

LASIGE-BioTM at MESINESP2: entity linking with semantic similarity and extreme multi-label classification on Spanish biomedical documents

Pedro Ruas¹, Vitor D. T. Andrade¹ and Francisco M. Couto¹

¹LASIGE, Faculdade de Ciências, Universidade de Lisboa, Lisbon 1749-016, Portugal

Abstract

Our team, LASIGE_BioTM, participated in the three sub-tracks of MESINESP2: (1) scientific literature, (2) clinical trials, and (3) patents. Our system comprises two modules: entity linking and extreme multi-label classification. The first module uses the entities recognized in text and then applies a graph-based entity linking model to link them to the DeCS vocabulary. In the end, it applies a semantic similarity-based filter to determine the most relevant entities in each document, which are then fed to the second module. The second module consists of an adapted version of the X-Transformer algorithm, and is responsible for associating each document with the top-20 relevant DeCS codes, which can be viewed as an extreme multi-label classification algorithm. The obtained results (micro F1-scores) were 0.2007, 0.0686, and 0.0314 for sub-tracks 1, 2, and 3, respectively. These represent low values when compared to other participants, mainly because of the lack of time our team had available to train the models. All of the used software is available in an open access repository.

Keywords

Named Entity Recognition, Named Entity Linking, Extreme Multi-Label Classification, Multilingual, Text Mining

1. Introduction

Automatic semantic indexing is essential to organise the growing text data that is available, which is particularly critical in scientific domains, including the biomedical one, where most of the findings are available in the text format. We can view this task as an extreme multi-label classification (XMC) problem, in which the goal is to tag a given data point with a subset of relevant labels from an extremely large label list. Therefore, the data points are the text documents to classify, and the label list provided by a knowledge base, such as an ontology. Most of the proposed XMC approaches focus on datasets including Wikipedia articles or on datasets with commercial application (e.g. dynamic search advertising) and less attention is devoted to the biomedical domain. Additionally, multilingual approaches focusing on other languages besides English are also scarce, such is the case of Spanish.

In this sense, initiatives such as BioASQ [1] are necessary to stimulate the development of biomedical, multilingual-focused approaches. In particular, the Medical Semantic Indexing


CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania

✉ psruas@fc.ul.pt (P. Ruas); fc49005@alunos.fc.ul.pt (V. D. T. Andrade); fcouto@di.fc.ul.pt (F. M. Couto)

🆔 0000-0002-1293-4199 (P. Ruas); 0000-0003-0627-1496 (F. M. Couto)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

In Spanish (MESINESP) task was first introduced in the BioASQ 2020 challenge and the goal was to perform semantic indexing of Spanish health-related documents, like scientific articles, clinical trials, and healthcare project summaries, with terms from the Spanish version of the *Descriptores en Ciencias de la Salud* (DeCS). The second edition, the MESINESP2 shared-task [2] was extended and included the following sub-tracks: **MESINESP-L** – Scientific Literature: Automatic indexing with DeCS terms of Spanish abstracts from two databases, IBECs and LILACS; **MESINESP-T** - Clinical Trials: Automatic indexing with DeCS terms of Spanish clinical trials from REEC (Registro Español de Estudios Clínicos); **MESINESP-P** – Patents: Automatic indexing with DeCS terms Spanish patents extracted from Google Patents.

In the past, named entities have been considered important features that aid the classification of texts. For instance, Gui et al [3] proposed a hierarchical text classification method that leverages named entities as features, and, according to the conclusions of the referred study, the features are responsible for the improvement of the method's performance. More recently, Anelic and co-workers [4] have argued that named entities do not improve the performance of text classification, and can even decrease it. However, none of these works attempted to normalise the recognised entities to concepts belonging to structured vocabularies, the approaches only used the surface form of the entities instead of the designations for the associated concepts. Besides, not every entity recognised in a given document has the same importance, i.e., some entities may not be related with the main topic of the document, which can be particularly true in documents containing a large number of different entities. Therefore, we explored the hypothesis that linking the recognised entities to concepts of a structured vocabulary and selecting only the most relevant entities to feed the text classification algorithm improve its performance.

