Semantic Space Transformations for Cross-lingual Document Classification

Jiří Martínek[1], Ladislav Lenc[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,llenc,pkral}@kiv.zcu.cz

Abstract. Cross-lingual document representation can be done by training monolingual semantic spaces and then to use bilingual dictionaries with some transform method to project word vectors into a unified space. The main goal of this paper consists in evaluation of three promising transform methods on cross-lingual document classification task. We also propose, evaluate and compare two cross-lingual document classification approaches. We use popular convolutional neural network (CNN) and compare its performance with a standard maximum entropy classifier. The proposed methods are evaluated on four languages, namely English, German, Spanish and Italian from the Reuters corpus. We demonstrate that the results of all transformation methods are close to each other, however the orthogonal transformation gives generally slightly better results when CNN with trained embeddings is used. The experimental results also show that convolutional network achieves better results than maximum entropy classifier. We further show that the proposed methods are competitive with the state of the art.

1 Introduction

The performance of many Natural Language Processing (NLP) systems is strongly dependent on the size and quality of annotated resources. Unfortunately, there is a lack of annotated data for particular languages / tasks and manual annotation of new corpora is a very expensive and time consuming task. Moreover, the linguistic experts from the target domain are often required. These issues can be solved by the usage of cross-lingual text representation methods. The classifiers are trained on resource-rich languages and the cross-linguality allows using the models with data in other languages with no available training data.

The text document representations are often created using multi-dimensional word vectors, often so called word embeddings (Levy and Goldberg [10]). One way of creating cross-lingual representations is to use transformed semantic spaces. Such approaches take a monolingual, independently trained, semantic space and project it into a unified space using some transformation method.

Several such transformation methods have been proposed. However, to the best of our knowledge, a comparative study of the role of different transformation methods / classifiers for the document classification across several languages is missing. Therefore, the main contribution of this paper consists in the thorough study of the impact of three promising transform methods, namely Least Squares Transformation (LST), Orthogonal Transformation (OT) and Canonical Correlation Analysis (CCA), for cross-lingual document classification. More information about linear transformations to build cross-lingual semantic spaces can be found in [2, 3]. In this context, we propose, evaluate and compare two cross-lingual document classification approaches. The first one uses directly the transformed embeddings in different languages while the second one realizes a simple word translation by choosing the closest word using cosine similarity of the embedding vectors.

For classification, we use popular convolutional neural network (CNN) and compare its performance with a standard maximum entropy classifier. The proposed methods are evaluated on four languages, namely English, German, Spanish and Italian from the Reuters corpus.

2 Literature Review

Recent work in cross-lingual text representation field is usually based on word-level alignments. Klementiev et al. [7] 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. [13] propose an alternative approach based on another neural network language models using different regularization.

Kočiský et al. [8] propose a bilingual word representation approach based on a probabilistic model. This method simultaneously learns alignments and distributed representations for bilingual data. Contrary to the prior work, which is based on parallel corpora or hard alignment, this method marginalizes out the alignments, thus captures a larger bilingual semantic context.

Chandar et al. [4] 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. [5] introduce an efficient method for bilingual word representations called Trans-gram. This approach extends popular skip-gram model to multilingual 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 Cross-lingual Document Classification

3.1 Document Representation

We use three document representations in our experiments. The first one is the Bag-of-Words (BoW). The second approach called *averaged embeddings* utilizes word embeddings. It averages the word vectors for all words occurring in the document. Its length corresponds to the embeddings dimensionality. The last method uses the sequence of words in the document and transforms it to the 2D representation suitable for the CNN. The words are one-hot encoded and are translated using a look-up table by the corresponding embeddings. Further we describe the three ways how we achieve the cross-linguality in our classification methods.

Machine Translation Machine translation (MT) is used as a strong baseline for comparison with the two other methods. The documents are translated using Google API. The translation is then used in the same way as if classifying documents in one language.

Transformed Embeddings This approach relies on the transformed word embeddings. The representations of the training documents are created from the original word embeddings in the language, which was used for the training of the model. The documents in the testing dataset are then represented by the embeddings transformed to the language of the model. This method will be hereafter called *transformed (emb)eddings*.

Embedding Translation This method is also based on the transformed embeddings. However, the embeddings are used for *per-word* translation of the documents instead of using it directly. It utilizes the non-transformed embedding in the target language and the transformed one from source to target language for similarity search. The most similar word in the target language is found for each word in the source language by *cosine similarity*. This method is in the following text referred as *(emb)edding translation* approach.

