On the Contribution of Per-ICD Attention Mechanisms to Classify Health Records in Languages With Fewer Resources than English

Alberto Blanco[1], Sonja Remmer[2,3], Alicia Pérez[1], Hercules Dalianis[2,3], and Arantza Casillas[1]

[1]HiTZ Center - Ixa, University of the Basque Country UPV/EHU, Donostia, Spain
[2]Department of Computer and Systems Sciences, Stockholm University, Sweden
[3]Norwegian Centre for E-health Research, Tromsø, Norway
{*alberto.blanco,alicia.perez,arantza.casillas*}*@ehu.eus*
{*remmer,hercules*}*@dsv.su.se*

Abstract

We introduce a multi-label text classifier with per-label attention for the classification of Electronic Health Records according to the International Classification of Diseases. We apply the model on two Electronic Health Records datasets with Discharge Summaries in two languages with fewer resources than English, Spanish and Swedish. Our model leverages the BERT Multilingual model (specifically the Wikipedia, as the model have been trained with 104 languages, including Spanish and Swedish, with the largest Wikipedia dumps[1]) to share the language modelling capabilities across the languages. With the per-label attention, the model can compute the relevance of each word from the EHR towards the prediction of each label. For the experimental framework, we apply 157 labels from Chapter XI – Diseases of the Digestive System of the ICD, which makes the attention especially important as the model has to discriminate between similar diseases.

1 Introduction

Electronic Health Records (EHRs) are classified by clinical experts for documentation, reporting global health vital statistics, insurance billing, etc. International Classification of Diseases (ICD) is used world-wide to define diagnostic terms and procedures and serves to encode EHRs. There are thousands of terms encoded within the ICD WHO (2016). For medical experts, reading EHRs, lengthy and technical documents, finding explicit and implicit mentions of diagnoses and procedures for then assigning standard ICD codes is cumbersome and requires specific training. In fact, it is well-known that manual encoding is not error-free, as an example, Jacobsson and Serdén (2013), estimated that 20% of them were either incorrect or were missing. In this context, natural language understanding brings opportunities to bridge the needs of the society in terms of computer aided coding approaches.

In 2006, it was argued that Natural Language Processing (NLP) tools could quickly help identify codes in discharge summaries Kukafka et al. (2006). Today NLP tools for classifying clinical documents written in English are widespread. Even more, languages with scarce resources for biomedical NLP like Spanish, Italian, Swedish, etc., are in the limelight in the last years to develop codification systems as has been done for English. In the context of working towards the codification of documents, in languages with a small number of resources for NLP, different tasks have been addressed. In 2018 CLEF Névéol et al. (2018b) worked with Italian, French and Hungarian for the automatic codification of death certificates. Each death certificate consisted of a few words (on average 20 words) with at least one main diagnosis. In 2020 the CodiEsp task at CLEF Miranda-Escalada et al. (2020) consisted on the automatic assignment of ICD-10 codes to Spanish Clinical Records with 350 tokens on average. For Swedish Henriksson et al. (2011) the authors mentioned that the corpus was compiled with documents that on average had a length of 96 words.

Admittedly, multi-label classification is **challenging**, particularly with extensive label-sets (as it is the case of the ICD) and domain-specific corpora, and even more when it comes to dealing with clinical information extraction on languages other than English Névéol et al. (2018a). Spanish and Swedish researchers are striving to bridge this gap, indeed, as the first and relevant step, they gathered corpora conveying patient records Oronoz et al. (2015); Dalianis (2018). Previous works showed that the multi-label classification problem of EHRs coded with ICD-10 can be tackled with an adapted

[1]https://github.com/google-research/bert/blob/master/multilingual.md#list-of-languages

Proceedings of Recent Advances in Natural Language Processing, pages 165–172
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_020

BERT architecture Amin et al. (2019); Zhang et al. (2020).