After participating in the first edition [5], this paper describes the participation of our team, LASIGE_BioTM, in the sub-tracks of MESINESP2. We developed a pipeline based on two modules: the first one performs entity linking, by mapping the recognised entities in text to terms of the DeCS vocabulary and then applying a semantic similarity-based filter to obtain the most relevant entities in each document; the second module is based on the X-Transformer algorithm [6], and is responsible to classify each document with the most relevant DeCS terms. The software used in the experiments is available on: <https://github.com/lasigeBioTM/MESINESP2>.

1.1. Related work

1.1.1. Entity Linking

The extraction of entities is carried out through the text mining process. This process can be executed by different approaches such as: rule-based methods, machine learning and deep learning.

Rule-based methods include a set of terms, regular expressions or sentence constructions defined by experts [7]. Rule-based methods also include dictionary approaches, in which a given text is matched against a lexicon using string matching [8].

Machine learning methods in text mining are trained on training and validation datasets to make predictions on a test dataset [7]. Deep learning is a subset of machine learning that consists of artificial neural networks that include multiple hidden layers between input and

output. An artificial neural network is composed of nodes, processing units with a similar function to the neurons in the brain. The input for the nodes in text mining applications are word embeddings, which are vector representations of words. According to the way the nodes are organised, deep neural networks can be classified as Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), among others.

Usually text mining approaches include the tasks of Named Entity Recognition (NER) and Named Entity Linking (NEL). NER corresponds to the recognition of entities mentioned in the text and NEL to the linking of the recognised entities to concepts of a given knowledge base.

For the NER task, state-of-art approaches usually have a bidirectional long short-term memory - conditional random fields (BiLSTM-CRF) architecture. However, approaches that use pre-trained language models have recently emerged and showed promising results. One of the pre-trained language models that has been highlighted in the tasks of text mining is BERT [9], which is organized in a multilayer bidirectional transformer encoder. This architecture is based on an attention mechanism and allows the finding of dependencies between input and output [10]. Several variations of the original BERT model are trained in different scientific corpora, such as BioBERT [11] which was trained in PubMed and PMC articles and SciBERT [12], that was trained in Semantic Scholar articles. After the pre-training, these variations and the original BERT model can also be fine-tuned for NEL tasks[13].

In addition to the pre-trained language models, NEL state-of-the-art approaches in the biomedical domain also include graph-based models. Usually, these build a disambiguation graph composed by candidates for entity mentions and then ranked according to their relevance and coherence in the graph. Models that use the Personalized PageRank algorithm to determine the relevance of the candidates in the graph have been proposed, such as Pershina et al. [14].

1.1.2. Semantic similarity

The calculation of the relevance of the candidates in a graph normally requires a similarity measure to compare its nodes, as was proposed by Lamurias et al. [15]. A semantic similarity measure is a metric to compare the similarity between sets of text based on their implicit and explicit semantics. In the present work, we measured the semantic similarity between each entity and the remaining entities of a given document through Resnik's metric [16]. This metric is based on the extrinsic information content (IC) of the most informative common ancestor (MICA) of two given concepts [17] and is defined as:

$$SSM_{resnik}(e_1, e_2) = IC_{shared}(e_1, e_2)$$

Being e_1 and e_2 the entity 1 and the entity 2, respectively.

1.1.3. Extreme multi-label classification and biomedical semantic indexing

Chang et al. [18] divided the approaches to the XMC task in four categories: one-vs-all, partitioning methods, embedding-based, and deep-learning-based.

The Parabel algorithm [19] follows a one-vs-all approach because it learns a separate classifier for each label in the label list. It also applies a tree-based method, since it learns a balanced hierarchy over labels, which helps identifying the most similar labels with respect to a given

Table 1

Number of documents in each corpus.

Corpus	Train	Dev	Test
Scientific literature (L)	249,474	1,065	10,179 (500 gold standard)
Clinical Trials (T)	3,560	147	8,919 (250 gold standard)
Patents (P)	—	115	68,404 (150 gold standard)

label, i.e. those that are present in the same leaves. It performs sub-sampling of data points by restricting a given label’s negative training examples to those examples that are annotated with similar or confusing labels, which decreases training and prediction times from linear to logarithmic. The approach then applies a hierarchical multi-label model, which is a generalisation of the multi-class hierarchical softmax model. Each classifier learns a joint probability distribution over the possible labels that is based on data point features and on the label hierarchy. Parabel was applied to Dynamic Search Advertising, which aims to predict the subset of search engine queries that will lead to a click on a given ad page.

