ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

Using Semantic Similarity for Multi-Label Zero-Shot Classification of Text Documents

Sappadla Prateek Veeranna[1] Jinseok Nam[2,3]
Eneldo Loza Mencía[2] Johannes Fürnkranz[2] *

1- Birla Institute of Technology and Science - Pilani - India
2- Knowledge Engineering Group - TU Darmstadt - Germany
3- Knowledge Discovery in Scientific Literature - DIPF - Germany

Abstract. In this paper, we examine a simple approach to zero-shot multi-label text classification, i.e., to the problem of predicting multiple, possibly previously unseen labels for a document. In particular, we propose to use a semantic embedding of label and document words and base the prediction of previously unseen labels on the similarity between the label name and the document words in this embedding. Experiments on three textual datasets across various domains show that even such a simple technique yields considerable performance improvements over a simple uninformed baseline.

1 Introduction

Multi-label text classification is the problem of assigning multiple labels or keywords to a given document. In many such tasks, such as news classification, the number of possible categories is unlimited and dynamically changing so that new categories will be added over time. In such cases we may not have training data for the newly introduced categories, but may still want to be able to classify documents into these categories. *Zero-shot learning* attempts to deal with such situations.

In this paper, we investigate simple unsupervised algorithms that classify documents without the use of any training data for learning the classifier. These unsupervised algorithms can then be used in the zero-shot setting for predicting the unseen labels, while the seen labels are predicted using existing state-of-the-art supervised approaches. The key idea is to base these predictions on the semantic similarity of the document text to the label, which can be computed based on semantic word embeddings.

The paper is organised as follows. Section 2 gives a brief introduction into the problems of multi-label text classification and zero-shot learning. Section 3 describes the similarity-based algorithms that we study in this paper in more detail. Section 5 shows the results of their experimental evaluation before we conclude in Section 6.

2 Preliminaries

A text document $\mathbf{x}_i$ is given as a sequence of words $\langle w_i^1, w_i^2, \ldots, w_i^{|\mathbf{x}_i|} \rangle$ from a vocabulary $w_i^j \in \mathcal{V} = \{1, 2, \ldots, |\mathcal{V}|\}$ of words. In *multi-label classification* (MLC),

*Parts of this work have been supported by the German Institute for Educational Research (DIPF) under the Knowledge Discovery in Scientific Literature (KDSL) program, and the German Research Foundation as part of the Research Training Group *Adaptive Preparation of Information from Heterogeneous Sources* (AIPHES) under grant No. GRK 1994/1.

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

each document $\mathbf{x}_i$ is associated with a subset of labels $\mathcal{y}_i \subseteq \mathcal{Y}$ of the set of possible labels $\mathcal{Y} = \{1, 2, \ldots\}$. The task is to learn these associations based on a training set $T_n = \{(\mathbf{x}_1, \mathcal{y}_1), (\mathbf{x}_1, \mathcal{y}_i), \ldots, (\mathbf{x}_n, \mathcal{y}_n)\}$ in order to predict the unknown label sets on a test set $T_m = \{(\mathbf{x}_{n+1}, \mathcal{y}_{n+1}), \ldots, (\mathbf{x}_{n+m}, \mathcal{y}_{n+m})\}$. Usually, it is assumed that the set of possible labels $\mathcal{Y}$ is fixed and that all labels in the test set already appeared at least once in the training set, i.e., $\mathcal{Y}_k = \bigcup_{i=1}^{n} \mathcal{y}_i, \mathcal{Y}_m = \bigcup_{i=n+1}^{n+m} \mathcal{y}_i, \mathcal{Y} = \mathcal{Y}_k \supset \mathcal{Y}_m$.

There are many learning algorithms for multi-label classification [3, 10]. In this paper, since our concern is on predicting unknown labels for which existing approaches do not offer solutions, we use the very straight-forward binary relevance decomposition (BR) of the original problem as our baseline multi-label method. Essentially, BR learns a separate classifier for each label, which predicts whether the label should or should not be predicted for a given instance $\mathbf{x}$.

In *zero-shot learning* (ZSL), it is assumed that new, unknown labels may appear during query-time, i.e., the assumption $\mathcal{Y}_m \subset \mathcal{Y}_k$ no longer holds, and a non-empty set of *unknown labels* $\mathcal{Y}_u = \mathcal{Y}_m \backslash \mathcal{Y}_k$ exists. Moreover, we assume in our particular setting that each label $l \in \mathcal{Y}$ has a name given by a sequence of words $\lambda_l = \langle w_{\lambda_l}^1, \ldots, w_{\lambda_l}^{|\lambda_l|} \rangle$.