Moreover, we focused just on a sub-set of the ICD, i.e. the Diseases of the Digestive System (the ICD codes starting with the letter K). Focusing on semantically related diseases poses an added challenge, since the Natural Language Understanding (NLU) in charge of encoding the input EHR must be able to cope with the nuances inherent to the distinction of similar diseases. Unarguably, it is easier to distinguish two diseases each belonging to a different body-part than two diseases within the same body-part (as it is this case distinguishing diseases all within the digestive system). In summary, distinguishing semantically different diseases (e.g. gastrointestinal vs cardio-pulmonary) would be easier than distinguishing two diseases within the same speciality. To that end, the LM and the attention mechanisms play the most critical role, so we opted for the transformers models. BERT-based approaches have been tested in this context, with attention mechanisms as a strength towards finding relationships between input text with output ICD codes. The attention is a mechanism whose effectiveness has also been shown with other architectures such as RNNs with LSTM units Hochreiter and Schmidhuber (1997) or Convolutional Neural Networks Du et al. (2017).

Nevertheless, in this context we are dealing with scarce resources and relatively similar codes. In this line, the main scientific **contribution** of this paper rests on the implementation of a head adapted for BERT with multiple label attention mechanisms (instead of a generic one) in order to delve deeper into the nuances of the understanding module. In this work, we have implemented a per-label attention mechanism, and given that regular BERT models also have the self-attention mechanism, it allowed us to compare the effect of different attention mechanisms. The per-label attention mechanism allows the model to give a different relevance to each word and ICD code pair, contrary to the regular attention mechanism. The experimental results support the approach's acceptable performance, so we decided to release the head for the scientific community.

2 Corpora

We have applied two datasets of languages with scarce resources for this work, i.e., languages with fewer resources than English, specifically, Spanish and Swedish. Both datasets are Electronic Health Records containing Discharge Summaries from patients. The Spanish EHRs are from the Emergency Services of the Basque Health Public System, conveying records, and therefore labels, from all the medical specialities Oronoz et al. (2015). However, the Swedish EHRs are only from the gastrosurgery medical specialisation and comes from the research infrastructure Health Bank - Swedish Health Record Research Bank[2], at Stockholm University. Therefore, to have equal label sets, we have selected the ICD codes shared between both datasets to carry out the experiments, obtaining 157 codes, all from the Chapter XI of the ICD-10, i.e., Diseases of the Digestive System. By selecting the codes of some specialities the number of available EHRs is reduced but the label sets are easier to handle. Training specific models on EHRs of specialities improves the performance against training general models Blanco et al. (2020). For the Swedish ICD-10 corpus data set the Swedish KB-BERT model Malmsten et al. (2020) has been applied with good results, see (Remmer et al., 2021).

Here we present a quantitative description and comparison between both datasets. Regarding the input, the Swedish dataset is more than twice larger in number of EHRs, with 8,909 records in contrast to the 3,891 available Spanish EHRs. Nevertheless, the vocabulary (i.e., number of unique words) is around three times bigger for the Spanish dataset. One explanation is that the Spanish EHRs come from several specialities, and therefore there is a higher lexical variability due to the specific terms of each medical specialisation. Also, the Spanish EHR contains lab tests, which could increase the number of unique words significantly.

Regarding the output, both datasets are equivalent, with the same set of 157 gastrointestinal ICD-10 codes. Although this is just a subset of the labels, there are still infrequent codes. For example, only 45 codes from the 157 appear in at least 1% of the EHRs. This fact makes the task even more challenging, as, for around 28% of the labels, there are only a few samples from where the model can learn. Even though the number of labels is the same, the distinct label sets (i.e., label combinations that are unique) are larger in the Swedish dataset than in the Spanish (1,288 and 558, respectively) due to the higher number of records. The ratio between the distinct label sets and the number of records

[2]http://dsv.su.se/healthbank

is similar, 6.97 for Spanish and 6.91 for Swedish, meaning that about the same number of EHRs lead to the same number of unique label sets.

The most significant differences come when evaluating the length of the EHRs, as the Spanish EHRs are significantly longer. While the Spanish records convey 984 words on average, the Swedish only have 74 words. The standard deviation is also more prominent in proportion, with 491 for the Spanish and 77 for the Swedish (note that the standard deviation is higher than the mean). Although the records from both datasets are Discharge Summaries, it seems that not all the Swedish records are complete summaries, but instead a summary or even one-sentence synopsis of the patient's outcome.

