Computer Methods and Programs in Biomedicine 188 (2020) 105264

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Boosting ICD multi-label classification of health records with contextual embeddings and label-granularity

Alberto Blanco*, Olatz Perez-de-Viñaspre, Alicia Pérez, Arantza Casillas

IXA Taldea. UPV-EHU, Manuel Lardizabal Ibilbidea, 1, Donostia 20018, Spain

ARTICLE INFO

Article history:
Received 1 August 2019
Revised 26 November 2019
Accepted 5 December 2019

Keywords:
Electronic health record
International classification of diseases
Multi-label classification
Recurrent neural networks
Contextual embeddings
Label-granularity

ABSTRACT

Background and objective: This work deals with clinical text mining, a field of Natural Language Processing applied to biomedical informatics. The aim is to classify Electronic Health Records with respect to the International Classification of Diseases, which is the foundation for the identification of international health statistics, and the standard for reporting diseases and health conditions. Within the framework of data mining, the goal is the multi-label classification, as each health record has assigned multiple International Classification of Diseases codes. We investigate five Deep Learning architectures with a dataset obtained from the Basque Country Health System, and six different perspectives derived from shifts in the input and the output.

Methods: We evaluate a Feed Forward Neural Network as the baseline and several Recurrent models based on the Bidirectional GRU architecture, putting our research focus on the text representation layer and testing three variants, from standard word embeddings to meta word embeddings techniques and contextual embeddings.

Results: The results showed that the recurrent models overcome the non-recurrent model. The meta word embeddings techniques are capable of beating the standard word embeddings, but the contextual embeddings exhibit as the most robust for the downstream task overall. Additionally, the label-granularity alone has an impact on the classification performance.

Conclusions: The contributions of this work are a) a comparison among five classification approaches based on Deep Learning on a Spanish dataset to cope with the multi-label health text classification problem; b) the study of the impact of document length and label-set size and granularity in the multi-label context; and c) the study of measures to mitigate multi-label text classification problems related to label-set size and sparseness.

1. Introduction

Methodical documentation of healthcare data is fundamental for public health. The **International Classification of Diseases (ICD)** is the standard diagnoses coding system for **Electronic Health Records (EHR)** classification. ICD serves, worldwide, for epidemiology, health management and documentation purposes. Over time, several versions have been developed, being the ICD-10th the current version. Regarding the hospital network associated with the Spanish "Ministerio de Sanidad, Servicios Sociales e Igualdad", from January the 1st 2016, the clinical modification of the ICD-10th is the reference version, adopting the Spanish translated CIE-10-ES variant as the coding standard. The ICD-10 is designed as an alphanumeric code and it is arranged hierarchically [1]. Each code is built by a set from 3 to 7 alphanumeric characters as shown in Fig. 1.

In this paper we tackle the **task** of automatically coding the diagnostic terms present in a free-text medical record according to the ICD coding system. The task is framed within the Natural Language Processing (NLP) field. The purpose is to determine which classes are present in the input text. Our approach rests on machine learning, specifically on supervised multi-label classification.

Classification based solely on text is an open **challenge** in artificial intelligence [2–4]. We aim to solve a text classification problem on medical free-text, EHRs that present medical jargon and clinical-specific language. Furthermore, EHRs often contain abbreviations (frequently non-standard), and misspellings are also common. The length of the texts plays an important role, here we face

* Corresponding author.
E-mail address: ablanco061@ikasle.ehu.eus (A. Blanco).

https://doi.org/10.1016/j.cmpb.2019.105264
0169-2607/

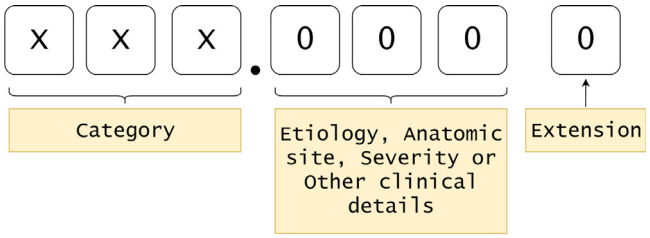

Fig. 1. ICD-10 code structure.

a broad spectrum, ranging from a few words to several tens of lines. EHRs seldom express clinical diagnoses as in the standard ICD.att

An EHR could entail many diagnostics henceforth, multiple ICD labels should be assigned. This task, **Multi-label Classification**, can be seen as a multi-class classification (not binary) task in which the classes are not mutually exclusive. Multi-label classification tends to be, by far, more challenging than mere multi-class classification. Its complexity lies in the exponential growth of label combinations. Note, as well, that the number of labels associated to each EHR is variable.

Multi-label classification can be tackled with the so-called binary relevance approach. This simplistic approach consists of using as many binary classifiers as ICD codes to determine if each ICD code is present or absent from the EHR. The drawback of this approach rests on the fact that the model is not able to capture label-dependencies. While some diagnostics are prone to co-appear others are incompatible. Learning label-dependencies is crucial to this task. To this end, we explore approaches based on Deep Learning [5–7].

The contribution of this work is to explore the impact of dataset characteristics, such as the characterization of the input text focused on (either full document or a part of it), on the predictive ability of the multi-label neural models and also to assess the performance with respect to label-set cardinality and granularity. We deal with real EHRs from Osakidetza (the Basque Public Health System) written in Spanish.[1]

2. Related work

Text classification of EHRs is a demanding task, hence most works have focused on short English texts, though on this work we deal with novel challenges including long EHRs written in Spanish with thousands of words.

Multi-label classification is a challenging task, especially when the number of labels is high [8–10]. The binary relevance approach transforms the multi-label problem in multiple binary classification problems [11], but disregard the dependencies among labels. Several works have addressed the EHRs classification according to the ICD [12–15]. Yet, little attention was paid to dense features and to the approaches that could take advantage of them. Furthermore, much uncertainty still exists about the inter-dependency of labels, that could enhance the prediction performance avoiding incongruities such as, for example, assigning an adult-specific disease simultaneously with a childhood condition. On this work, we tackle the model and capture of label dependencies through Deep Learning models, leveraging the dense output layer with Sigmoid activation function.