ZSL has so far been primarily investigated in computer vision. A common solution to ZSL is to represent labels by attributes which are shared by known as well as unknown labels [1, 8]. For instance, Lampert et al. [4] use semantic features of visual objects in a visual object classification task. Recent approaches for object recognition also use textual information such as the labels' names. For example, Frome et al. [2] learn d-dimensional word representation from large textual corpora such as Wikipedia and then use the same embedding space for representing the images. Recently, Nam et al. [7] proposed an approach for text classification which produces a joint embedding of words, documents, labels and associated (longer) label descriptions.

A common limitation for most techniques relying on (joint) embeddings is the large amount of training data necessary. In contrast, in this paper we investigate the potential of simple techniques that solely rely on estimating the semantic similarity between a label and the given documents. Thus, these techniques do not use any training documents at all, and may also be used to complement any given multi-label text classification algorithm.

3 Similarity-Based Zero-Shot Prediction

Label Presence. The most straight-forward approach to predict a label l as relevant given a document $\mathbf{x}_i$ is to check whether the label name λ_l appears in the document. More formally, we include l in the set of predicted labels $\hat{y}$ for $\mathbf{x}_i$ if for any $j = 1, \ldots, |\mathbf{x}_i| - |\lambda_l|$ it holds that

$$\langle w_i^j, w_i^{j+1}, \ldots, w_i^{j+|\lambda_l|} \rangle = \langle w_{\lambda_l}^1, \ldots, w_{\lambda_l}^{|\lambda_l|} \rangle \tag{1}$$

This simple unsupervised approach may work well in the case when the labels are very specific and rather short entities (e.g., "*iphone*", "*European Investment Bank*"). In such cases, the documents will generally contain the label in the text, and ideally not many non-relevant document would be covered. However, if the labels are very general, e.g., "*dollar*", "*cough*", it is very likely that they appear in too many documents.

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

Moreover, for very long and descriptive label names such as "*fever and other physiologic disturbances of temperature regulation*" it is unlikely that they will occur exactly in this form in the text of the relevant documents (label names from datasets in Sec. 4).

Label Word Similarity. To circumvent this problem, we propose to relax the strict equation in (1) and substitute it by a more general formulation making use of a textual similarity measure. We set label l as true if

$$t \leq \max_{1 \leq c \leq c_{\max}} \max_{1 \leq j \leq |\mathbf{x}_i| - c} \sigma(\langle w_i^j, w_i^{j+1}, \ldots, w_i^{j+c} \rangle, \langle w_{\lambda_l}^1, \ldots, w_{\lambda_l}^{|\lambda_l|} \rangle) \qquad (2)$$

for a user-defined threshold t and a maximum window or n-gram size $c_{\max}$. Obviously, (1) would be obtained by setting $c_{\max} = |\lambda_l|$, $t = 1$, and $\sigma(\mathbf{w}_1, \mathbf{w}_2) = 1$ if $\mathbf{w}_1 = \mathbf{w}_2$ otherwise 0.

Semantic Similarity. Obviously, the performance of (2) highly depends on the proper choice of the similarity function. Recently, methods such as skip-gram [6], which compute low-dimensional continuous vector representations of words and phrases based on word co-occurrences in large corpora, became very popular. These so-called word embeddings are capable of capturing syntactic and semantic characteristics of the words and were shown to be very effective on word similarity tasks. Moreover, it is suggested that a non-obvious degree of language understanding can be obtained by using basic mathematical operations on the word vector space. For instance, it was found that relations such as $vec(\text{"}\textit{Germany}\text{"}) + vec(\text{"}\textit{capital}\text{"}) \approx vec(\text{"}\textit{Berlin}\text{"})$ hold very often. Relying on these compositionality characteristic of word embeddings, we obtain the embedding of a label by adding up the embedding of their constituent words. We also include compound words such as "*New York*" in our vocabulary (cf. Sec. 4) instead of using the composed variant. The similarity between two word sequences is then given by the Cosine similarity $\sigma(\mathbf{w}_1, \mathbf{w}_2) = \frac{vec(\mathbf{w}_1)^T \cdot vec(\mathbf{w}_2)}{|vec(\mathbf{w}_1)| \cdot |vec(\mathbf{w}_2)|}$ with $\sigma(\mathbf{w}) \in \mathbb{R}^d$ as either the embedding of a compound word or the sum of the embeddings of the words in $\mathbf{w}$, and with d as the dimensionality of the embedding space.

Combination with Supervised Multi-Label Algorithms. The above algorithms are meant to complement conventional supervised multi-label algorithms. For the purpose of the evaluation in this paper, the simple but effective BR technique (Sec. 2) is used for making the predictions of the known labels, whereas the techniques discussed above are used for making predictions for unknown labels (i.e., labels that have not been seen in training).