3 Methodological Approach

Focusing the attention on the methodology, in Amin et al. (2019) the authors demonstrate the effectiveness of transfer learning with pre-trained language representation model BERT without attention for the multi-label classification of German non-technical summaries (NTSs) of animal experiments. In e-Health 2020 the authors of López-Garcıa et al. (2020) tackled the task as a multi-label classification problem using BERT model Devlin et al. (2019) for the automatic clinical coding of medical cases in Spanish. NLU results crucial to this task and Transformers-based Language Models (LM) are, doubtlessly, the key strength of most recent approaches such as multi-label biomedical text classification Gu et al. (2020). All this and the inherent challenges related to our work (e.g. the ability to distinguish concepts leveraging semantically related diseases) **motivated** us towards BERT-based approaches. Another fact in favour to this choice rests on the ability to the transfer learning between the two languages and, if possible, get benefits from one Language Model to the other. That is, the resources from one language can boost the LM of the other one, while the system remains decoupled from the data.

In order to tackle the multi-label text classification task, we applied a model with a Transformer-based architecture. The problem to solve is the mapping between the input of the EHRs (the raw text, X) and a subset of ICDs from the entire label set, $\mathcal{C}$, where $|\mathcal{C}|$ is the number of codes. The Deep Learning model is trained for the downstream task with pairs of input and output (i.e., EHR texts and ICD codes). The Transformer-based neural network model is trained with instances comprising pairs of input (EHR text) and output (ICD codes). The j-th instance is described formally as $(X_j, \mathbf{c}^j) \subseteq \Sigma^* \times \{0,1\}^{|\mathcal{C}|}$. The input-output pair is as follows: X_j is the string of any length (comprised by tokens from the vocabulary Σ), i.e. the EHR. $\mathbf{c}^j$ is a presence-bits array. $\mathbf{c}^j_i$ encodes the absence or presence of each code $C_i \in \mathcal{C}$ linked to the instance X_j, i.e. the ICDs assigned to the EHR.

From the input text, X_j, fed to the Language Model part of the model, a hidden document representation is obtained. The importance of this rests in that our multi-label classifier is built on top of a BERT model (see Section 3.1). The LM is the core of the Transformer-based NLP models. The principal contribution of this work is the use of the hidden representation to compute attention weights that are label-specific for each input token. After computing the attention, the final output (label predictions) is computed with a fully connected layer that is fed with another document representation got from the label-specific attention layers. To support the reproducible research, we release the code of the per-label attention mechanism with this article.

3.1 Baseline: BERT to Boost LM

The Language Models based on Transformers, specifically BERT models Devlin et al. (2019), have been acknowledged due to their ability to generate contextual representations. In this work, we have to differentiate between very similar diagnoses (all from the gastrointestinal service), which motivated the chosen BERT model as the LM part of our multi-label text classification system to generate the representation of the EHRs. A BERT model is also suitable because of its built-in self-attention function, which can connect different locations of a single input sequence to one another. We also turned to BERT because it has been shown to expand Recurrent Neural Networks' ability to model dependencies to long-distance patterns Hochreiter and Schmidhuber (1997).

In an attempt to encompass Spanish and Swedish, EHRs were represented with shared LMs. The transfer learning approach of sharing the LM poses two advantages. On the one hand, it alleviates the training process for each language since just the task-dependent module (i.e. ICD multi-

label classification) has to be trained. On the other hand, this bypasses the lack of in-domain data for languages other than English. Indeed, the multi-lingual LM, with English, leverages other languages such as Spanish and Swedish in a synergistic effect since cross-language regularities are captured Pires et al. (2019).

