Neural Networks for Multi-lingual Multi-label Document Classification

Jiří Martínek[1,2], Ladislav Lenc[1,2], and Pavel Král[1,2]

[1] Dept. of Computer Science & Engineering
Faculty of Applied Sciences
University of West Bohemia
Plzeň, Czech Republic
[2] NTIS - New Technologies for the Information Society
Faculty of Applied Sciences
University of West Bohemia
Plzeň, Czech Republic
{jimar,pkral,llenc}@kiv.zcu.cz @kiv.zcu.cz

Abstract. This paper proposes a novel approach for multi-lingual multi-label document classification based on neural networks. We use popular convolutional neural networks for this task with three different configurations. The first one uses static word2vec embeddings that are let as is, while the second one initializes it with word2vec and fine-tunes the embeddings while learning on the available data. The last method initializes embeddings randomly and then they are optimized to the classification task. The proposed method is evaluated on four languages, namely English, German, Spanish and Italian from the Reuters corpus. Experimental results show that the proposed approach is efficient and the best obtained F-measure reaches 84%.

Keywords: convolutional neural network, cnn, document classification, multi-label, multi-lingual

1 Introduction

Nowadays the importance of multi-lingual text processing increases significantly due to the extremely rapid growth of data available in several languages particularly on the Internet. Without multi-lingual systems it is not possible to acquire information across languages. Multi-label classification is also often beneficial because, in the case of real data, one sample usually belongs to more than one class.

This paper focuses on the multi-lingual multi-label document classification in a frame of a real application designed for handling texts from different sources in various languages. There are several possibilities how to perform a classification in multiple languages. Most of them learn one model in a mono-lingual space and then use some transformation method to pass across the languages. The usual document representation are word embeddings created for instance by the word2vec approach [8]. Contrary to this idea, we suggest one general model trained on all available languages. Therefore, this model is able to classify more languages without any transformation.

We use popular convolutional networks for this task with three different settings. The first one uses static word2vec embeddings that are not trained. The second one initializes the embeddings with word2vec and fine-tunes it on the available data. The last method initializes embeddings randomly and then they are, as in the previous case, optimized to the given task using available data. All these methods use the same vocabulary.

To the best of our knowledge, there is no previous study, which uses one classifier on multi-lingual multi-label data as proposed in this paper. The proposed approach is evaluated on four languages (English, German, Spanish and Italian) from the standard Reuters corpus.

2 Related Work

This section first presents the usage of neural networks for document classification and then focuses on multi-linguality.

Feed-forward neural networks were used for multi-label document classification in [16]. The authors have modified the standard backpropagation algorithm for multi-label learning which employs a novel error function. This approach is evaluated on functional genomics and text categorization.

Le and Mikolov propose [8] so called *Paragraph Vector*, an unsupervised algorithm that addresses the issue of necessity of a fixed-legth document representation. This algorithm represents each document using a dense vector. This vector is trained to predict words in the document. The authors obtain new state of the art results on several text classification and sentiment analysis tasks.

A recent study on the multi-label text classification was presented by Nam et al. [12]. The authors use the cross-entropy algorithm instead of ranking loss for training and they also further employ recent advances in deep learning field, e.g. the rectified linear units activation and AdaGrad learning with dropout [11, 14]. Tf-idf representation of documents is used as a network input. The multi-label classification is done by thresholding of the output layer. The approach is evaluated on several multi-label datasets and reaches results comparable or better than the state of the art.

Another method [7] based on neural networks leverages the co-occurrence of labels in the multi-label classification. Some neurons in the output layer capture the patterns of label co-occurrences, which improves the classification accuracy. The architecture is basically a convolutional network and utilizes word embeddings as inputs. The method is evaluated on the natural language query classification in a document retrieval system.

An alternative multi-label classification approach is proposed by Yang and Gopal [15]. The conventional representations of texts and categories are transformed into meta-level features. These features are then utilized in a learning-to-rank algorithm. Experiments on six benchmark datasets show the abilities of this approach in comparison with other methods.