The text classification field has leapt forward, from linear and probabilistic models over hand-crafted engineered features [16,17] to non-linear Neural Network models and end-to-end learnt inherent high-level text representations. It is shown good performance with NN [18], as Convolutional Neural Networks [19], Recurrent Neural Networks [20] and Bidirectional Long Short-Term Memory [21].

Methods of **meta-embeddings** aim to conduct a complementary combination of information from an ensemble of distinct word embeddings to yield an embedding set with enhanced quality and characteristics of the semantics captured. Yin and Schütze [22] presented, among others, the “concatenation” method, wherein the meta-embedding is the concatenation of several embeddings. Coates and Bollegala [23] assured that direct averaging of embedding can provide an approximation of the efficiency of concatenation without increasing the dimension of the embeddings.

Context representations are vital to NLP tasks such as text classification. To alleviate this weakness present in generic word embeddings the **contextual embeddings** emerged. Melamud et al. [24] presented an unsupervised model for learning context embedding of wide contexts of sentences using bidirectional LSTMs. These embeddings are dependent on the entire corpus from which they were inferred and carry reinforced contextual meaning. The ELMo [25] and BERT [26] have become state-of-the-art in contextual word representations. Much uncertainty still exists about the advantages of applying meta and contextual embeddings over the standard options for clinical text classification tasks, and we have found that the contextual embeddings may give an extra edge on the ICD classification.

In the automatic ICD coding, there are also works that point towards the Neural Network trend but seems to fall short on the field. These models manage to handle large amounts of text through a dense representation of words. Nigam [27] took advantage of Recurrent Neural Networks to perform multi-label classification. Both works were carried out with discharge summaries from the MIMIC-III [28] corpus. Recently, this task has gained more attention through the CLEF eHealth evaluation labs. Suominen et al. [29] presented an overview of the sixth annual edition. The goal of one of the tasks is to automatically assign ICD-10 codes to few words length texts from free-text descriptions of causes of death as reported by physicians [30,31]. The task is similar to what we have presented on this work with the Diagnostic input perspective, and the finding is that the performance of the classifiers could be improved employing the full documents.

Spanish NLP is under strong growth, among others, driven by the Plan de Tecnologías del Lenguage.[2] EHRs in Spanish are currently being collected [32,33], as well as complimentary corpora including abstracts [34]. These data sets enable to develop several tasks e.g., Negation Extraction [35], Extraction of Adverse Drug Reactions [36], Text Classification [30,31,37], and Negation Cue Detection [38].

3. Methods

We explored four unique RNN model instances plus the baseline model, a Feed Forward Neural Network with Neural-Net Language Model (NNLM) as the text representation layer. The core architecture is a *Bidirectional Recurrent Neural Network with GRU* units and pooling techniques [39] (explained in Section 3.1). The cornerstone of the model is the word embedding layer, as it is responsible for the expressiveness of the input. Thus, we explored three variants: standard embeddings, meta-embeddings and contextual embeddings (explained in depth in Section 3.2). Together

[1] The dataset contains sensitive, confidential data, and therefore can not be released.

[2] https://www.plantl.gob.es/tecnologias-lenguaje/actividades/infraestructuras/Paginas/infraestructuras-linguisticas.aspx.

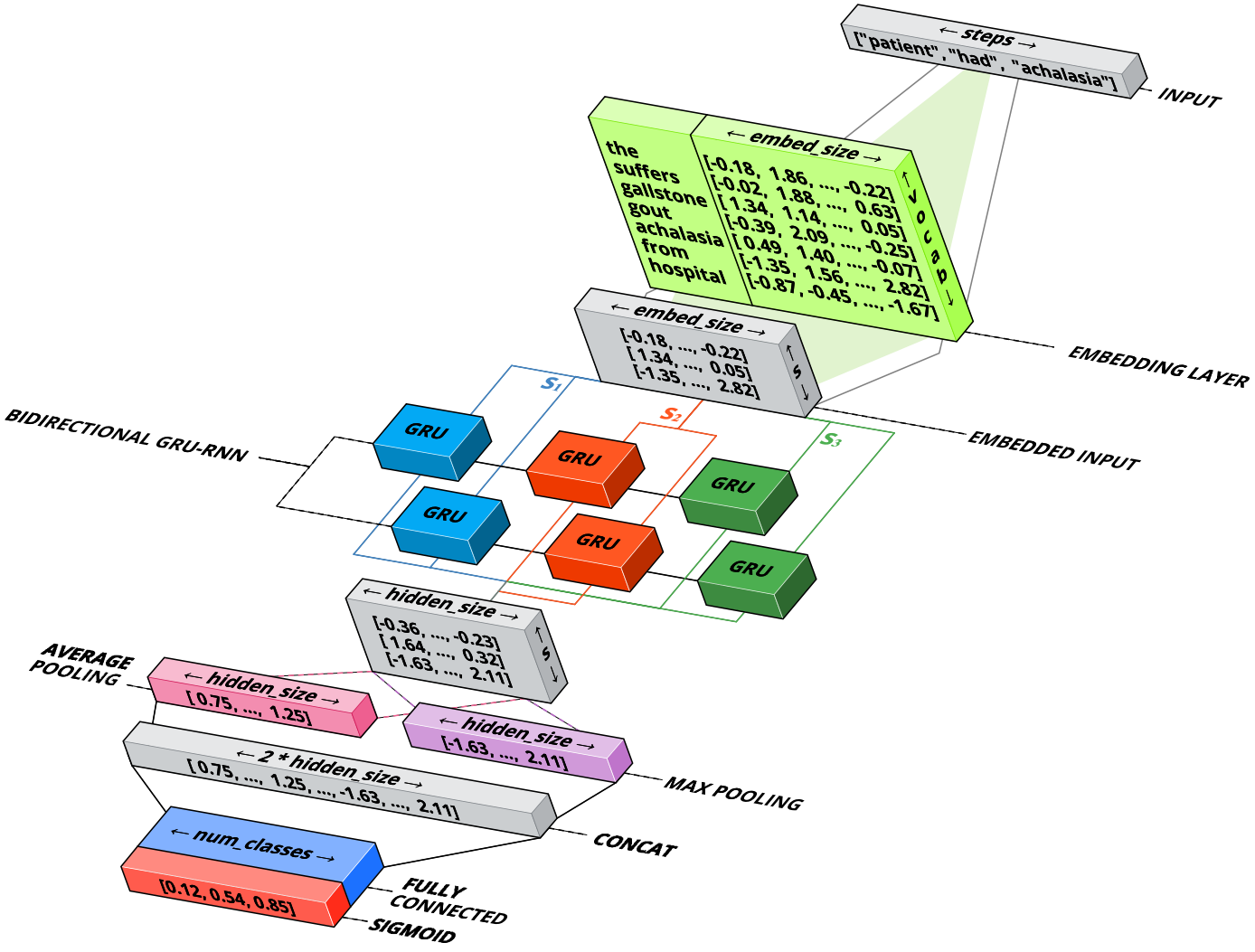

Fig. 2. Architecture: Bidirectional RNN with GRU units and pooling model.