The LM part is the core of the BERT models, but coupling different heads on top of the LM is what concedes the ability to tackle numerous downstream tasks, as multi-label classification. Since there are many parameters to describe both the LM and the head for the downstream task, training a BERT model is challenging. The LM module contains the broad majority of the parameters that must be inferred during the training stage. The ICD multi-label classification head built for this study, for example, accounts for less than 1% of the total model parameters (even though using the smallest variant of BERT, which has $110M$ of parameters). With this in mind, we opted to train the multi-label heads from scratch while fine-tuning the LMs instead of training the LMs from scratch.

Because of memory and computational limitations, we used the $\text{BERT}_{\textbf{BASE}}$ as the baseline BERT model (our GPUs are limited to 8GB of DRAM memory). The $\text{BERT}_{\textbf{BASE}}$ model comprises 12 Transformers blocks, 12 self-attention heads, and an internal embedding layer size (d) of 768, totaling $110M$ parameters. The pre-trained $\text{BERT}_{\textbf{BASE}}$ Multilingual model was used. The downstream tasks' attention and output layers are connected to the output of LM, the hidden document representation, ($\mathbf{H}$), of the EHR.

3.2 Contribution: Per-ICD Attention Head

Having opted for the multilingual BERT to cope with the LM, next we proposed to improve the task-dependant head. The aim was to leverage ICD-dependant attention mechanisms in an attempt to enhance the model with added NLU capability when it comes to distinguishing ICDs within the same hospital-service (Digestive in our case).

Our multi-label classification head incorporates a per-label (per-ICD) attention mechanism. The model can classify the EHRs with respect to the ICD labels that are present through the text while also calculating the importance that each input token (word) has in relation to each of the ICDs.

Here, N is the number of tokens of the EHR (length) and d is the BERT hidden layer dimension (i.e., the representation of documents, being $d = 768$ for $\text{BERT}_{\textbf{BASE}}$ models). Then, rather than perform the pool operation (across the document length, N), as in the original BERT Devlin et al. (2019) for classification, our head uses a per-ICD attention mechanism. The per-ICD attention mechanism allows the classifier to discover the correct relationships between the input tokens and each label.

For each ICD label, C_i, the attention vector $\boldsymbol{\alpha}_{C_i} \in \mathbb{R}^{|\mathcal{C}|\times N}$ is computed from the learnable vector parameter $\mathbf{u}_{C_i} \in \mathbb{R}^d$, following (1), where C is the full set of ICD labels.

$$\boldsymbol{\alpha}_{C_i} = Softmax(\mathbf{H}^T \mathbf{u}_{C_i}) \tag{1}$$

The attention scores must be computed as a probability distribution, representing the importance between each token and ICD label pair, and to that end, the model leverages the Softmax function. The matrix multiplication between $\boldsymbol{\alpha}$ and $\mathbf{H}$ is calculated to get an ICD representation for each class from the attention weights. In the end, the maximum through the labels' dimension is taken, obtaining the document representation on the final layer ($\mathbf{v} \in \mathbb{R}^d$), which combines the per-ICD attention representation.

The final layer of the head for multi-label classification is a regular one that allows getting the probabilities for each ICD label. It is a linear layer that takes the document representation ($\mathbf{v}$) as input, which takes into account the attention weights for each input token and label pair. After that, a Sigmoid function is applied to get the actual probabilities of each ICD, as in (2).

$$\hat{\mathbf{y}}_\mathbf{i} = \sigma(\mathbf{W}_\mathbf{i} v_i + b_i) \tag{2}$$

The probability of each ICD class ($C_i \in C$) being on the given input text is $\hat{y}_i$. The parameters of the final layer are the weights matrix ($\mathbf{W}$) and bias ($\mathbf{b}$). Regarding the training of the model, it is carried out by minimising the loss function, precisely, the Binary Cross-Entropy (BCE) loss, as in (3). On this equation, the $\hat{\mathbf{y}}$ is the output of the previous final layer, and $\mathbf{y}$ is the vector that encloses the ICD codes present on the EHR (i.e., the appearance or lack of ICD codes). Figure 1 shows an architectural outline of the system.

$$BCE(\hat{\mathbf{y}}, \mathbf{y}) = -W[\mathbf{y}\log(\hat{\mathbf{y}}) + (1-\mathbf{y})\log(1-\hat{\mathbf{y}})] \tag{3}$$