The current state-of-the-art in XMC consists of approaches that leverage pre-trained deep language models. The first approach of this type was X-BERT (*BERT for eXtreme Multi-label Text Classification*) [18], later renamed to X-Transformer [6], which fine-tunes BERT, RoBERTa, and XLNet for the XMC task. The main challenges of applying Transformer to the XMC problem are the extremely large set of possible labels and the label sparsity, which arises from the fact that too few labels are associated with a large number of training instances. The model includes three components: a semantic label indexer, a deep neural matcher, and a ranker. The authors applied the developed algorithm to four datasets, Eurlex-4K, Wiki10-28K, AmazonCat-13K and Wiki-500K, obtaining the following precision@1 values: 86.00%, 85.75 %, 95.17 %, 67.87 %.

1.1.4. MESINESP1

With respect to the MESINESP task, six teams have participated in the first edition, including our team, which have developed a pipeline [5] based on the X-Transformer algorithm [6] and the MER tool [20] for the named entity recognition and linking step. The approach with best performance was based on AttentionXML with multilingual-BERT [21], which achieved a micro F-measure value of 0.4254, whereas our approach achieved a micro F1-score of 0.2507.

Besides DeCS and MeSH vocabularies, there are also related works that focus on the classification or coding of clinical content with codes belonging to other vocabularies, in particular the International Classification of Diseases (ICD) terminology [22, 23, 24, 25].

2. Methodology

2.1. Data description

The target label list consisted of 34,046 codes belonging to the DeCS vocabulary¹ (2020 edition), complemented with additional COVID-related descriptors added by the organisation. Both corpora (JSON files) and the DeCS vocabulary (TSV file) were provided by the organisation and downloaded from the following link: <https://zenodo.org/record/4634129#.YHcShxIo9an>.

2.2. Entity Linking

Our approach consisted in using the recognised entities from the documents of each subtrack that were provided in the folder “Additional Data”. The entities of these files were then further linked to the respective DeCS codes through an entity linking model. This model searches for the ten best candidates of DeCS through string matching and then develops a disambiguation graph with those candidates. The Personalized PageRank algorithm is applied to the disambiguation graph and estimates the coherence of each node, i.e. candidate, to the graph. The coherence is associated with the node degree, meaning that nodes linked to a high number of other candidate nodes are probable candidates for their respective entities compared with more isolated nodes. Besides coherence, the IC of the DeCS code associated with the nodes is used for ranking: nodes associated with DeCS codes with higher IC receive higher ranking scores. IC corresponds to the presence of an entity in a corpus, if an entity is not common in a corpus its IC will be high. The higher the IC of a candidate is, the better ranking that candidate will have in the graphic. After ranking all the candidates, the PPR selects the candidate with better ranking to map each entity. At the end, all entities in a given document are linked to their respective DeCS concepts.

To explore the guiding hypothesis of this work, we filtered the number of entities to include each document by applying a semantic similarity-based filter, more concretely, by selecting the entities for which there were other similar entities recognised in the same document.

After this step, the average of the several semantic similarity values obtained for an entity corresponded to the final score of that entity. The entities were then sorted by their score. At the end, we explored two values for the semantic similarity-based filter: 1.0 and 0.25. Considering the filter 0.25, we only included the top 25% entities according to their score, and for the filter 1.0, we included all the entities in the document. This way, we could determine the impact of choosing the most relevant entities in the performance of the classifier algorithm.

2.3. Extreme Multi-Label Classification

We approached the sub-tracks as an Extreme Multi-Label Classification (XMC) problem. Our starting point was a pipeline based on the X-Transformer algorithm [6] that was adapted to the biomedical domain by our group in the context of past competitions, such as BioASQ [5] and CANTEMIST [26]. The pipeline was further adapted to the present competition, and includes the following modules: entity linking (subsection 2.2), preprocessing, semantic label indexer, deep neural matcher, and ranker. The main modifications were made in the entity linking and

¹<http://red.bvsalud.org/decs/en/>

preprocessing modules. The complete description of the entity linking component is available in the previous subsection 2.2.