with this work, in an attempt to promote reproducibility, we released the software package that we implemented.[3]

3.1. Bidirectional recurrent neural network with GRU units and pooling

We applied a Bidirectional layer with GRU units, which leverages sequences of text in forward and reverse order with separate hidden states, and whose mathematical formulation for the forward and backward hidden state and its combination is shown in (1).

$$\begin{aligned} \overrightarrow{h^{(t)}} &= \sigma(\overrightarrow{W}x^{(t)} + \overrightarrow{V}\,\overrightarrow{h}^{(t-1)} + \overrightarrow{b}) \\ \overleftarrow{h^{(t)}} &= \sigma(\overleftarrow{W}x^{(t)} + \overleftarrow{V}\,\overleftarrow{h}^{(t-1)} + \overleftarrow{b}) \\ h^{(t)} &= [\overrightarrow{h}^{(t)}, \overleftarrow{h}^{(t)}] \end{aligned} \tag{1}$$

The parameters are the weight matrices $[\overrightarrow{W}, \overleftarrow{W}]$ and $[\overrightarrow{V}, \overleftarrow{V}]$, and the bias terms $[\overrightarrow{b}, \overleftarrow{b}]$. The hidden-states are computed through the non-linear activation (σ) applied to the weighted sum between previous hidden-states $[\overrightarrow{h}^{(t-1)}, \overleftarrow{h}^{(t-1)}]$ and current input ($x^{(t)}$) with their corresponding matrices. Then, both hidden states are combined with concatenation to provide the resulting hidden state (h^t).

The output of the Bidirectional RNN layer could be fed to the dense layer. However, this can be computationally challenging, due to the high number of parameters. Learning a classifier with too many parameters can be unwieldy, and can also be prone to overfitting. A popular technique to deal with the high dimensionality of the Bidirectional RNN layer output is Pooling [40]. We applied average and max-pooling, known as 1-dimensional global pooling. The pooled features are concatenated and fed into a final fully-connected layer. This layer is responsible for computing the probability estimation of the labels i.e. ICD codes. Fig. 2 shows the full architecture of the Bidirectional Recurrent Neural Network with GRU units and pooling techniques, i.e. BiGru. The figure shows a forward pass for an example text. The output of the Sigmoid function is the probability estimation of each label. The depth of every layer indicates the batch size. The Recurrent layer is unrolled, so $s_i \forall i \in s$ brings the embedded representation of the input token $\{s_1 = emb(\text{“patient”}),\ s_2 = emb(\text{“had”}),\ s_3 = emb(\text{“achalasia”})\}$.

The BiGru model can handle all the labels at once, instead of following a binary relevance approach, training independent classifiers for each label. The final dense layer is able to capture and model the label dependencies, producing a non-mutually exclusive

[3] The software is available at http://ixa2.si.ehu.es/prosamed/cmpICD_soft and can be downloaded with user `CMPB` and password `IXAcmpb`. Provided that the software is used anyhow, this article should be cited.

probability estimation for each label with the Sigmoid activation function [41].

3.2. Comprehensive input characterization: embedding layer variations

A comprehensive input characterization is crucial for attaining competitive performance. In the training stage, the embedding layer holds more than 90% of the model's complexity in terms of parameter count. What is more, the predictive capacity rests on the ability of the model to extract knowledge from the source provided in the input stage. Thus, we paid special attention to this layer. The embedding layer from the Fig. 2 shows just a vanilla embedding layer that we enhanced later. Indeed, in this work we explored three variations of the embedding layer: i) Standard embeddings. ii) Meta embeddings (Sections 3.2.1 and 3.2.2). iii) Contextual embeddings (Section 3.2.3)

Moreover, according to Yin et al. [42] and Coates and Bollegala [23], different pre-trained word embeddings have substantial differences in quality and characteristics of the word representations. The consequence is some word embeddings performing better on some tasks than in others. Bearing all this in mind, in addition to a standard pre-trained embedding, we tried *meta-embeddings*, which are ensemble approaches (embedding concatenation and blending) with the hope to get an embedding set with the improved overall quality.

We turned to embeddings derived from fastText [43] as the standard embeddings setup. As for meta-embeddings setup, we employed fastText, Word2Vec [44] and GloVe [45]. Every embedding set is trained on the same corpus, the Spanish Billion Word Corpus [46].

3.2.1. Embedding concatenation

The meta-embedding is computed as the concatenation of word embeddings, based on the work by Yin and Schütze [22]. Before the concatenation, each embedding set must be L2-normalized [6], so that all the values are in the range $[-1, 1]$ and, therefore, every set contributes equally.

The dimensionality of the resulting meta-embeddings is $\hat{d}_{s_k} = d_{s_1} + \ldots + d_{s_i} + d_{s_n}$ with d_{s_i} being the dimension of the ith set concatenated. It is important to note that the model's complexity increases with each added embedding set, as it increases the dimension of the features of the embedding layer.

3.2.2. Embedding blending

The meta-embedding variant is computed as the average of the embeddings involved, based on the work by Coates and Bollegala [23]. Note that even having embedding sets with matching number of dimensions ($d_{s_i} = d_{s_j} \forall\ i, j$), each dimension among embeddings is not related. In any case, averaging can provide an approximation of the performance of concatenation without the expense of increasing the dimension [23].

