the best BERT multi-label models. In both ensembling variants the model reaches an F_1 score of 0.815. This represents an increase of 0.6% over the single model.

Table 2. Evaluation results of our model (last three rows) and the five baseline approaches (first five rows) on the provided development set. We report precision, recall and F_1 scores in two data scenarios: (left) using only the provided training data and (right) using documents from German Clinical Trail Register as additional training instances.
*: Ensembling trained on development set.

Model	**NTS data**			**NTS + GCTR data**		
	P	**R**	**F1**	**P**	**R**	**F1**
Dictionary matching	0.416	0.894	0.568	-	-	-
SVM group	0.778	0.565	0.655	0.813	0.554	0.659
SVM sub-group	0.804	0.646	0.717	0.815	0.653	0.725
BERT group	0.810	0.624	0.705	0.820	0.640	0.719
BERT sub-group	0.811	0.748	0.778	0.833	0.737	0.782
BERT multi-label	0.901	0.747	0.776	0.834	0.788	0.810
BERT multi-label Avg	-	-	-	0.850	0.782	0.815
BERT multi-label LogReg	-	-	-	0.808*	0.822*	0.815*

3.3 Development predictions

We further analysed the predictions made by the different approaches. Figure 1 (left) compares the true positives of our BERT multi-label model as well as the

SVM and BERT sub-group baseline (all with the GCTR corpus as additional training data). We exclude the dictionary matching baseline for this investigation, since the approach predicts too optimistically and thereby distorts the picture.

First of all it can be noted that, in total 1,422 of the 1,682 gold standard ICD-10 codes are identified by at least one of the three methods. This corresponds to 84.5% of the complete development data set. The intersection of all three methods consists of 1,001 true positives. This represents 70.4% of all correctly identified codes. Additionally, 1,240 (87.2%) labels are predicted by two of the three methods. Furthermore, it can be seen that 110 true positives are exclusively identified by our multi-label approach. This constitutes 7.7% of all correctly found codes. In contrast, 98 codes (6.9%) were predicted by (at least) one of the two classification baseline and not detected by our BERT multi-label approach. We tried to investigate the differences between the multi-label and the classification models but can't come up with a clear (error) pattern.

We also perform the investigation using the best ensembled version of our approach (BERT multi-label LogReg). Figure 1 (right) highlights the results of this comparison. Through the ensembling we are able to additionally identify 20 labels correctly. Moreover, 38 ICD-10 codes that were exclusively predicted by the classification baselines previously are now detected by the multi-label approach too. However, when interpreting the figures one has to keep in mind that the logistic regression model that ensembles the predictions of the individual model instances is trained on the development set and hence may tend to represent an over-optimistic picture.

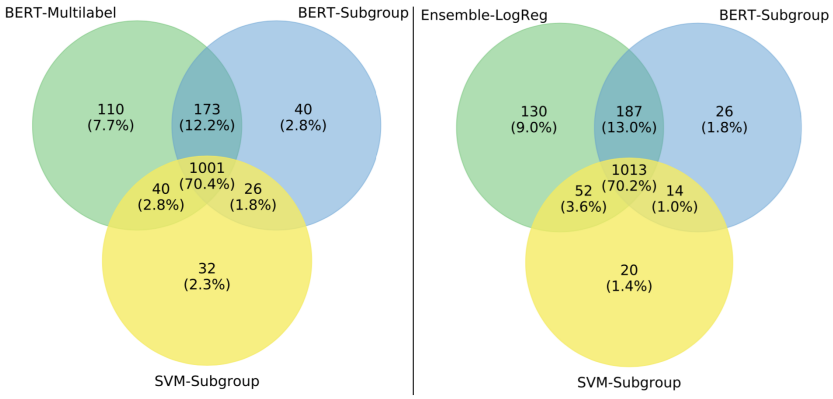

Fig. 1. Comparison of the predicted true positive ICD-10 codes of the evaluated models. On the left the best (single) instance of our BERT multi-label model is contrasted with the best SVM and BERT classification baseline. The diagram on the right shows the changes when using the best ensemble model of our approach (BERT multi-label LogReg).