4 Experimental Setup

We had access to the full texts of three MLC datasets: The REUTERS corpus[1] of news wire articles was split into 7769 training and 3019 test examples and contains 90 labels

[1] `http://www.daviddlewis.com/resources/testcollections/reuters21578/readme.txt`

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

Table 1: Zero-shot results of the proposed method on the three datasets.

Dataset	REUTERS				MEDICAL				EURLEX			
Threshold	0.7	0.8	0.9	1	0.7	0.8	0.9	1	0.7	0.8	0.9	1
Ma. Prec.	.2228	.3713	.4708	.4706	.0494	.1513	.2244	.2031	.0368	.0624	.1020	.1748
Ma. Rec.	.6654	.5633	.5181	.5169	.7842	.5205	.2178	.2201	.7520	.6973	.5191	.2803
Ma. F1	.2667	.3805	.4378	.4365	.0816	.1826	.1980	.2102	.0577	.0826	.1176	.1886

with an average of 1.23 labels assigned to each document. Acronym label names were expanded, such as "*bop*" to "*balance of payments*". MEDICAL consists of 1953 radiology reports associated to 45 diagnoses ($\sim$1.24 per doc.) [9]. The EURLEX [5] is a collection of legal documents about the European Union associated to 201 *subject matters* ($\sim$2.21 per doc.). All used texts were converted to lowercase and tokenized. Numbers were converted to the symbol "0". For the MLC datasets, we additionally removed stop words. For MEDICAL and EURLEX, label names range from short and general ("*cough*", "*coffee*') to long and specific ("*fever and other physiologic disturbances of temperature regulation*", "*quantitative restrictions and measures of equivalent effect*").

The skip-gram word embeddings used for our experiments have been learnt with word2vec[2] using the English Wikipedia as the raw text corpus. The final vocabulary consisted of $\sim$2.5 million entries. The BR baseline was trained with LibLinear[3] on TF-IDF bag-of-words vectors. We used the original train-test split for REUTERS and three-fold and ten-fold cross validation for MEDICAL and EURLEX, respectively. Parameters $t = 0.9$ and $c_{\max} = 3$ were used for REUTERS and MEDICAL and $c_{\max} = 1$ for EURLEX.

For evaluating the different approaches we focus on the macro-averaged variants of recall, precision and the F1-measure [cf. 10]. These metrics give equal weight to all labels whereas micro-averaged measures are dominated by the most frequent labels. As unseen labels generally are rare labels, we consider the macro-averaged measures to be a better indicator for the performance in the analyzed setting. More specifically, we compute the measures as

$$\text{Ma. Prec.: } \frac{1}{|\mathcal{Y}|}\sum_{l=1}^{|\mathcal{Y}|}\frac{\sum_{i=n+1}^{m}|y_i \cap \hat{y}_i \cap l|}{\sum_{i=n+1}^{m}|\hat{y}_i \cap l|} \qquad \text{Ma. Rec.: } \frac{1}{|\mathcal{Y}|}\sum_{l=1}^{|\mathcal{Y}|}\frac{\sum_{i=n+1}^{m}|y_i \cap \hat{y}_i \cap l|}{\sum_{i=n+1}^{m}|y_i \cap l|}$$

Macro F1 uses the recall and precision values for each label to compute the harmonic means.

5 Experimental Results

In our experiments we analyze the case where all labels are known (MLC) to the case were all labels are unknown during classification (full ZSL). More specifically, we ordered the labels by their ascending frequency and switched their training information off one by one. For predicting such ignored labels, we use the output of our proposed ZSL approach. We compare to the case where the learned classifier predicts these labels as false since no positive examples were ever seen.

[2] `https://code.google.com/p/word2vec/`. Dimensionality of 300, hierarchical softmax, negative sampling, window size of 10 and probability of 10^{-4} of subsampling frequent words was used. Phrases were considered as one token if they appeared as wiki links in the raw text.

[3] `https://www.csie.ntu.edu.tw/~cjlin/liblinear/`. Default parameters were used.

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

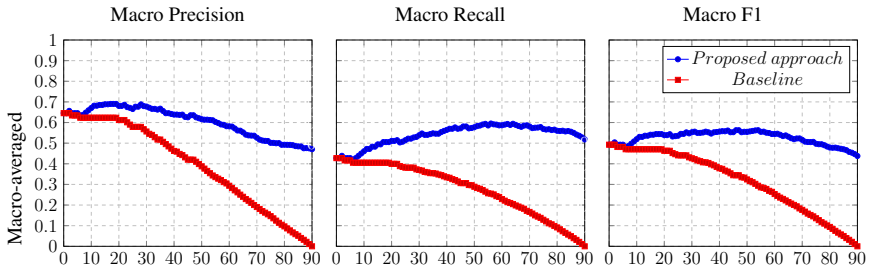

Fig. 1: Removal of labels for REUTERS. The x-axis indicates the number of switched off labels, the y-axis the corresponding measure.