3.2.3. Contextual embeddings

Recently, approaches that improve the semantic word representation by leveraging the context to encode syntactical meaning and handle polysemy are pushing the state-of-the-art. Regular word embedding techniques use all the occurrences of a word to extract a joint representation. However, depending on the context, words could have different meanings. Recent models exploit this reasoning and propose contextual word embeddings. There is no longer a lookup table between words and dense representations. Instead, the word embedding is computed on the fly, taking advantage of the context.

Embeddings from Language Models (ELMo) [25] representations are obtained from a bidirectional Language Model (biLM) that has recently produced state-of-the-art results in several NLP tasks like Coreference Resolution [47] or Natural Language Inference and Sentiment Analysis [25]. The embedding for a given word varies from one sentence or document to another with its context. As it cannot be pre-computed, the embedding computation is done computing a forward propagation of the model for each token of each input sequence [48].

Table 1
Statistical characterization of the Osakidetza dataset and every perspective. The input can comprise either a small part of the document denoted as "Diagnostic" or the entire "Document" (full EHR). The output (ICD code) can be explored at different granularity levels: Chapter, Block, Full-code.

	Corpus	EHRs		
$\mathcal{S}$	Samples	10,707		
	Input	Diagnostic	Document	
$\mathcal{X}$	Vocabulary size	12,811	60,197	
	Words per doc	37.9 ± 73.8	770.5 ± 351.2	
	Output	Chapter	Block	Full
$\mathcal{Y}$	Distinct labels	24	991	3572
	Avg. Cardinality	3.7	5.4	5.5

4. Experimental framework

4.1. Data

The datasets used in our experiments consist of EHRs written in Spanish from the Basque public health system (Osakidetza). Specifically, emergency services discharge summaries from hospitals. The EHRs are not structured and were not written using templates with sections. Table 1 introduces the details of the dataset used. There are 10,707 EHRs. As revealed by the table, we considered several perspectives of the **dataset** by varying two factors, the input and the output explained in what follows.

According to the **input**, the shift consists of the retained proportion of text from the full EHR. Our aim is to determine whether the neural models were able to extract the information from entire documents or, could be benefited from small pieces of text conveying meaningful information. As a result, we explored, on the one hand, the full EHR (referred to as "Document") and on the other hand a short part of the document (referred to as "Diagnostic") as shown in Fig. 3. Note that the mean text length of the Diagnostic perspective is ≈ 38 words, while for the Document perspective, it

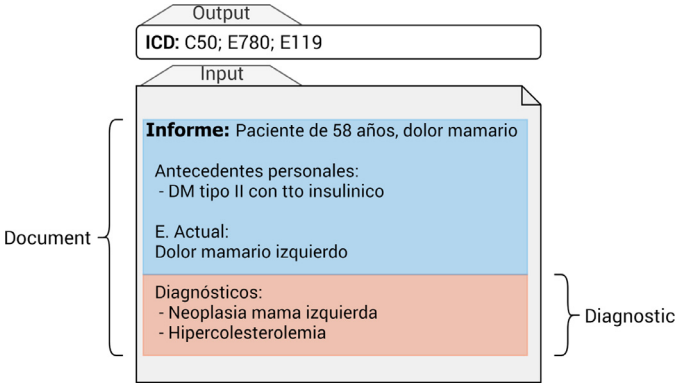

Fig. 3. Health record from the Osakidetza datasets. The ICD codes are listed right to the ''ICD:'' keyword. Every character after the ''Informe:'' keyword is part of the text of the report.

Table 2
Evaluation results for the *{**Corpus:** Big, **Out:** Full-code}* varying the input (full document or diagnostic section).

Inp	Model	Characterization		Precision	Recall	F-Score
Document	NNLM	Document embeddings		50.968	30.976	31.513
	BiGru	Standard fastText emb.		51.485	55.825	53.307
		Meta embeddings	Avg	64.824	57.416	59.533
			Conc	65.345	54.428	58.576
		Contextual ELMo emb.		**67.283**	**60.490**	**63.165**
Diagnostic	NNLM	Document embeddings		54.971	34.974	39.257
	BiGru	Standard fastText emb.		51.257	49.785	49.838
		Meta embeddings	Avg	58.482	50.860	53.864
			Conc	**58.752**	50.688	53.992
		Contextual ELMo emb.		56.648	**52.472**	**54.301**

rises to $\approx$ 770, but in both cases, the standard deviation is high, as shown in Table 1.

Regarding the **output**, the shift is in the granularity of the labels, with a three-level alternative taken into account, as follows: The *"Full-code"* level preserves the original code (e.g., "`M1A.1421`": *Chronic gout, lead induced in hand left with tofus*), the *"Block"* level keeps the first three characters (e.g., "`M1A`": *Chronic gout*), and the *"Chapter"* level keeps only the first character (e.g., "`M`": *Diseases of the muskuloskeletal system and connective tissue*).

Fig. 4 shows the resulting label distributions for each output alternative of the Osakidetza datasets. The diseases are not uniformly distributed, as there are frequent and rare conditions. Indeed, the class imbalance is one of the challenges of ICD classification, as machine learning models struggle to handle classification tasks with classes that present large disparities in prevalence. To cope with the high imbalance and high scarcity of some labels Dermouche et al. [49] set a threshold of minimum occurrences per label i.e. keeping only those labels that appear in more than 15 records. Following similar reasoning, and to keep consistency across perspectives, we set a relative threshold, based on the percentage of appearances. Specifically, we keep only those labels that appear in at least 5% of EHRs. The relative threshold enables to keep every perspective label-set coherent concerning the label distribution and minimum class support, while leaving enough samples with each label to evaluate on the test set.

4.2. Results

The experimental set is designed to provide a full range of insights about the application of neural networks to an EHR-ICD based multi-label Spanish text classification task. To that end, we have explored the performance of the 5 models over 6 dataset perspectives.

4.2.1. Assessing the models and the impact of the input text

To evaluate the impact of document length, we make use of the Diagnostic and Document perspectives. Here, the intuition is that extracting the most relevant part of the documents may improve the results by focusing the attention and preventing long-range sequence problems [50,51], but also may harm, due to loss of information, like the mention of symptoms or drugs, on the discarded text.