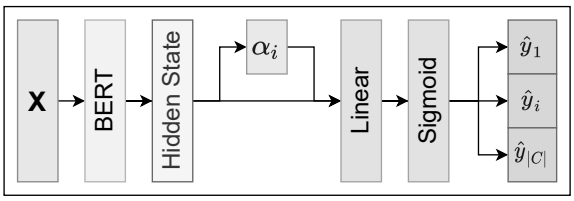


Figure 1: Architectural outline of the developed per-ICD BERT model

4 Experimental Framework

We propose the following experimental setup to evaluate our BERT model's performance with per-ICD attention compared to the benchmark (standard BERT model) on the multi-label ICD classification downstream task. The experimental setup comprises the two minority languages (in terms of in-domain clinical data available), Spanish and Swedish, and a gastrointestinal label set of 157 labels. Each experiment is carried out twice, with the same experimental and training parameters, one with the regular multi-label classification head (as the baseline) and the other with our head with per-ICD attention. We show the results from the experimental results in Table 1 and Figures 2 and 3, for Spanish and Swedish, respectively.

The model with our per-ICD attention head obtains better results in both languages. It is important to note that the results improve considerably even in this context with a considerably large label set (157 labels). This finding is consistent with the following hypothesis: many terms can be important when dealing with a wide number of ICD codes at once and long EHRs, but probably only a few of them are relevant for each ICD code individually.

Multi-label ICD classification is often assessed by means of the Area Under the ROC Curve (AUC) micro averaging the metric for all the ICDs involved (denoted as AUCm in Table 1). For Spanish, the per-ICD model surpasses the base BERT model by 9.16 points, also improves slightly for the Swedish, with an improvement of around 1 point. In Figures 2 and 3 we show the confusion matrices for each experiment. Each confusion matrix is the average of the matrices of each ICD class, and we have computed two versions, i.e. one with arithmetic averaging (aka samples average) and the other with weighted averaging. In both, the darker the colour, the higher the metric, always in the range $[0 - 100]$. The weighted averaged matrices are computed considering the support (relative frequency) of each ICD class. Note that the TPR (True Positive Rate) and FNR (False Negative Rate) shown in Table 1 are also the arithmetic average of each corresponding model, but the CM show also the FPR (False Positive Rate) and TNR (True Negative Rate), while the weighted average of each metric. Regarding the per-class performance, there is a positive association with the support; the more frequent the label, the better are the results.

If we analyse the matrices, it can be observed that the source of improvement of the per-ICD model can be broken down; while the True Negatives stay close (as with a large label set, the majority of classes are negative), the True Positives improves considerably, with an increment of almost 100%. In the same way, the False Negatives decrease by around 20%. Although the Swedish results are in general weaker, this behaviour is appreciated similarly for both languages. Therefore, given the results, it seems that our per-ICD attention head is able to improve the Precision of the regular BERT models for ICD multi-label classification with large label sets. Nevertheless, the per-ICD model outperforms regular BERT in terms of performance, but also in interpretability capabilities, as it has the ability to export the attention weights, allowing its visualisation.

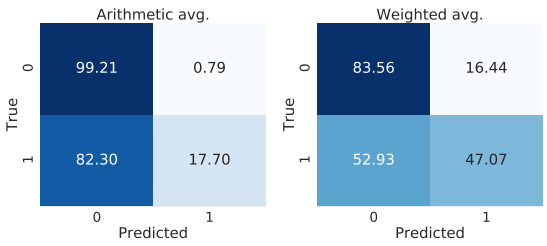
L	Model	AUCm	TPR	FNR
SP	baseline	58.16	17.70	99.21
	per-ICD	**67.32**	**34.92**	**99.38**
SW	baseline	54.92	15.49	92.24
	per-ICD	**55.96**	**27.91**	**82.45**

Table 1: Comparison of results on the Spanish (SP) and Swedish (SW) datasets ("L" stands for "Language") obtained with the baseline BERT and BERT enhanced with per-ICD attention head. TPR is the True Positive Rate and FNR the False Negative Rate.