Recent work in the multi-lingual text representations field is usually based on word-level alignments. Klementiev et al. [5] train simultaneously two language models based on neural networks. The proposed method uses a regularization which ensures that pairs of frequently aligned words have similar word embeddings. Therefore, this approach needs parallel corpora to obtain the word-level alignment. Zou et al. [17] propose an alternative approach based on neural network language models using different regularization.

Kovčisky et al. [6] propose a bilingual word representations approach based on a probabilistic model. This method simultaneously learns alignments and distributed representations from bilingual data. This method marginalizes out the alignments, thus captures a larger bilingual semantic context. Chandar et al. [1] investigate an efficient approach based on autoencoders that uses word representations coherent between two languages. This method is able to obtain high-quality text representations by learning to reconstruct the bag-of-words of aligned sentences without any word alignments.

Coulmance et al. [2] introduce an efficient method for bilingual word representations called Trans-gram. This approach extends popular skip-gram model to multi-lingual scenario. This model jointly learns and aligns word embeddings for several languages, using only monolingual data and a small set of sentence-aligned documents.

3 Multi-lingual Document Classification

3.1 Multi-lingual Document Representation

The documents are represented as sequences of word indexes in a shared vocabulary $\boldsymbol{V}$ which is constructed in a following way. Let N be a number of the available languages. $\boldsymbol{V_n}$ represents the vocabulary of most frequent words in the given language. The shared vocabulary $\boldsymbol{V}$ is then constructed by the following equation

$$\boldsymbol{V} = \bigcup_{n=1}^{N} \boldsymbol{V_n} \tag{1}$$

The convolutional network we use for classification requires that the inputs have the same dimensions. Therefore, the documents with fewer words than a specified limit are padded, while the longer ones must be shortened. This is different from Kim's approach [3] where documents are padded to the length of the longest document in the training set. We are working with much longer documents where the lengths vary significantly. Therefore, the shortening of some documents and thus losing some information is inevitable in our case. However, based on our preliminary experiments, the influence of document shortening is insignificant to document classification score.

3.2 Neural Network Architecture

Neural network learns a function $f : d \rightarrow \boldsymbol{C_d}$ which maps document $d \in \boldsymbol{D}$ to a set of categories $\boldsymbol{C_d} \subset \boldsymbol{C}$. $\boldsymbol{D}$ is the set of classified documents and $\boldsymbol{C}$ is the set of all possible categories.

We use a CNN architecture that was proposed in [9]. This architecture utilizes one-dimensional convolutional kernels which is the main difference from the network proposed by Kim in [3] where 2D kernels over the entire width of the word embeddings are used. The input of our network is a vector of word indexes of the length M where M is the number of words used for document representation. The second layer is an embedding layer which represents a look-up table for the word vectors. It translates the word indexes into word vectors of length E. The document is then represented as a matrix with M rows and E columns. The next layer is the convolutional one. We use N_C convolution kernels of the size $K \times 1$ which means we do 1D convolution over one position in the embedding vector over K input words. The following layer performs a max-pooling over the length $M - K + 1$ resulting in N_C $1 \times E$ vectors. The output of this layer is then flattened and connected to a fully-connected layer with E nodes. The output layer contains $|\boldsymbol{C}|$ nodes where $|\boldsymbol{C}|$ is the cardinality of the set of classified categories.

The output of the network is then thresholded to get the final results. The values greater than a given threshold indicate the labels that are assigned to the classified document. The architecture of the network is depicted in Figure 1. This figure shows the processing of two documents in different languages (English and German) by our network. Each document is handled in one training step. The key concept is the shared vocabulary and the corresponding shared embedding layer.

4 Experiments

4.1 Reuters RCV1/RCV2 Dataset

The Reuters RCV1 dataset [10] contains a large number of English documents. The RCV2 is a multi-lingual corpus that contains news stories in 13 languages. The distribution of the document lengths is shown in Figure 2. We use four languages, namely English, German, Spanish and Italian. We prepare two settings: single- and multi-label ones.