Table 2 assess the models with either diagnostic or full document as input and full-code labels as output.

Comparing the **models and representation**, we can derive the best performing approach. The baseline is outperformed by every model by a noteworthy amount, besides, the BiGru ELMo outperformed the others in terms of F-score. The BiGru with standard embeddings obtained average results, and the meta-embeddings surpassed the model using just standard one-sided embeddings derived from fastText.

Comparing the amount of information conveyed by the **input text** and relevant to the models to make their predictions, the results are of much interest. All the recurrent models are favoured when providing the full document as input (just the baseline is superior for the diagnostic input). Indeed, the mean difference in terms of F-score in favour of the document input is around 5 points. These results could lead to an extensive discussion. We now bestow an argument: the recurrent models can take advantage of longer sequences with more information and larger vocabulary successfully. Notwithstanding, if the model is not suitable for sequential data (like the baseline model, NNLM, which is not recurrent), those long sequences and large vocabulary weaken the performance, allowing an improvement by keeping shorter text fragments. With these results in mind, we recommend text summarization or similar techniques for extracting the most relevant fragments of texts as a mitigation measure in case of non-recurrent models for document classification tasks. The results attained by the best models for each document length (i.e. Diagnostic and Document perspectives) are marked in bold.

4.2.2. Assessing the impact of label-set size and granularity

Regarding the influence of label-set size and granularity, the intuition is that as the label-set size decreases the performance increases, as the inherent difficulty of the problem diminishes. Besides, it is interesting to explore if the granularity, the degree of label detail, by itself, has an impact on performance.

It is important to remember that due to the relative threshold for label-set reduction, the block label-set of the big dataset has more labels than the full-code label-set, as shown in Fig. 4. Hence, this scenario allows us to check the impact of label-granularity alone.

Due to the high number of entries that a table would require ($n = 30$), and for the sake of clarity, we have chosen to show the outcome of this experiment by a line plot. Fig. 5 shows the F-score (y-axis) for the full-code, block and chapter labels (x-axis) achieved by each of the five models explored for both input perspectives, and we can observe similar behaviours.

Focusing on the document input, we can observe that the behaviour for every model is also similar, improving results as the granularity decreases. One key finding is that the granularity has an impact alone. With less granularity, the performance increases, even with more number of labels. This finding is depicted by the situation between the full labels ($n = 16$) and the block labels ($n = 19$), where with the block labels the performance improves despite having 3 more labels. This suggests that is possible to get models performing better with the same number of labels by just decreasing the label granularity.

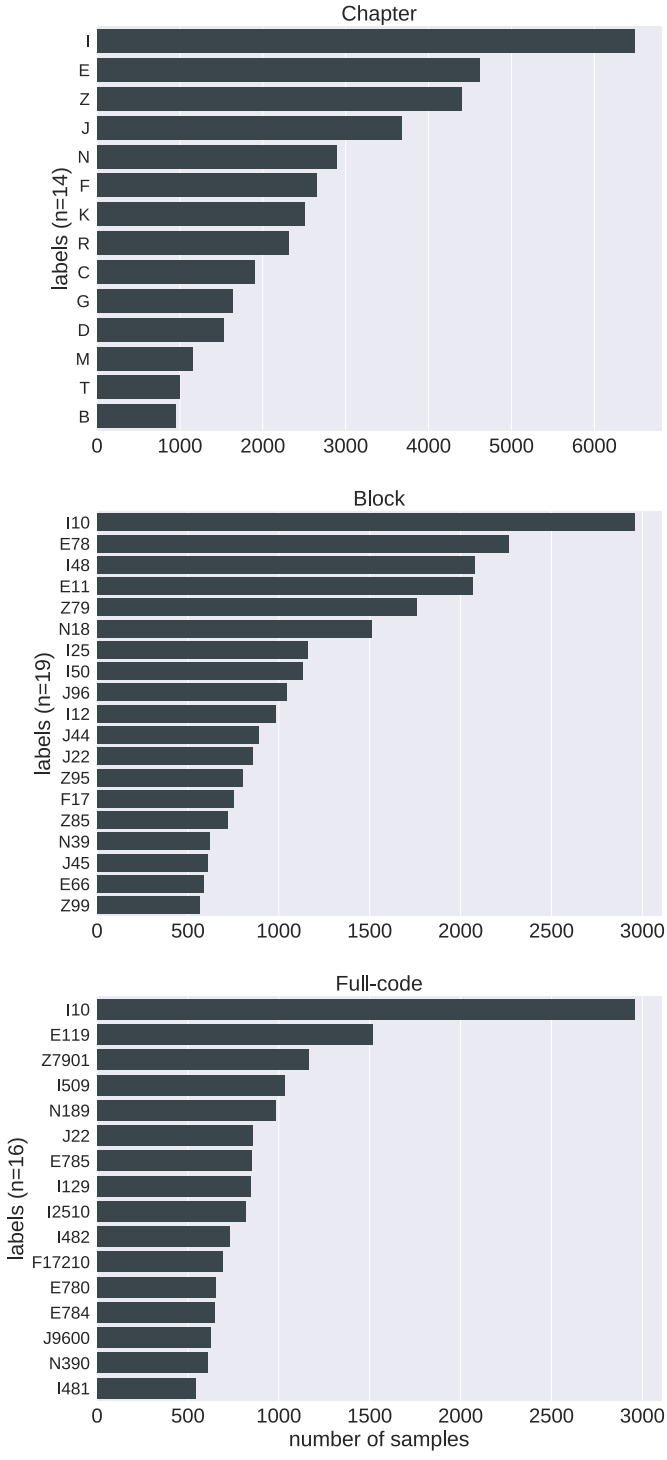

Fig. 4. Label distributions obtained after the reduction of the label-set size with a relative threshold of 5% carried for each output perspective of the Osakidetza datasets.