The preprocessing module imports the retrieved dataset JSON files (train, dev, and text subsets) and the DeCS TSV file, the JSON files with the output from the entity linking (subsection 2.2), and, for each dataset, it generates several files:

1. vocabulary file ("label_vocab.txt"): it includes the internal numerical identifier for each DeCS term. For example, the term "calcimicina" has the internal numerical identifier "0".
2. label correspondence file ("label_correspondence.txt"): it includes the correspondence between the internal numerical identifiers, and the respective DeCS labels and terms. For example, "0" corresponds to "D000001", which corresponds to "calcimicina".
3. subset files ("*subset.txt*", "*subset_raw_text.txt*", "*subset_raw_labels.txt*"): for each subset (train, dev, and test) it is generated the three aforementioned files. The file "*subset.txt*" includes the DeCS labels that are associated with the respective documents, separated by commas, the stemmed texts of documents' titles, and the DeCS terms that were extracted in the documents appended to the end of the stemmed titles. The file "*subset_raw_text.txt*" includes only the stemmed titles, and the file "*subset_raw_labels.txt*" only the DeCS terms relative to the labels associated with the documents.

We only considered the titles of the documents based on the results described by Neves et al. [5]: the performance of the models using titles is similar to that of models using abstracts, so it is more efficient to use titles since they have less text. The limited time that we had to train models also influenced our decision to only use the titles, since the required time is lower. The titles were stemmed using the Snowball Stemmer implementation for Spanish text provided by the NLTK package². As the documents belonging to the test sets were unlabeled, we added the placeholder "0" to each document in the "*subset.txt*" files. The module was also modified in order to integrate extracted entities independently of the tool employed.

The X-Transformer algorithm includes three modules: semantic label indexer, deep neural matcher, and ranker. The semantic label indexer first obtain meaningful representations for labels that are based on embeddings of the text descriptions associated with the labels, and on Positive Instance Feature Aggregation (PIFA), which is a type of label embeddings based on the TF-IDF features that are relevant instances for the labels. Then, it applies k-means clustering in order to generate label clusters according to the semantic representations described before. The deep neural matcher performs fine-tuning of BERT to encode an instance embedding, which is then used to find the most relevant clusters for the instance. At the end of this step, only a small subset of clusters are considered for the next step, which is performed by the ranker. The ranker determines the relevance of the labels in the chosen clusters to the instance, which is substantially more efficient than performing the ranking of all the initial labels. For a more complete description of the X-Transformer algorithm please refer to the original publication by Chang et al. [6].

The models developed for the different sub-tracks are shown in Table 2. We explored the fine-tuning of different deep neural matchers. The BERT Base Multilingual Cased model was trained on the Wikipedia dumps of the top 104 largest languages in Wikipedia and has the following

²<https://www.nltk.org/>

Table 2

Models used for the three sub-tracks, with the respective target datasets, thresholds (top entities to consider according to their relevance), and deep neural matcher.

Model	Target dataset	Threshold	Deep neural matcher
LASIGE_BioTM-1	L	1.0	CANTEMIST
LASIGE_BioTM-2		0.25	
LASIGE_BioTM-3	T	1.0	BERT Multilingual Base Cased
LASIGE_BioTM-4		0.25	
LASIGE_BioTM-5	P	1.0	BERT Multilingual Base Cased

characteristics: 12-layer, 768-hidden, 12-heads, 110M parameters. The X-Transformer algorithm uses the Pytorch implementation from HuggingFace Transformers [27]. The CANTEMIST model corresponds to the Model 7 described by Ruas et al. [26]. It is also based on the the BERT Base Multilingual Cased model and was first fine-tuned on 318,658 Spanish biomedical articles from the IBECS, LILACS and PubMed databases, jointly with extracted entities in the context of the participation in the first edition of MESINESP [5].