3.4 Test results

Table 3 shows the results of the final evaluation performed by the task organisers. Every team was allowed to submit up to 3 runs of their approaches. We submitted three different runs: the best single model instance (according to the development results) of BERT multi-label (*WBI-run1*) as well as the Avg- and LogReg-ensemble (*WBI-run2* / *WBI-run3*). All models are trained on the GCTR-extended data.

Table 3. Results of the final evaluation performed by the task organisers. They report precision, recall and F1 scores. We submitted three runs: BERT multi-label (*WBI-run1*), Avg- (*WBI-run2*) and LogReg-ensemble (*WBI-run3*). Our models achieve the best performance in the challenge. Bold figures highlight the highest value per column.

Team	Run	P	R	F1
DEMIR	run1	0.46	0.50	0.48
	run2	0.49	0.44	0.46
	run3	0.46	0.49	0.48
IMS-UNIPD	run1	0.00	0.00	0.00
	run2	0.009	0.50	0.017
	run3	0.10	0.05	0.07
MLT-DFKI		0.64	**0.86**	0.73
SSN-NLP	run1	0.19	0.27	0.22
	run2	0.19	0.27	0.23
	run3	0.13	0.34	0.36
TALP-UPC		0.37	0.35	0.36
WBI	run1	0.83	0.77	**0.80**
	run2	**0.84**	0.74	0.79
	run3	0.80	0.78	0.79

The overall best performance is accomplished by the single BERT multi-label model. In this setting the model achieves an F_1 score of 0.80. The model shows a slightly better precision (0.83) than recall (0.77). Comparing the model with both ensembling variants it can be seen that all models perform almost on par and just leverage slightly different precision-recall trade-offs. The Avg-ensemble of the best models (*run2*) predicts more conservatively reaching the highest precision (0.84) of all evaluated models, but offers lower recall scores. In contrast, the LogReg-ensemble provides well-balanced precision and recall scores. Moreover, it has to be noted that the final evaluation scores are virtually the same as the development scores. However, no positive effects can be observed through ensembling of multiple models (at least in the considered way).

Comparing our method with the other submissions, it can be seen that our model outperforms the other team's approaches by a large extend. The second best team (MLT-DFKI) reaches a higher recall (0.86) than our multi-label model

(0.77). However, their approach has a lower precision compared to our model (0.64/0.83). This allows our model to achieve a 9.6% higher F_1 score.

4 Conclusion

This paper presents our contribution to Task 1 of the CLEF eHealth competition 2019. The task challenges the automatic assignment of ICD-10 codes to German non-technical summaries of animal experiments.

We approach the task as multi-label classification problem and leverage the multi-lingual version of BERT [5] to represent the summaries. We extend the model with a single output layer to predict probabilities for each ICD-10 code. Furthermore, we utilise additional data from the German Clinical Trails Register to built an extended training data set and hereby improve the overall performance of the approach. Evaluation results highlight the advantage of our proposed approach. Our model achieves the highest performance figures of all submission with an F_1 score of 0.80. Moreover, experiments on the development set illustrate that the model outperforms several strong classification baselines by a large extend.

There are several research questions worth to investigate following this work. Due to the multi-lingual nature of the used BERT encoding model it would be interesting to evaluate our approach in an cross-lingual setup, e.g. apply the learned model to non-German clinical documents or animal trail summaries. For this purpose we want to use the data from the previous editions of the CLEF eHealth challenges, i.e. Italian, English, French and Hungarian death certificates. This is especially interesting, because of the different text format of the certificates. They are much shorter than the animal experiment summaries and contain a lot of abbreviations of medical terms. It is an open question how well our trained model can be transferred to this type of texts. Furthermore, we also plan to inspect other approaches to the task, e.g. modelling the task as question-answering problem. Recently, versions of BERT trained on English biomedical literature have been published [12, 3]. It would be worthwhile to investigate whether an extension of such models to multi-lingual biomedical texts would improve results further.