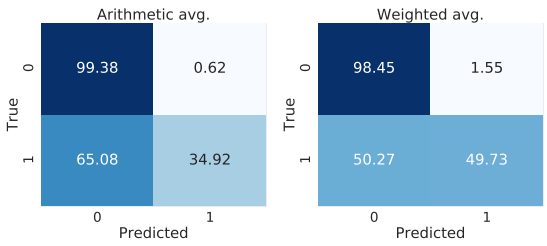
5 Discussion

Within the clinical text mining field, the main weakness tends to be the availability of corpora due to the natural patient's confidentiality policy Cohen and Demner-Fushman (2014). As a result, for the research to make progress, the so important comparability might get compromised. By contrast, through this work the authors are glad to make available their own implementation of the per-ICD attention approach[3] as a secondary contribution of

[3]To get the source code of the implementation, simply e-mail the first author.

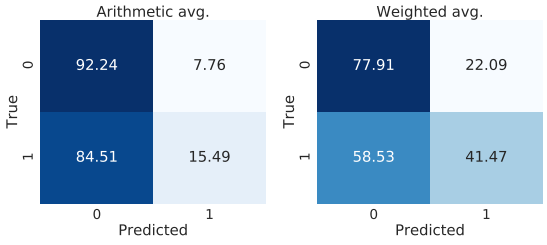


(a) Baseline BERT

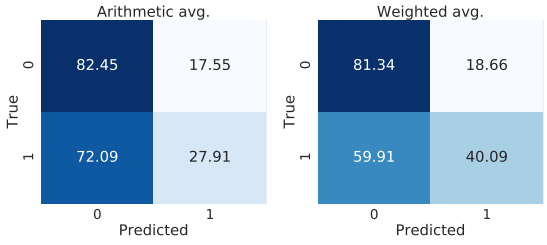


(b) Per-ICD BERT

Figure 2: Average heatmaps of Spanish models



(a) Baseline BERT



(b) Per-ICD BERT

Figure 3: Average heatmaps of Swedish models

this paper.

Another aspect related with the corpus is the complexity and length of the input EHR. The average length of the input of the works mentioned Névéol et al. (2018b); Cappellato et al. (2019) are variable from a few words in the case of Italian, Hungarian and French to 350 words for the documents written in Spanish Miranda-Escalada et al. (2020). By contrast, in our paper we deal with documents in Spanish and Swedish with an average length of 800 (exceeding the aforementioned ones) and 70 respectively.

According to these results, the per-label attention mechanism improves Precision. While more performance is still necessary for a fully automated system, the results suggest that it is suitable for multi-label classification of EHRs according to the ICD standard, specifically applying it as a clinical DSS, as the per-ICD attention can aid the expert in the EHR codification process.

6 Conclusions

We have dealt with the codification of EHRs of the gastrointestinal service for Swedish and Spanish hospitals. We have developed a BERT model for multi-label classification incorporating a per-label attention mechanism.

The results obtained have revealed that the proposed model outperforms the regular BERT. We have proved this fact for two languages with minority resources in clinical NLP, showing that solutions of language independent nature work. Moreover our proposal generates an interpretable output that helps to know the relevance of the tokens with respect to each ICD assigned to the EHR. To sum up, the per-label attention mechanism differentiates semantically ICDs that are related and aids to explain the core of each label. Future work may include testing BERT models trained for the specific languages, as the BETO model Cañete et al. (2020) for Spanish.

Acknowledgments

This work was partially funded by the Spanish Ministry of Science and Innovation (DOTT-HEALTH/PAT-MED PID2019-106942RB-C31), European Commission (FEDER) and the Basque Government (IXA IT-1343-19, Predoctoral Grant PRE-2019-1-0158) and by the ClinCode project, project number 318098, from the Norwegian Research Council. This research has been approved by the Regional Ethical Review Board in Stockholm under permission no. 2007/1625-31/5. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

References

Saadullah Amin, Günter Neumann, Katherine Dunfield, Anna Vechkaeva, Kathryn Annette Chapman, and Morgan Kelly Wixted. 2019. MLT-DFKI at CLEF eHealth 2019: Multi-label Classification of ICD-10 Codes with BERT. In *CLEF (Working Notes)*, pages 1–15.