Single-label Configuration The single-label setting was prepared so that we can compare the proposed approach with the state of the art. Similarly as the other studies, we follow the set-up proposed by Klementiev et al. [5]. Four main categories are used in this setting: *Corporate/industrial – CCAT, Economics – ECAT, Government/social – GCAT* and *Markets – MCAT.*

Documents containing more than one or zero main categories are filtered out. In total we randomly sample 15,000 documents for each language. 10,000 documents are used for training while the remaining 5,000 is reserved for testing.

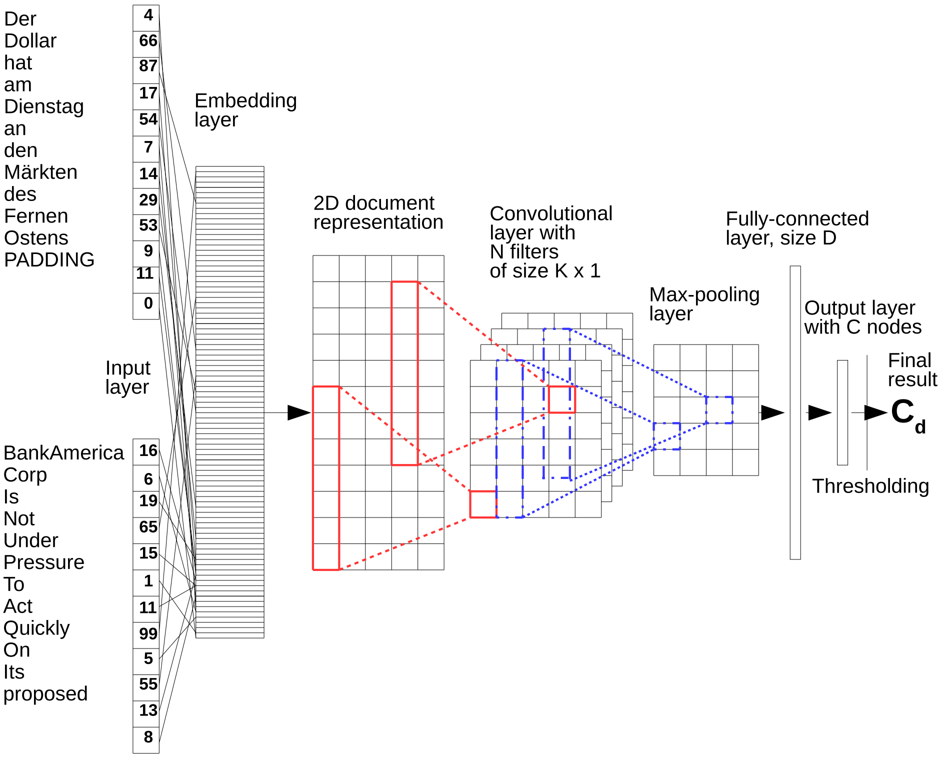

Fig. 1. The architecture of the CNN network used for multi-lingual classification. Two example documents are used as network input. Each document is handled in one training step.

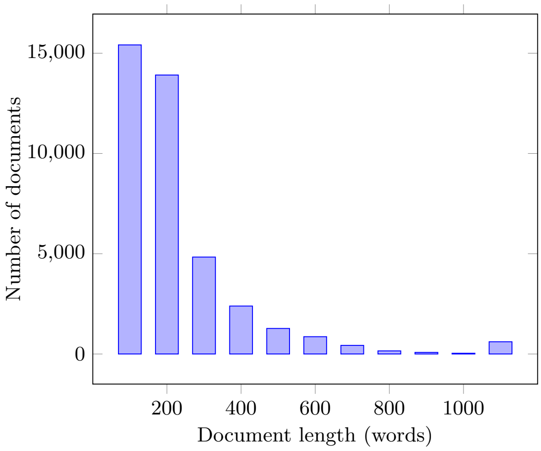

Fig. 2. Distribution of the document lengths in word tokens.