2.4. Training approach

We explored several training approaches according to the target corpus:

- L corpus: Fine-tuning of the model CANTEMIST using the provided training dataset of 249,474 documents and the provided test set with 10,179 documents.
- T corpus: Training of the model BERT Multilingual Base Cased using the provided training dataset of 249,474 documents from the L corpus and a generated test set built from the 3560 clinical trials of the training set, the 147 clinical trials of the development set, and the 8919 clinical trials of the test set (total of 12,627 documents).
- P corpus: Training of the model BERT Multilingual Base Cased using the provided training dataset of 249,474 documents from the L corpus and a generated test set built from the 115 patents of the development set and the 68,404 patents from the test set.

The training of the deep neural matcher is the limiting step of the algorithm in terms of time. Each model was trained during a single epoch then evaluated on the respective test set. The training and evaluation time was approx. 2 days for each model using a single NVIDIA Tesla P4 GPU. The values for the hyper-parameters are the following: `depth=6`, `train_batch_size=4`, `eval_batch_size=4`, `learning_rate=0.00005`, `warmup_rate=0.1`.

3. Results and discussion

The results obtained for each sub-track are shown on Table 3. The official evaluation metric of the competition was the micro F1-score (MiF). Our best models achieved a MiF of 0.2007, 0.0686, and 0.0314 in the sub-tracks L, T, and P, respectively. These results are low when compared to

Table 3

Results on test sets for the three sub-tracks. Performance for the baseline models, the best models, and our models are shown according to the metrics: MiF-micro F1-score, MiP-micro precision, MiR-micro recall, MaF-macro F1-score, MaP-macro precision, MaR-macro recall.

Sub-track	Model	MiF	MiP	MiR	MaF	MaP	MaR
L	Baseline	0.2876	0.2335	0.3746	0.3438	0.2335	0.3746
	BERTDeCS version 4	0.4837	0.5077	0.4618	0.3926	0.5237	0.3990
	LASIGE_BioTM-1	0.2007	0.1584	0.2738	0.0941	0.1016	0.1232
	LASIGE_BioTM-2	0.1886	0.1489	0.2573	0.0920	0.0950	0.1219
T	Baseline	0.1288	0.0781	0.3678	0.2403	0.0977	0.3619
	BERTDeCS version 2	0.3640	0.3666	0.3614	0.3102	0.4177	0.3391
	LASIGE_BioTM-3	0.0679	0.0575	0.0828	0.0056	0.0050	0.0136
	LASIGE_BioTM-4	0.0686	0.0581	0.0838	0.0061	0.0054	0.0133
P	Baseline	0.2992	0.4293	0.2296	0.2518	0.5290	0.2497
	BERTDeCS version 2	0.4514	0.4487	0.4541	0.4138	0.5041	0.4271
	LASIGE_BioTM-5	0.0314	0.0239	0.0459	0.0071	0.0060	0.0135

the top results in each sub-track, more concretely, there is a difference of 0.2830, 0.2961, and 0.4200 in terms of MiF in the sub-tracks L, T, and P, respectively.

With respect to the initial hypothesis, the obtained results were mixed. In the sub-track L, the LASIGE_BioTM-1 model, which included all the entities recognised in the documents, obtained slightly better results (0.2007 MiF) compared with LASIGE_BioTM-2 model (0.1886 MiF), which only included 25% of the top relevant entities. However, in the sub-track T, the opposite happened, since LASIGE_BioTM-4 (top 25% entities) obtained marginally better results (0.0686 MiF) than LASIGE_BioTM-3 (0.0679 MiF). Consequently, we cannot confirm the initial hypothesis that feeding only the most relevant entities to the classifier algorithm improves its performance.

Assuming that there were no coding errors that may have undermined the results, there are several possible reasons behind the relatively low performance that our models achieved in the three sub-tracks.

Arguably, the main one is related with the impossibility of carrying out an optimisation of the hyper-parameters of the classifier algorithm, in particular the number of training epochs. Each model was only trained or fine-tuned during one epoch in the respective training dataset, which is not enough to accurately learn relevant features. The limited time we had available made it impossible to extend the training process during more epochs. Additionally, we were not able to train the models in a multi-gpu setting due to unresolved errors, so the duration of each training epoch was approximately two days using a single gpu. Beyond the number of training epochs, the optimization of other hyperparameters such as `train_batch_size`, `eval_batch_size`, and `learning_rate`, would probably lead to a better performance.