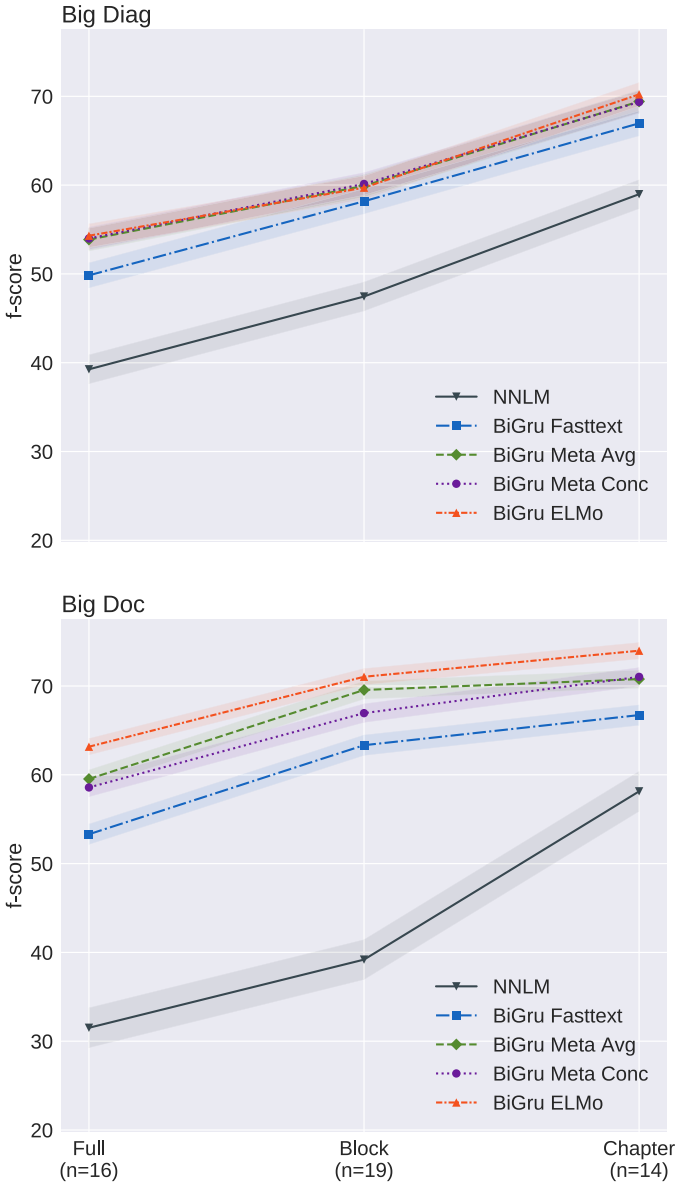

Fig. 5. Line plots for performance comparison among the full-code, block and chapter labels for both perspectives: diagnostic (top), document (bottom).

4.2.3. Discussion

With this work we gained the following **insights**: Despite is a difficult task, Deep Learning recurrent models exhibit strong predictive capabilities and can be enhanced by more robust text representation techniques such as the meta or contextual embeddings.

We argue that our experimental results throw one key **finding**: the granularity of the labels alone has an impact on performance. The significance lies in the possibility of performance improvement by reducing the granularity without reducing the label-set size.

BiGru powered by ELMo is the dominant model in practically every situation from both the input and output perspectives (shown in Table 2 and Fig. 5). Accordingly, a **per-class evaluation** on the best-performing dataset perspective is shown in Fig. 6.

Draw attention to the fact that the worse performing label T (“*Injury, poisoning and certain other consequences of external causes*”) gets 41.7% F-score, while the best-performing label C (“*Neoplasms*”) reaches an outstanding 91%. Half of the labels are above 70% and the $\approx$ 30% of labels are above 80%.

To assess the stability of the models and the statistical significance of the results, we performed five runs repeating the experimental set with random seeds and found that Stdev. among runs remained under 0.5 for precision and recall and 0.25 for F-score for every model and setup, which means that the given experimental results are both reproducible and representative.

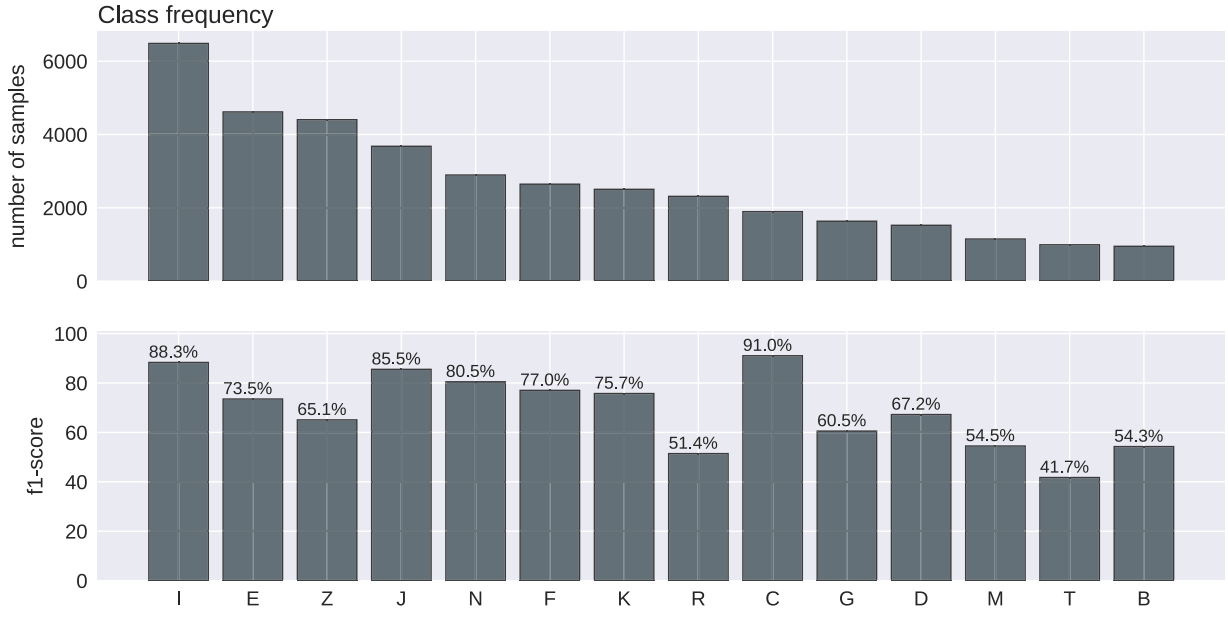

Fig. 6. Per-class evaluation of BiGru ELMo based on F-score and class frequency for the {***Inp:*** *Document*, ***Out:*** *Chapter}* subtask.