Multi-label Configuration In this setting we use all 103 topic codes available in the English documents. The number of documents for each language corresponds to the minimal number across the utilized languages which is Spanish in our case. Therefore we have 18,655 documents for each language where three fifths are used for training and the remaining two fifths for development and test set respectively.

4.2 Neural Network Set-up

In all experiments with multi-label classification we use the same configuration of the CNN. We use 20,000 most frequent words from each language to create the vocabulary. The document length is adjusted to $M = 100$ words with regard to the distribution of the document lengths according to Figure 2. The embedding length E is set to 300 which allows a direct usage of pre-trained word2vec vectors. The number of convolutional kernels N_C is 40 and its shape is set to 16×1. We use a *valid* mode for the convolutions. The number of neurons in the fully-connected layer is 256. Before the output layer and before the fully-connected one we add dropout layers with the probabilities set to 0.2 in both cases. Relu activation function is used in all layers except the output one. The output layer employs sigmoid function in the multi-label classification scenario. The model is optimized using Adaptive moment estimation (Adam) [4] algorithm and cross-entropy loss function.The data is shuffled in all experiments. We set the number of epochs to 10 in all experiments.

The single-label model is nearly the same as the multi-label one. The only difference is that softmax activation function is used in the output layer.

4.3 Single-label Results

Table 1 summarizes the results of the single-label classification experiments. We use the standard Precision ($Prec$), Recall (Rec), F-measure ($F1$) and Accuracy (ACC) metrics [13] and the confidence interval is ±0.3% at the confidence level of 0.95.

We present all three possible settings of the embedding layer. The first one uses static word2vec embeddings (*Word emb notrain*), the second one uses word2vec embeddings which are fine-tuned during the network training (*Word emb train*) and the last one uses randomly initialized vectors that are trained (*Random init*).

The results show that the training of the embeddings is beneficial and allows achieving significantly higher recognition scores. However, the usage of static pre-trained embeddings also reaches reasonable accuracy while dramatically lowering the time needed for the network training.

Table 2 compares the accuracies of the proposed methods with the state-of-the-art. As the other studies we use the standard accuracy metric in this experiment.

This table clearly shows that our methods outperform significantly all the other approaches. This is particularly evident in the case of English language

	Word emb notrain				**Word emb train**				**Random init**			
	Prec	**Rec**	**F1**	**ACC**	**Prec**	**Rec**	**F1**	**ACC**	**Prec**	**Rec**	**F1**	**ACC**
en	93.0	89.7	91.3	90.2	96.1	93.9	95.0	94.4	96.6	96.3	96.4	96.3
de	95.3	94.8	95.1	95.0	97.0	96.9	96.9	96.8	96.6	96.3	96.4	96.3
es	98.7	98.1	98.4	98.3	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9
it	88.8	86.7	87.8	86.9	91.9	91.6	91.7	90.7	91.5	91.2	91.3	90.6
avg	94.0	92.3	93.2	92.6	96.2	95.6	95.9	95.5	96.2	95.9	96.0	95.8

Table 1. Results of the single-label classification experiments [in %].

where the increase of accuracy is almost by 20%. We must note that the set-up of the other approaches slightly differ. However, the reported methods are the most similar set-ups we found. Moreover, to the best of our knowledge, there are no studies with exactly the same configuration as we use.

Method [ACC in %]	**de**	**en**
Klementiev et al. [5]	77.6	71.1
Kovčisky et al. [6]	83.1	76.0
Chandar A P et al. [1]	91.8	74.2
Coulmance et al. [2]	91.1	78.7
Word emb notrain	95.0	90.2
Word emb train	**96.8**	**94.4**
Random init	96.3	96.3

Table 2. Comparison with the state of the art [accuracy in %].

4.4 Multi-label Results

Table 3 shows the results of our network in the multi-label scenario. We use the standard Precision ($Prec$), Recall (Rec), F-measure ($F1$) metrics in this experiment. The confidence interval is ±0.35% at the confidence level of 0.95.

We can summarize the results in this table in a similar way as the previous one for the single-label classification. The training of the embeddings improves the obtained classification results. However, the training of randomly initialized vectors has worse results than the fine-tuned word2vec vectors. The best obtained F-measure 86.8% is, as in the previous case, for Spanish using word2vec initialized embeddings with a further training.