Alberto Blanco, Alicia Pérez, and Arantza Casillas. 2020. Extreme multi-label icd classification: Sensitivity to hospital service and time. *IEEE Access*, 8:183534–183545.

Linda Cappellato, Nicola Ferro, David E. Losada, and Henning Müller, editors. 2019. *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019*, volume 2380 of *CEUR Workshop Proceedings*. CEUR-WS.org.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish Pre-Trained BERT Model and Evaluation Data. In *Practical ML for Developing Countries at ICLR 2020*.

Kevin Bretonnel Cohen and Dina Demner-Fushman. 2014. *Biomedical natural language processing*, volume 11. John Benjamins Publishing Company.

Hercules Dalianis. 2018. *Clinical text mining: Secondary use of electronic patient records*. Springer Nature.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv*, pages 4171–4186.

Jiachen Du, Lin Gui, Ruifeng Xu, and Yulan He. 2017. A convolutional attention model for text classification. In *National CCF conference on natural language processing and Chinese computing*, pages 183–195. Springer.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing. *arXiv preprint arXiv:2007.15779*.

Aron Henriksson, Martin Hassel, and Maria Kvist. 2011. Diagnosis code assignment support using random indexing of patient records – a qualitative feasibility study. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 348–352. Springer.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Anders Jacobsson and Lisbeth Serdén. 2013. Kodningskvalitet i patientregistret, (In Swedish).

Rita Kukafka, Michael E Bales, Ann Burkhardt, and Carol Friedman. 2006. Human and automated coding of rehabilitation discharge summaries according to the International Classification of Functioning, Disability, and Health. *Journal of the American Medical Informatics Association*, 13(5):508–515.

Guillermo López-Garcıa, José M Jerez, and Francisco J Veredas. 2020. ICB-UMA at CLEF e-Health 2020 Task 1: Automatic ICD-10 coding in Spanish with BERT. In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020.*

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with Words at the National Library of Sweden – Making a Swedish BERT. *arXiv preprint arXiv:2007.01658*.

Antonio Miranda-Escalada, Aitor Gonzalez-Agirre, Jordi Armengol-Estapé, and Martin Krallinger. 2020. Overview of automatic clinical coding: annotations, guidelines, and solutions for non-english clinical cases at codiesp track of CLEF eHealth 2020. In *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings*.

Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum. 2018a. Clinical natural language processing in languages other than English: opportunities and challenges. *Journal of biomedical semantics*, 9(1):1–13.

Aurélie Névéol, Aude Robert, Francesco Grippo, Claire Morgand, Chiara Orsi, Laszlo Pelikan, Lionel Ramadier, Grégoire Rey, and Pierre Zweigenbaum. 2018b. CLEF eHealth 2018 Multilingual Information Extraction Task Overview: ICD10 Coding of Death Certificates in French, Hungarian and Italian. In *CLEF (Working Notes)*, pages 1–18.

Maite Oronoz, Koldo Gojenola, Alicia Pérez, Arantza Díaz de Ilarraza, and Arantza Casillas. 2015. On the creation of a clinical gold standard corpus in Spanish: Mining adverse drug reactions. *Journal of biomedical informatics*, 56:318–332.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Sonja Remmer, Anastastios Lamproudis, and Hercules Dalianis. 2021. Multi-label Diagnosis Classification of Swedish Discharge Summaries – ICD-10 Code Assignment Using KB-BERT. *In the Proceedings of Recent Advances in Natural Language Processing, RANLP 2021, Varna, Bulgaria*.

WHO. 2016. International Classification of Diseases (ICD). Accessed 2021-04-14.

Zachariah Zhang, Jingshu Liu, and Narges Razavian. 2020. BERT-XML: Large Scale Automated ICD Coding Using BERT Pretraining. *arXiv preprint arXiv:2006.03685*.