3.2 Classification Models

Maximum Entropy The first classifier is the Maximum Entropy (ME) model Berger et al. [1]. It takes for each document an input with a fixed number of features, represented as a feature vector F, and outputs the probability distribution $P(Y = y|F)$ where $y \in \mathcal{C}$ (set of all possible document classes). This model is popular in the natural language processing field, because it usually gives good classification scores.

Convolutional Neural Network The second classifier is a popular Convolutional Neural Network (CNN). It also outputs normalized scores interpreted as a probability distribution $P(Y = y|F)$ over all possible labels. The network we use was proposed by Lenc and Král [9] and it was successfully used for multi-label classification of Czech documents. The architecture of the network is inspired by Kim [6]. The main difference from Kim's network is that this net uses different number and size of convolutional kernels.

We perform a basic preprocessing which detects all numbers and replaces them by one "NUMERIC" token. Then the document length is adjusted to a fixed value. Longer documents are shortened while shorter ones are padded so that they have fixed length L.

A vocabulary of the most frequent words is prepared from the training data. The words are then represented by their indexes in the vocabulary. The words that are not in the vocabulary are assigned to a reserved index ("OOV") and the "PADDING" token has also a reserved index.

The input of the network is a vector of word indexes of the length L where L is the number of words used for document representation. The second layer is an embedding layer which represents each input word as a vector of a given length. The document is thus represented as a matrix with L rows and E columns where E is the embeddings dimensionality. The embedding layer can be initialized either randomly and trained during the network training process or use the pre-trained word embeddings as its weights. The third layer is the convolutional one. N convolutional kernels of the size $K \times 1$ are used which means that a 1D convolution over one position in the embedding vector over K input words is performed. The following layer performs max pooling over the length $L - K + 1$ resulting in N $1 \times E$ vectors. This layer is followed by a dropout layer Srivastava et al. [12] for regularization. The output of this layer is then flattened and connected with a fully connected layer with D neurons. After another dropout layer follows the output layer with C neurons which corresponds to the number of the document categories. The architecture of the network is depicted in Figure 1.

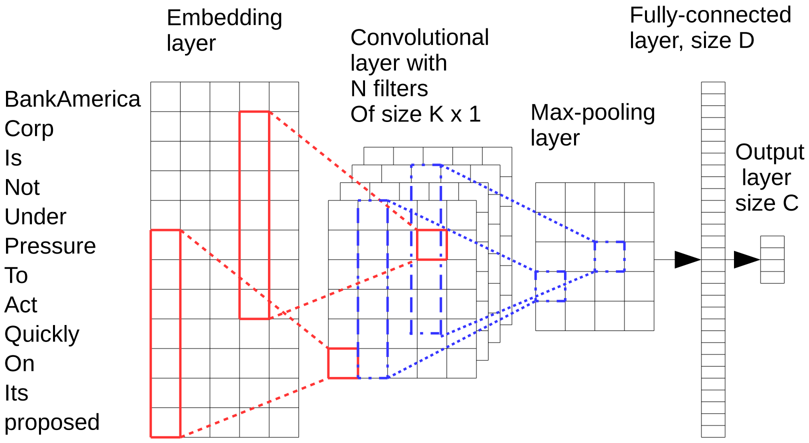

Fig. 1. Convolutional neural network architecture.

4 Experiments

4.1 Reuters Corpus Volume I

We use four languages, namely English (en), German (de), Spanish (es) and Italian (it) from Reuters Corpus Volume I (RCV1-v2) Lewis et al. [11] with similar setup as used by Klementiev et al. [7]. The documents are classified into four following categories: *Corporate/industrial – CCAT, Economics – ECAT, Government/social – GCAT* and *Markets – MCAT*.

As the other studies we use the standard accuracy metric in our experiments. The confidence interval is $\pm 0.3\%$ at the confidence level of 0.95.

4.2 Baseline Approaches Results

Our first baseline method is a *majority class (MC)* classifier which determines the distribution of categories in the training dataset and chooses the most frequent class. In testing phase, all test documents are classified into this most frequent class. The accuracy of this classifier is depicted in third column of Table 1. These results show that the corpus is unbalanced and that there are significant differences among different languages.

The second baseline is the *machine translation (MT)* approach. The results with the ME classifier are reported in Table 1, while the accuracy of the CNN is shown in Table 2 (column *MT*). Classification accuracies of this approach are very high and show that the translation results have a strong impact on document classification.