4.5 Word Similarity Experiment

The last experiment analyzes the quality of the resulting embeddings obtained by the three neural network settings.

	Word emb notrain			**Word emb train**			**Random init**		
	Prec	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
en	84.3	62.7	71.9	85.4	89.2	82.2	83.6	75.1	79.2
de	84.2	69.8	76.3	87.5	81.2	84.2	86.5	77.3	81.6
es	90.4	77.1	83.2	89.4	84.3	86.8	89.4	81.5	85.3
it	84.9	68.4	75.8	86.5	81.2	83.8	85.2	77.8	81.3
avg	86.0	69.5	76.8	87.2	81.5	84.3	86.2	77.9	81.9

Table 3. Precision (*Prec*), Recall (*Rec*), F-measure (*F*1) of the multi-label classification [in %].

Table 4 shows 10 most similar words to the English word "accident" across all languages based on the cosine similarity. These words are mainly in English when word2vec initialization without any training is used (the first column). Further training of the embeddings (middle column) causes that also German and Spanish words with a similar meaning are shifted closer to the word "accident" in the embedding space. On the other hand, when training from randomly initialized vectors, the ten most similar words have often quite a different meaning. However, as shown in the classification results, this fact has nearly no impact on the resulting F-measure. We can conclude that word2vec initialization is not necessary for the classification task. This table further shows that the similarity between Germanic (English and German) languages is clearly visible.

Word emb notrain		**Word emb train**		**Random init**	
word	**cos sim**	**word**	**cos sim**	**word**	**cos sim**
accidents	0.860	accidente	0.685	ruehe	0.248
incident	0.740	unglück (*de*, misfortune)	0.632	bloccando (*es*, blocking)	0.239
accidente (*es*, accident)	0.600	estrelló (*es*, crashed)	0.609	compelled	0.236
incidents	0.574	accidents	0.599	numerick	0.219
accidentes (*es*, accidents)	0.546	geborgen (*de*, secure)	0.585	fiduciary	0.217
disaster	0.471	absturz	0.584	barriles (*es*, barrels)	0.216
explosions	0.461	unglücks (*de*, misfortunes)	0.576	andhra	0.214
incidence	0.452	abgestürzt (*de*, crashed)	0.567	touring	0.212
personnel	0.452	trümmern (*de*, rubble)	0.560	versicherers (*de*, insurers)	0.209
unfall (*de*, accident)	0.450	unglücksursache (*de*, ill cause)	0.551	oppositioneller (*de*, oppositional)	0.203

Table 4. Ten closest words to the English word "accident" based on the cosine similarity; English translation in brackets including the language of the given word.

Table 5 shows 10 most similar words to the English word "czech" using the cosine similarity. The table is very similar to the previous one. For instance, if we take a look at the *Word emb train* column, we observe that there is (as in the previous case) a significant decrease of the cosine similarity. However on the other hand, some new words, which are more related to the word "czech", are included. The inapplicability to find similar words of randomly initialized embeddings has been confirmed. It is worth noting that although the Czech language is not a part of our corpus, some Czech words (*praha, dnes, fronta*) are also included due to the Czech citations available.

Word emb notrain		**Word emb train**		**Random init**	
word	**cos sim**	**word**	**cos sim**	**word**	**cos sim**
czechoslovakia	0.757	czechoslovakia	0.399	festakt (*de*, ceremony)	0.273
slovakia	0.634	praga (*es*, prague)	0.335	val	0.250
polish	0.569	republic	0.329	provence	0.235
hungary	0.539	brno (*cz*, brno - czech city)	0.315	sostiene (*es*, hold)	0.222
hungarian	0.537	slovak	0.314	larry	0.216
prague	0.533	praha (*cz*, prague)	0.313	köpfigen (*de*, headed)	0.212
slovak	0.509	dnes (*cz*, today)	0.307	überschreiten (*de*, exceed)	0.206
praha (*cz*, praha)	0.509	checa (*es*, czech)	0.307	aktienindex (*de*, share index)	0.205
austrian	0.506	fronta (*cz*, queue)	0.304	councils	0.205
lithuanian	0.496	tschechoslowakei (*de*, czechoslovakia)	0.297	bancario (*it*, banking)	0.205

Table 5. Ten closest words to the word "czech" based on the cosine similarity; English translation in brackets including the language of the given word.