With respect to the sub-track 2 and sub-track 3, the developed models were trained on documents belonging to the L corpus (sub-track 1), and not on documents of the respective sub-tracks corpora. The text present in scientific literature has different characteristics compared with the text associated with clinical trials and patents, so the models fine-tuned in a certain

type of text will necessarily have a worse performance when their evaluation occurs over a different type of text. For sub-track 3, there was no training dataset available, but for sub-track 2 probably it would have been better if we had trained models 3 and 4 over the training dataset of the task and not over the training dataset for sub-track 1.

4. Conclusion

Our approach including an entity linking model and the X-Transformer algorithm obtained a micro F1-score of 0.2007, 0.0686, and 0.0314 in sub-tracks 1, 2, and 3, respectively, which is a low performance compared with the top participants, and even with the baseline approaches. In order to improve the performance, we need to perform a careful error-analysis to identify any coding errors that may have undermined the results. Next, we need to spend more time in the training process, more concretely, by training the models during more epochs, to perform hyper-parameter optimisation, to solve the problems associated with multi-gpu training, to explore the use of summarisation tools to feed only the relevant content to the classifier, and to explore less resource-demanding pre-trained models, such as DistilBERT. Besides, we only used the titles of the articles based on previous studies, but in the future we will explore the impact of using more text in the performance of the classification algorithm.

Acknowledgments

This work was supported by FCT through funding of Deep Semantic Tagger (DeST) project (ref. PTDC/CCI-BIO/28685/2017) and LASIGE Research Unit (ref. UIDB/00408/2020 and ref. UIDP/00408/2020); and FCT through funding of PhD Scholarship, ref. 2020.05393.BD.

References

- [1] A. Nentidis, G. Katsimpras, E. Vandorou, A. Krithara, L. Gasco, M. Krallinger, G. . Paliouras, Overview of BioASQ 2021: The ninth BioASQ challenge on Large-Scale Biomedical Semantic Indexing and Question Answering. (2021).
- [2] L. Gasco, A. Nentidis, A. Krithara, D. Estrada-Zavala, , R.-T. Murasaki, E. Primo-Peña, C. Bojo-Canales, G. Paliouras, M. Krallinger, Overview of BioASQ 2021-MESINESP track. Evaluation of advance hierarchical classification techniques for scientific literature, patents and clinical trials. (2021).
- [3] Y. Gui, Z. Gao, R. Li, X. Yang, Hierarchical text classification for news articles based-on named entities, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 7713 LNAI (2012) 318–329. doi:10.1007/978-3-642-35527-1_27.
- [4] S. Andelic, M. Kondic, I. Peric, M. Jovic, A. Kovacevic, Text Classification Based on Named Entities, in: 7th International Conference on Information Society and Technology ICIST 2017, 2017.