5. Conclusions

We presented a set of Deep Learning methods to tackle the NLP challenge of multi-label text classification with medical free-text: EHRs written in Spanish with datasets from the Basque Country Health System and classified according to the ICD. Each EHR is assigned multiple ICD codes, leading to multi-label classification of text.

In this work we turned to deep neural models and we found that contextual information conveyed by the BiGru ELMo achieved competitive results. BiGru, by contrast to main approaches seen in the literature, has a mechanism to cope with label co-apparitions and regard diseases as related.

We wondered if the neural models were able to extract the information from entire EHRs of nearly a thousand words or could be boosted by selecting a small though representative section (diagnoses). Experimental results showed that it is worthy providing the model with the full document as it might convey meaningful information. Particularly, BiGru powered with contextual embeddings form ELMo(BiGru+ELMo) outperformed the rest of the models explored. In fact, BiGru+ELMooutperformed every model in all the setups. The difficulty of correctly predicting a label is not the same across labels. A per-class evaluation revealed the competitive performance of this approach on minority classes. That is, BiGru+ELMo resulted robust regarding the class imbalance and, obviously, leveraged frequent ICDs.

Finally, we explored the performance attained varying the output label granularity (fully-specified code, block, chapter) and label-set cardinality (from 14 to 19). This is of interest to decide whether to create a fully automatic ICD classification engine or, depending on the performance required, make the decision to let the model just predict a higher order in the hierarchy.

There are several open directions for future work. First, our models leverage ELMo based contextual embeddings, but there are other novel approaches to contextual embeddings based on Language Models, like BERT [26]. Second, the core architecture of this work is the Recurrent Neural Network, but there are other intriguing architectures like Convolutional Neural Networks, especially Capsule Network [52] or the architecture behind BERT, the new RNN alternative promising approach called Transformer [53]. Third, the methods to address the relation among labels, such as statistical driven approaches (e.g., correlation analysis [54]) and strategies leveraging the hierarchically structured ICD and related ontologies (e.g., Hierarchical Multi-label Classification [55] and SNOMED-CT [56]).

Ethical

This article does not contain any studies with human participants or animals performed by any of the authors.

Declaration of Competing Interest

The authors declare that there is no conflict of interest.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.cmpb.2019.105264

References

[1] W. H. Organization, International Statistical Classification of Diseases and Related Health Problems, 1, World Health Organization, 2004.
[2] H.T. Madabushi, M. Lee, High accuracy rule-based question classification using question syntax and semantics, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 2016, pp. 1220–1230.
[3] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. St. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, B. Strope, R. Kurzweil, Universal sentence encoder for English, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 169–174.
[4] J. Howard, S. Ruder, Universal language model fine-tuning for text classification, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1, 2018, pp. 328–339.
[5] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016. http://www.deeplearningbook.org
[6] Y. Goldberg, A primer on neural network models for natural language processing, J. Artif. Intell. Res. 57 (2016) 345–420.
[7] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.
[8] K. Bhatia, H. Jain, P. Kar, M. Varma, P. Jain, Sparse local embeddings for extreme multi-label classification, in: Advances in Neural Information Processing Systems, 2015, pp. 730–738.
[9] H. Jain, Y. Prabhu, M. Varma, Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2016, pp. 935–944.