5 Conclusions

In this paper we presented a novel approach for the multi-label document classification in multiple languages. The proposed method builds on the popular convolutional networks. We added a simple yet efficient extension that allows using one network for classifying text documents in more languages.

We evaluated our method on four languages from the Reuters corpus in both multi- and single-label classification scenarios. We showed that the proposed approach is efficient and the best obtained F-measure in multi-label scenario reaches 84%. We also showed that our methods outperform significantly in the single-label settings all the other approaches. Another added value of this approach is also that no language identification is needed as in the case of the use of the single networks.

Acknowledgements

This work has been partly supported from ERDF "Research and Development of Intelligent Components of Advanced Technologies for the Pilsen Metropolitan Area (InteCom)" (no.: CZ.02.1.01/0.0/0.0/17_048/0007267), by Cross-border Cooperation Program Czech Republic - Free State of Bavaria ETS Objective 2014-2020 (project no. 211) and by Grant No. SGS-2016-018 Data and Software Engineering for Advanced Applications.

References

1. Chandar A P, S., Lauly, S., Larochelle, H., Khapra, M., Ravindran, B., Raykar, V.C., Saha, A.: An autoencoder approach to learning bilingual word representations. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 27, pp. 1853–1861. Curran Associates, Inc. (2014)
2. Coulmance, J., Marty, J.M., Wenzek, G., Benhalloum, A.: Trans-gram, fast cross-lingual word-embeddings. arXiv preprint arXiv:1601.02502 (2016)
3. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014)
4. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
5. Klementiev, A., Titov, I., Bhattarai, B.: Inducing crosslingual distributed representations of words. Proceedings of COLING 2012 pp. 1459–1474 (2012)
6. Kočiský, T., Hermann, K.M., Blunsom, P.: Learning bilingual word representations by marginalizing alignments. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). vol. 2, pp. 224–229 (2014)
7. Kurata, G., Xiang, B., Zhou, B.: Improved neural network-based multi-label classification with better initialization leveraging label co-occurrence. In: Proceedings of NAACL-HLT. pp. 521–526 (2016)
8. Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents. In: ICML. vol. 14, pp. 1188–1196 (2014)
9. Lenc, L., Král, P.: Deep neural networks for Czech multi-label document classification. In: Gelbukh, A. (ed.) Computational Linguistics and Intelligent Text Processing. pp. 460–471. Springer International Publishing, Cham (2018)
10. Lewis, D.D., Yang, Y., Rose, T.G., Li, F.: Rcv1: A new benchmark collection for text categorization research. Journal of machine learning research 5(Apr), 361–397 (2004)
11. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th international conference on machine learning (ICML-10). pp. 807–814 (2010)
12. Nam, J., Kim, J., Mencía, E.L., Gurevych, I., Fürnkranz, J.: Large-scale multi-label text classification - revisiting neural networks. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 437–452. Springer (2014)
13. Powers, D.: Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. Journal of Machine Learning Technologies 2(1), 37–63 (2011)
14. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research 15(1), 1929–1958 (2014)
15. Yang, Y., Gopal, S.: Multilabel classification with meta-level features in a learning-to-rank framework. Machine Learning 88(1-2), 47–68 (2012)
16. Zhang, M.L., Zhou, Z.H.: Multilabel neural networks with applications to functional genomics and text categorization. Knowledge and Data Engineering, IEEE Transactions on 18(10), 1338–1351 (2006)
17. Zou, W.Y., Socher, R., Cer, D., Manning, C.D.: Bilingual word embeddings for phrase-based machine translation. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. pp. 1393–1398 (2013)