4.3 Proposed Approaches Results

The embedding translation approach needs repeatedly searching the target semantic space which is computationally demanding. In order to reduce the computational burden we set the vocabulary size $|\boldsymbol{V}| = 20,000$. The vocabulary is constructed from the most frequent words in the training set. To increase efficiency of searching, we created vocabulary mapping dictionary between each pair of languages. There is a mapping onto target language vocabulary for each word in the source language. This dictionary is the centerpiece of the embedding translation. If the source word is not present in the vocabulary, the out of vocabulary token (“OOV”) is used. Each proposed method is experimentally validated on two classification models.

Maximum Entropy Results The last six columns in Table 1 show the results of the maximum entropy classifier with *transformed (emb)eddings* and *(emb)edding translation* methods. Three linear transformations are used.

This table shows that the grammatically close languages (same family) give usually better results than the other ones. More concrete, $en \leftrightarrow de$ and $es \leftrightarrow it$ have generally better results than for instance $en \leftrightarrow es\ (it)$ or $de \rightarrow es\ (it)$.

The results further show, that the three transformation are comparable in many cases. However, LST gives the best results in several other cases (e.g. $en \rightarrow es\ (it)$ or $de \rightarrow es\ (it)$) and OT gives the significantly worst results for some cases (e.g. $it \rightarrow en$). Based on this experiment we can propose generally to use LST with ME classifier and static word embeddings.

		Baselines [%]		Proposed Approaches [%]					
Languages		MC	MT	transformed emb			emb translation		
Train	Test			LST	CCA	OT	LST	CCA	OT
en	de	30.4	91.9	54.4	62.0	52.2	75.6	75.8	76.3
en	es	14.7	81.5	52.9	39.0	49.9	57.2	48.8	49.8
en	it	36.0	71.2	48.4	42.2	56.0	58.1	51.0	51.5
de	en	23.9	76.7	59.9	60.4	57.4	66.6	69.0	70.2
de	es	8.76	81.1	35.9	32.5	29.4	73.2	58.5	63.8
de	it	9.50	67.0	58.7	57.5	57.0	47.2	47.7	46.4
es	en	23.3	74.3	47.8	53.2	45.4	67.3	70.5	69.6
es	de	22.6	85.7	51.7	43.8	47.6	74.5	70.1	70.3
es	it	36.4	67.9	22.8	19.8	29.7	71.2	72.0	72.1
it	en	23.3	69.7	68.5	69.8	63.8	56.8	55.6	50.2
it	de	22.6	86.9	48.7	45.1	52.6	76.6	77.4	76.8
it	es	67.7	80.8	61.4	49.4	59.5	75.3	75.2	70.5

Table 1. ME classifier results. Columns 3 and 4 represent the majority class (MC) and machine translation (MT) baselines. The rest of the table shows results for the proposed methods with different embedding transformations.

CNN Results In all our experiments we use the vocabulary size $|V| = 20,000$. The document length L is set to 100 tokens and the embedding length E is 300 in all cases. We use $N = 40$ convolutional kernels of size 16×1 ($K = 16$). The dropout probability is set to 0.2. The size of the first fully connected layer is 256. The output layer has 4 neurons ($C = 4$) while we are classifying into 4 classes. All layers except the output one use *relu* activation function. The output layer uses the *softmax* activation function.

		Baselines [%]		Proposed Approaches [%]								
Languages		MT		transformed emb			emb translation (stat)			emb translation (rnd)		
Train	Test	rnd	stat	LST	CCA	OT	LST	CCA	OT	LST	CCA	OT
en	de	89.7	86.5	62.2	64.4	56.0	78.9	79.1	81.3	80.4	80.4	82.7
en	es	85.7	69.8	24.4	26.7	23.6	82.0	77.0	76.9	81.9	75.7	72.9
en	it	74.7	65.8	27.7	26.9	18.0	68.1	70.2	68.6	71.1	70.5	68.3
de	en	61.3	59.4	66.4	64.8	58.4	69.3	70.0	70.2	72.3	75.6	75.6
de	es	64.7	55.7	65.0	57.6	55.2	51.0	55.3	54.5	81.4	79.3	80.7
de	it	47.8	48.8	39.1	50.9	58.4	44.8	48.6	49.2	68.7	71.1	71.2
es	en	60.7	67.6	49.2	41.1	41.5	51.9	54.9	55.0	59.0	63.1	63.0
es	de	76.8	81.8	42.9	54.1	58.0	58.5	72.2	81.5	54.7	69.2	82.0
es	it	62.4	61.9	20.2	35.5	37.5	68.0	70.9	71.5	73.0	76.2	76.7
it	en	69.1	65.7	42.5	44.8	46.4	41.4	41.8	41.1	54.8	54.7	51.3
it	de	85.4	81.5	37.0	38.3	44.5	59.6	72.7	74.1	63.0	76.0	60.8
it	es	80.1	73.8	68.5	68.8	68.1	61.3	61.3	62.1	78.6	78.6	78.7