- [5] A. Neves, A. Lamurias, F. M. Couto, Extreme Multi-Label Classification applied to the Biomedical and Multilingual Panorama, in: CLEF 2020 Working Notes, 2020. URL: http://ceur-ws.org/Vol-2696/paper_67.pdf.
- [6] W.-C. Chang, H.-F. Yu, K. Zhong, Y. Yang, I. Dhillon, Taming Pretrained Transformers for Extreme Multi-label Text Classification (2020). URL: <https://doi.org/10.1145/3394486.3403368>. doi:10.1145/3394486.3403368. arXiv:1905.02331v4.
- [7] A. Lamurias, F. Couto, Text Mining for Bioinformatics Using Biomedical Literature, 2019, p. 602–611. doi:10.1016/B978-0-12-809633-8.20409-3.
- [8] F. M. Couto, A. Lamurias, Mer: a shell script and annotation server for minimal named entity recognition and linking, *Journal of Cheminformatics* 10 (2018) 58.
- [9] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
- [10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: *Advances in neural information processing systems*, 2017, pp. 5998–6008.
- [11] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, *Bioinformatics* 36 (2020) 1234–1240.
- [12] I. Beltagy, K. Lo, A. Cohan, SciBERT: A pretrained language model for scientific text, in: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3615–3620. URL: <https://www.aclweb.org/anthology/D19-1371>. doi:10.18653/v1/D19-1371.
- [13] Z. Ji, Q. Wei, H. Xu, Bert-based ranking for biomedical entity normalization, *AMIA Summits on Translational Science Proceedings 2020* (2020) 269.
- [14] M. Pershina, Y. He, R. Grishman, Personalized page rank for named entity disambiguation, in: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Denver, Colorado, 2015, pp. 238–243. URL: <https://www.aclweb.org/anthology/N15-1026>. doi:10.3115/v1/N15-1026.
- [15] A. Lamurias, P. Ruas, F. M. Couto, PPR-SSM: Personalized PageRank and semantic similarity measures for entity linking, *BMC Bioinformatics* 20 (2019) 1–12. doi:10.1186/s12859-019-3157-y.
- [16] P. Resnik, Using information content to evaluate semantic similarity in a taxonomy, arXiv preprint [cmp-lg/9511007](https://arxiv.org/abs/1905.02331) (1995).
- [17] F. Couto, A. Lamurias, Semantic similarity definition, *Encyclopedia of bioinformatics and computational biology* 1 (2019).
- [18] W.-C. Chang, H.-F. Yu, K. Zhong, Y. Yang, I. Dhillon, X-BERT: eXtreme Multi-label Text Classification with using Bidirectional Encoder Representations from Transformers (2019) 1–12. URL: <http://arxiv.org/abs/1905.02331>. arXiv:1905.02331.
- [19] Y. Prabhu, A. Kag, S. Harsola, R. Agrawal, M. Varma, Parabel: Partitioned label trees for extreme classification with application to dynamic search advertising, in: *Proceedings of the World Wide Web Conference, WWW 2018, ACM, New York, NY, USA, April 23-27, 2018, Lyon, France, 2018*, pp. 993–1002. doi:10.1145/3178876.3185998.

- [20] F. M. Couto, A. Lamurias, MER: a shell script and annotation server for minimal named entity recognition and linking, *Journal of Cheminformatics* 10 (2018) 58. URL: <https://jcheminf.biomedcentral.com/articles/10.1186/s13321-018-0312-9>. doi:10.1186/s13321-018-0312-9.
- [21] R. You, Z. Zhang, Z. Wang, S. Dai, H. Mamitsuka, S. Zhu, AttentionXML: Label Tree-based Attention-Aware Deep Model for High-Performance Extreme Multi-Label Text Classification, in: *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, Vancouver, Canada, 2019, pp. 1–11. arXiv:1811.01727.
- [22] P. Xie, H. Shi, M. Zhang, E. P. Xing, A Neural Architecture for Automated ICD Coding, in: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, Association for Computational Linguistics, 2018, pp. 1066–1076.
- [23] H. Shi, P. Xie, Z. Hu, M. Zhang, E. P. Xing, Towards Automated ICD Coding Using Deep Learning, Technical Report, 2017. arXiv:1711.04075v3.
- [24] S. Silvestri, F. Gargiulo, M. Ciampi, G. De Pietro, Exploit Multilingual Language Model at Scale for ICD-10 Clinical Text Classification, *Proceedings - IEEE Symposium on Computers and Communications 2020-July (2020)*. doi:10.1109/ISCC50000.2020.9219640.
- [25] C. Sen, B. Ye, J. Aslam, A. Tahmasebi, From Extreme Multi-label to Multi-class: A Hierarchical Approach for Automated ICD-10 Coding Using Phrase-level Attention (2021). URL: <http://arxiv.org/abs/2102.09136>. arXiv:2102.09136.
- [26] P. Ruas, A. Neves, V. D. Andrade, F. M. Couto, Lasigebiotm at cantemist: Named entity recognition and normalization of tumour morphology entities and clinical coding of Spanish health-related documents, in: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)*, 2020, pp. 422–437. URL: http://ceur-ws.org/Vol-2664/cantemist_paper11.pdf.
- [27] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Transformers: State-of-the-art natural language processing, in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: <https://www.aclweb.org/anthology/2020.emnlp-demos.6>.