Acknowledgments

Leon Weber acknowledges the support of the Helmholtz Einstein International Berlin Research School in Data Science (HEIBRiDS). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

References

1. Almagro, M., Montalvo, S., de Ilarraza, A.D., Pérez, A.: Mamtra-med at clef ehealth 2018: A combination of information retrieval techniques and neural networks for icd-10 coding of death certificates

2. Atutxa, A., Casillas, A., Ezeiza, N., Goenaga, I., Fresno, V., Gojenola, K., Martinez, R., Oronoz, M., Perez-de Vinaspre, O.: Ixamed at clef ehealth 2018 task 1: Icd10 coding with a sequence-to-sequence approach. CLEF (2018)
3. Beltagy, I., Cohan, A., Lo, K.: Scibert: Pretrained contextualized embeddings for scientific text. arXiv preprint arXiv:1903.10676 (2019)
4. Bert, B., Dörendahl, A., Leich, N., Vietze, J., Steinfath, M., Chmielewska, J., Hensel, A., Grune, B., Schönfelder, G.: Rethinking 3r strategies: Digging deeper into animaltestinfo promotes transparency in in vivo biomedical research. PLoS biology **15**(12), e2003217 (2017)
5. Bishop, C.M.: Pattern recognition and machine learning. springer (2006)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
7. Habibi, M., Weber, L., Neves, M., Wiegandt, D.L., Leser, U.: Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics **33**(14), i37–i48 (2017)
8. Hsu, C.W., Chang, C.C., Lin, C.J., et al.: A practical guide to support vector classification (2003)
9. Jeblee, S., Budhkar, A., Milic, S., Pinto, J., Pou-Prom, C., Vishnubhotla, K., Hirst, G., Rudzicz, F.: Toronto cl at clef 2018 ehealth task 1: Multi-lingual icd-10 coding using an ensemble of recurrent and convolutional neural networks
10. Kelly, L., Suominen, H., Goeuriot, L., Neves, M., Kanoulas, E., Li, D., Azzopardi, L., Spijker, R., Zuccon, G., Scells, H., ao Palotti, J.: Overview of the CLEF eHealth evaluation lab 2019. In: Cappellato, L., F.N.L.D.E., Müller, H. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Tenth International Conference of the CLEF Association (CLEF 2019). Lecture Notes in Computer Science. Springer, Berlin Heidelberg, Germany (2019)
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
12. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: pre-trained biomedical language representation model for biomedical text mining. arXiv preprint arXiv:1901.08746 (2019)
13. Neveol, A., Goeuriot, L., Kelly, L., Cohen, K., Grouin, C., Hamon, T., Lavergne, T., Rey, G., Robert, A., Tannier, X., et al.: Clinical information extraction at the clef ehealth evaluation lab 2016. In: Proceedings of CLEF 2016 Evaluation Labs and Workshop: Online Working Notes. CEUR-WS (September 2016) (2016)
14. Névéol, A., Robert, A., Anderson, R., Cohen, K.B., Grouin, C., Lavergne, T., Rey, G., Rondet, C., Zweigenbaum, P.: Clef ehealth 2017 multilingual information extraction task overview: Icd10 coding of death certificates in english and french. In: CLEF (Working Notes) (2017)
15. Névéol, A., Robert, A., Grippo, F., Morgand, C., Orsi, C., Pelikán, L., Ramadier, L., Rey, G., Zweigenbaum, P.: Clef ehealth 2018 multilingual information extraction task overview: Icd10 coding of death certificates in french, hungarian and italian. In: CLEF 2018 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS (2018)
16. Neves, M., Butzke, D., Dörendahl, A., Leich, N., Hummel, B., Schönfelder, G., Grune, B.: Overview of the CLEF eHealth 2019 Multilingual Information Extraction. In: Crestani, F., Braschler, M., Savoy, J., Rauber, A., et al. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Tenth International Conference of the CLEF Association (CLEF 2019). Lecture Notes in Computer Science. Springer, Berlin Heidelberg, Germany (2019)

17. Organization, W.H., et al.: The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines. Geneva: World Health Organization (1992)
18. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
20. Wang, S., Manning, C.D.: Baselines and bigrams: Simple, good sentiment and topic classification. In: Proceedings of the 50th annual meeting of the association for computational linguistics: Short papers-volume 2. pp. 90–94. Association for Computational Linguistics (2012)
21. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al.: Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 (2016)