Table 2. CNN results. Columns 3 and 4 represent the MT baseline. The rest of the table presents result of the proposed methods with different embedding transformations. Term *stat* means the static word embeddings while the term *rnd* means the using of randomly initialized embeddings with a subsequent training.

The direct usage of embedding vector is depicted in the leftmost columns of the Proposed Approaches part of Table 2. The results of this method are the worsts one among the other proposed approaches, however it is the simplest one.

The last six columns *emb translation* in Table 2 show the results of CNN on the *(emb)edding translation* method. In Table 2 there are two sets of results. The first one is the set of results, when embedding layer was excluded from learning (*stat*), while in the second case the embeddings layer are further fine-tuned by a training (*rnd*). In the table we can observe, that the embedding training has a positive impact for classification. Moreover, the the impact of the transformation differ from the previous case (see Table 1). We can suggest to use OT as the best transformation method when CNN with trained embeddings are used.

4.4 Comparison with the State of the Art

In this experiment, we compare the results of our best approach with the state of the art (see Table 3). These results show that the state-of-the-art methods slightly outperform the proposed approaches, however we must emphasize that our main goal consists in the comparison of several different methods. Moreover, the proposed approaches are very simple.

Method	**en → de [%]**	**de → en [%]**
Klementiev et al. [7]	77.6	71.1
Kočiský et al. [8]	83.1	76.0
Chandar et al. [4]	91.8	74.2
Coulmance et al. [5]	91.1	78.7
Best proposed configuration	82.7	75.6

Table 3. Comparison with the state of the art.

5 Conclusions

This paper presented a thorough study of the impact of three promising transform methods, namely least squares transformation, orthogonal transformation and canonical correlation analysis, for cross-lingual document classification. In this context, we proposed and evaluated two cross-lingual document classification approaches. The first one uses directly the transformed embeddings in different languages without any modification while the second one realizes the simple word translation choosing the closest word using cosine similarity of the embeddings. We compared the performance of standard maximum entropy classifier with our architecture of convolutional neural network for this task.

We evaluated the proposed approaches on four languages including English, German, Spanish and Italian from Reuters corpus. We have shown that the results of all transformation methods are close to each other, however the orthogonal transformation gives generally slightly better results when CNN with trained embeddings is used. We have also demonstrated that convolutional neural network achieves significantly better results than maximum entropy classifier. We have further presented that the proposed methods are competitive with the state of the art.

Acknowledgements

This work has been partly supported by the project LO1506 of the Czech Ministry of Education, Youth and Sports and by Grant No. SGS-2016-018 Data and Software Engineering for Advanced Applications.

References

1. Berger, A.L., Pietra, V.J.D., Pietra, S.A.D.: A maximum entropy approach to natural language processing. Computational linguistics 22(1), 39–71 (1996)
2. Brychcin, T.: Linear transformations for cross-lingual semantic textual similarity. CoRR abs/1807.04172 (2018), http://arxiv.org/abs/1807.04172
3. Brychcin, T., Taylor, S.E., Svoboda, L.: Cross-lingual word analogies using linear transformations between semantic spaces. CoRR abs/1807.04175 (2018), http://arxiv.org/abs/1807.04175
4. 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)
5. Coulmance, J., Marty, J.M., Wenzek, G., Benhalloum, A.: Trans-gram, fast cross-lingual word-embeddings. arXiv preprint arXiv:1601.02502 (2016)
6. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014)
7. Klementiev, A., Titov, I., Bhattarai, B.: Inducing crosslingual distributed representations of words. Proceedings of COLING 2012 pp. 1459–1474 (2012)
8. 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)
9. Lenc, L., Král, P.: Deep neural networks for Czech multi-label document classification. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 9624 LNCS, 460–471 (2018)
10. Levy, O., Goldberg, Y.: Dependency-based word embeddings. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). vol. 2, pp. 302–308 (2014)
11. 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)
12. 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)
13. 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)