[10] K. Jasinska, K. Dembczynski, R. Busa-Fekete, K. Pfannschmidt, T. Klerx, E. Hullermeier, Extreme F-measure maximization using sparse probability estimates, in: International Conference on Machine Learning, 2016, pp. 1435–1444.
[11] M.-L. Zhang, Y.-K. Li, X.-Y. Liu, X. Geng, Binary relevance for multi-label learning: an overview, Front. Comput. Sci. 12 (2) (2018) 191–202.
[12] P. Franz, A. Zaiss, S. Schulz, U. Hahn, R. Klar, Automated coding of diagnoses–three methods compared., in: Proceedings of the AMIA Symposium, American Medical Informatics Association, 2000, p. 250.
[13] A. Perotte, R. Pivovarov, K. Natarajan, N. Weiskopf, F. Wood, N. Elhadad, Diagnosis code assignment: models and evaluation metrics, J. Am. Med. Inform. Assoc. 21 (2) (2013) 231–237.
[14] M. Saeed, M. Villarroel, A.T. Reisner, G. Clifford, L.-W. Lehman, G. Moody, T. Heldt, T.H. Kyaw, B. Moody, R.G. Mark, Multiparameter intelligent monitoring in intensive care II (MIMIC-II): a public-access intensive care unit database, Crit. Care Med. 39 (5) (2011) 952.
[15] J. Pérez, A. Pérez, A. Casillas, K. Gojenola, Cardiology record multi-label classification using latent dirichlet allocation, Comput. Methods Programs Biomed. 164 (2018) 111–119.
[16] T. Joachims, Text categorization with support vector machines: learning with many relevant features, in: European Conference on Machine Learning, Springer, 1998, pp. 137–142.
[17] A. McCallum, K. Nigam, et al., A comparison of event models for naive Bayes text classification, in: AAAI-98 Workshop on Learning for Text Categorization, 752, Citeseer, 1998, pp. 41–48.
[18] J. Nam, J. Kim, E.L. Mencía, I. Gurevych, J. Fürnkranz, Large-scale multi-label text classification revisiting neural networks, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2014, pp. 437–452.
[19] Y. Kim, Convolutional neural networks for sentence classification, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1746–1751, doi:10.3115/v1/D14-1181.
[20] D. Tang, B. Qin, X. Feng, T. Liu, Target-dependent sentiment classification with long short term memory, CoRR, abs/1512.01100(2015).
[21] P. Zhou, Z. Qi, S. Zheng, J. Xu, H. Bao, B. Xu, Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 3485–3495.
[22] W. Yin, H. Schütze, Learning word meta-embeddings, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 1351–1360, doi:10.18653/v1/P16-1128.
[23] J. Coates, D. Bollegala, Frustratingly easy meta-embedding – computing meta-embeddings by averaging source word embeddings, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 194–198, doi:10.18653/v1/N18-2031.
[24] O. Melamud, J. Goldberger, I. Dagan, Context2Vec: learning generic context embedding with bidirectional LSTM, in: Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning, 2016, pp. 51–61.
[25] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 2227–2237, doi:10.18653/v1/N18-1202.
[26] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805(2018).
[27] P. Nigam, Applying deep learning to ICD-9 multi-label classification from medical records, 2016.
[28] A.E. Johnson, T.J. Pollard, L. Shen, H.L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L.A. Celi, R.G. Mark, MIMIC-III, a freely accessible critical care database, Sci. Data 3 (2016) 160035.
[29] H. Suominen, L. Kelly, L. Goeuriot, A. Névéol, L. Ramadier, A. Robert, E. Kanoulas, R. Spijker, L. Azzopardi, D. Li, et al., Overview of the CLEF eHealth evaluation lab 2018, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2018, pp. 286–301.
[30] A. Atutxa, A. Casillas, N. Ezeiza, V. Fresno, I. Goenaga, K. Gojenola, R. Martínez, M.O. Anchordoqui, O. Perez-de Viñaspre, IxaMed at CLEF eHealth 2018 task 1: ICD10 coding with a sequence-to-sequence approach, in: CLEF (Working Notes), 2018, p. 1.
[31] M. Almagro, S. Montalvo, A.D. de Ilarraza, A. Pérez, MAMTRA-MED at CLEF eHealth 2018: a combination of information retrieval techniques and neural networks for ICD-10 coding of death certificates, in: CLEF (Working Notes), 2018, p. 1.
[32] M. Oronoz, K. Gojenola, A. Pérez, A.D. de Ilarraza, A. Casillas, On the creation of a clinical gold standard corpus in Spanish: mining adverse drug reactions, J. Biomed. Inform. 56 (2015) 318–332.
[33] M. Marimon, B. Fisas, N. Bel, J. Vivaldi, S. Torner, M. Lorente, S. Vázquez, M. Villegas, The IULA Treebank, in: Lrec, 2012, pp. 1920–1926.
[34] A. Duque, M. Stevenson, J. Martinez-Romo, L. Araujo, Co-occurrence graphs for word sense disambiguation in the biomedical domain, Artif. Intell. Med. 87 (2018) 9–19.
[35] S.M. Jiménez-Zafra, M. Taulé, M.T. Martín-Valdivia, L.A. Ureña-López, M.A. Martí, SFU review SP-NEG: a Spanish corpus annotated with negation for sentiment analysis. A typology of negation patterns, Lang. Resour. Eval. 52 (2) (2018) 533–569.
[36] S. Santiso, A. Pérez, A. Casillas, Exploring joint AB-LSTM with embedded lemmas for adverse drug reaction discovery, IEEE J. Biomed. Health Inform. 23 (5) (2019) 2148–2155.
[37] M. Almagro, R. Martínez Unanue, V. Fresno Fernández, S. Montalvo Herranz, Estudio preliminar de la anotación automática de códigos CIE-10 en informes de alta hospitalarios, SEPLN, 2018.
[38] H. Fabregat, A. Duque, J. Martinez-Romo, L. Araujo, Extending a deep learning approach for negation cues detection in Spanish, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019). CEUR Workshop Proceedings, CEUR-WS, Bilbao, Spain, 2019, p. 1.
[39] H. Sak, A. Senior, F. Beaufays, Long short-term memory recurrent neural network architectures for large scale acoustic modeling, in: Fifteenth Annual Conference of the International Speech Communication Association, 2014, p. 1.
[40] Y.-T. Zhou, R. Chellappa, Computation of optical flow using a neural network, in: IEEE International Conference on Neural Networks, 1998, 1988, pp. 71–78.
[41] J. Liu, W.-C. Chang, Y. Wu, Y. Yang, Deep learning for extreme multi-label text classification, in: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2017, pp. 115–124.
[42] Y. Yin, Y. Song, M. Zhang, Nnembs at semeval-2017 task 4: neural twitter sentiment classification: a simple ensemble method with different embeddings, in: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), 2017, pp. 621–625.
[43] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Trans. Assoc. Comput. Linguist. 5 (2017) 135–146, doi:10.1162/tacl_a_00051.
[44] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems, 2013, pp. 3111–3119.
[45] J. Pennington, R. Socher, C. Manning, Glove: global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543.
[46] C. Cardellino, Spanish Billion Words Corpus and Embeddings, 2016.
[47] K. Lee, L. He, L. Zettlemoyer, Higher-order coreference resolution with coarse-to-fine inference, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 687–692, doi:10.18653/v1/N18-2108.
[48] M. Fares, A. Kutuzov, S. Oepen, E. Velldal, Word vectors, reuse, and replicability: Towards a community repository of large-text resources, in: Proceedings of the 21st Nordic Conference on Computational Linguistics, Association for Computational Linguistics, Gothenburg, Sweden, 2017, pp. 271–276.
[49] M. Dermouche, J. Velcin, R. Flicoteaux, S. Chevret, N. Taright, Supervised topic models for diagnosis code assignment to discharge summaries, in: International Conference on Intelligent Text Processing and Computational Linguistics, Springer, 2016, pp. 485–497.
[50] C. Manning, P. Raghavan, H. Schütze, Introduction to information retrieval, Nat. Lang. Eng. 16 (1) (2010) 100–103.
[51] R. Pascanu, T. Mikolov, Y. Bengio, On the difficulty of training recurrent neural networks, in: International Conference on Machine Learning, 2013, pp. 1310–1318.
[52] S. Sabour, N. Frosst, G.E. Hinton, Dynamic routing between capsules, in: Advances in Neural Information Processing Systems, 2017, pp. 3856–3866.
[53] 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.
[54] Y. Zhang, J. Schneider, Multi-label output codes using canonical correlation analysis, in: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, 2011, pp. 873–882.
[55] J. Wehrmann, R. Cerri, R. Barros, Hierarchical multi-label classification networks, in: International Conference on Machine Learning, 2018, pp. 5225–5234.
[56] K. Donnelly, SNOMED-CT: the advanced terminology and coding system for ehealth, Stud. Health Technol. Inform. 121 (2006) 279.