Table 1 shows the ZSL performance on the three datasets w.r.t. to the chosen threshold parameter. As expected, precision generally decreases and recall increases with lower thresholds. However, there are some exceptions and remarkable differences between the datasets. For instance, macro precision drops for MEDICAL for $t = 1$. This can be attributed to the fact that MEDICAL contains many infrequent labels with long names, denoting very specific diagnosis such as "*spina bifida without mention of hydrocephalus, unspecified region*".

The results for REUTERS are shown in Fig. 1. We can see that our proposed method of replacing the predictions clearly outperforms ignoring the labels for any number of affected labels on the macro-averaged metrics. Interestingly, the highest values for precision, recall and F1 are found when a considerable number of rare labels are missing. This result suggests that it might be more beneficial in some cases to abstain from learning models for some rare labels even if training data is available.

The results in micro-averaged precision (not shown) are quite similar except that here, as expected, precision values values suffer considerably when the most common labels are removed, even more so for the proposed approach. However, the advantage in recall is able to make up and result in an overall gain in terms of F1.

The results for MEDICAL and EURLEX are quite similar, we only show the macro-averaged F1 values in Fig. 2. Again, we can observe a consistent improvement of the proposed technique over the baseline on these datasets, too, although it is not as pronounced as for REUTERS. This can be attributed to having longer and more complex labels as compared to the REUTERS dataset (cf. Sec. 4). These long labels are generally not part of the vocabulary and their vectors have to be created by composition.

We obtained similar results if we remove the labels in descending frequency. In fact, the curves start from the same point at $x = 0$ but are rotated by 180° since label-wise metrics are removed from the average in reversed order (not shown due to space constraints).

6 Conclusion

In this paper, we investigated a simple and straight-forward approach for complementing conventional multi-label text classification algorithms with the capability of making

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

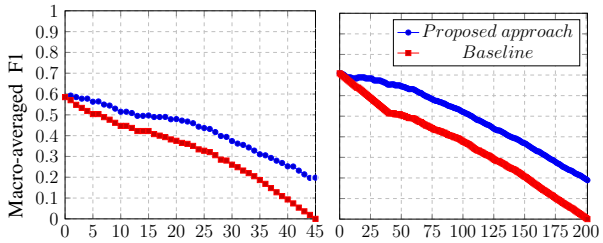

Fig. 2: Removal of labels for MEDICAL (left) and EURLEX (right).

predictions on labels that have been unseen during training time. The idea is to use word embeddings to compute a semantic similarity between the label and the document text. Although the approach is quite limited and cannot be expected to yield the same performance as a supervised approach, the results show that despite its simplicity, it can successfully complement multi-label classifiers with such functionality without additional training.

References

[1] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1778–1785, 2009.

[2] A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov. Devise: A deep visual-semantic embedding model. *Advances in Neural Information Processing Systems 26*, pp. 2121–2129. 2013.

[3] E. Gibaja and S. Ventura. Multi-label learning: a review of the state of the art and ongoing research. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 4(6):411–444, 2014.

[4] C. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2014.

[5] E. Loza Mencía and J. Fürnkranz. Efficient multilabel classification algorithms for large-scale problems in the legal domain. In *Semantic Processing of Legal Texts* , pp. 192–215. Springer-Verlag, 2010. `http://www.ke.tu-darmstadt.de/resources/eurlex`.

[6] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems 26*, pp. 3111–3119. Curran Associates, Inc., 2013.

[7] J. Nam, E. Loza Mencía, and J. Fürnkranz. All-in text: Learning document, label, and word representations jointly. In *Proc. 30th AAAI Conference on Artificial Intelligence*, 2016. To appear.

[8] M. Palatucci, D. Pomerleau, G. Hinton, and T. Mitchell. Zero-shot learning with semantic output codes. *Advances in Neural Information Processing Systems*, pp. 1410–1418. 2009.

[9] J. Pestian, C. Brew, P. Matykiewicz, D. Hovermale, N. Johnson, K. Bretonnel Cohen, and W. Duch. A shared task involving multi-label classification of clinical free text. In *Proceedings of the Workshop on BioNLP 2007 at ACL 2007*, 2007.

[10] G. Tsoumakas, I. Katakis, and I. Vlahavas. Mining multi-label data. In *Data Mining and Knowledge Discovery Handbook*, pp. 667–685. Springer, 2010.