Cross-lingual Transfer for News Article Labeling: Benchmarking Statistical and Neural Models*

Khalil Mrini
École Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne
Switzerland
khalil.mrini@epfl.ch

Nikolaos Pappas
Idiap Research Institute
Rue Marconi 19
CH-1920 Martigny
Switzerland
nikolaos.pappas@idiap.ch

Andrei Popescu-Belis
School of Management and Engineering Vaud (HEIG-VD)
CH-1401 Yverdon-les-Bains
Switzerland
andrei.popescu-belis@heig-vd.ch

Abstract

Cross-lingual transfer has been shown to increase the performance of a text classification model thanks to the use of Multilingual Hierarchical Attention Networks (MHAN), on which this work is based. Firstly, we compared the performance of monolingual and mulitilingual HANs with three types of bag-of-words models. We found that the Binary Unigram model outperforms the HAN model with DENSE encoders on the full vocabulary in 6 out of 8 languages, and ties against MHAN with the DENSE encoders, when it uses the full vocabulary i.e. many more parameters than neural models. However, this is not true when we limit the number of parameters and (or) we increase the sophistication of the neural encoders to GRU or biGRU. Secondly, new configurations of parameter sharing were tested. We found that sharing attention at the sentence level was the best configuration by a small margin when transferring from 5 out of 7 languages to English, as well as for cross-lingual transfer between English and Spanish, Russian, and Arabic. The tests were performed on the Deutsche Welle news corpus with 8 languages and 600k documents.

1 Introduction

Neural networks have been gaining popularity for natural language tasks. Research studies have shown that the performance of neural models can be improved with attention mechanisms (Shen et al., 2015; Luong et al., 2015; Pappas and Popescu-Belis, 2017a) and cross-lingual transfer (Firat et al., 2016; Zou et al., 2013; Zhang et al., 2014). Neural models have been applied to text classification using various kinds of word representations, such as word embeddings (Huang et al., 2012) or bags-of-words (Clark et al., 2003; Johnson and Zhang, 2014). The latter remains one of the most common kinds of word representations (Le and Mikolov, 2014). Previous work shows that combining bigrams with a bag-of-words representation can improve performance (Wang and Manning, 2012), although it usually comes with a high dimensionality that needs to be reduced (Martins et al., 2003).

This study is based on the general neural architecture proposed by Pappas and Popescu-Belis (2017b) for multilingual text classification. This novel text classification model relies on Multilingual Hierarchical Attention Networks (MHANs). The study has shown that cross-lingual transfer provides higher performance than a monolingual HAN, thus corroborating a set of findings that demonstrate the benefits of language transfer for text classification (Jarvis and Crossley, 2012; Bel et al., 2003; Ni et al., 2011; Rigutini et al., 2005).

Moreover, Pappas and Popescu-Belis (2017b) evaluated their proposal on a new multilingual corpus they obtained from Deutsche Welle (DW), Germany's public international broadcaster (http://www.dw.com). The corpus contains nearly 600,000 news articles in 8 languages: English (*en*), German (*de*), Spanish (*es*), Portuguese (*pt*), Ukrainian (*uk*), Russian (*ru*), Arabic (*ar*) and Persian (*fa*). The news articles are tagged and their labels come in two types: *general* and *specific*. For this project, we have only considered the *general* labels, as they form a smaller set (e.g., 327 out of 1385 labels for English) that is better aligned across languages.

In this study, we will first examine in Section 2 how competitive is a bag-of-words model with re-

* Work done while the first and last authors were at the Idiap Research Institute.

spect to the monolingual and multilingual neural models investigated by Pappas and Popescu-Belis (2017b). We will do so first by comparing on a frequency-filtered vocabulary the binary unigram model with a binary model combining unigrams and bigrams and a TF-IDF-weighted bag-of-words model (Section 2.1), then by estimating the influence of features such as stop words (Section 2.2), and finally comparing the performance of a bag-of-words model on the full vocabulary with the one of the neural model (Section 2.3). Then, in Section 3, we will experiment with several hierarchical sharing patterns in the neural model that are suggested as future work by Pappas and Popescu-Belis (2017b). For our experiments, we use the code and data provided by the same authors: `https://github.com/idiap/mhan`.

2 Bag-of-Words Models

In this section, we attempt to quantify the performance of Bag-of-Words representations with a Logistic Regression model in classifying texts in a monolingual manner. Instead of using word embeddings, this model uses a vector representation for each sentence. The dimension of each sentence vector is equal to the number of terms in the whole corpus of the respective language. A document vector is then simply a vector of the sentence vectors. This method does not apply any maximum or minimum on the number of words per sentence or the number of sentences per document and hence does not use zero-padding. However, its computational time depends heavily on the size of the vocabulary used. The text classification is done using logistic regression, with a decision threshold kept at 0.4. In the following experiments, the training is monolingual, and the labels used are the *general* ones from the DW Corpus (see Pappas and Popescu-Belis, 2017b, Table 2).

2.1 Comparing Bag-of-Words Models

In this subsection, we will compare three bag-of-words models: the binary unigram one, the TF-IDF-weighted one, and the binary one using both bigrams and unigrams.

First, for computational purposes, the dimensionality has to be reduced by reducing the vocabulary used. One of the most frequent approaches to do that is term-weighting (Salton and Buckley, 1988). Term frequency, a term-weighting approach, has been often shown to increase performance in text classification (e.g., Xu and Chen, 2010). We computed the term frequency for each term present in the DW corpus. As the number of terms differs considerably from one language to another, rather than keeping a constant number of most frequent terms, we decided to keep the 10% most frequent terms in a given vocabulary, after having removed the stop words using the list provided by Python's NLTK package (Loper and Bird, 2002).

The binary unigram model is such that for sentence s with vector v_s and containing words $w_1, w_2, \ldots w_n$, vector v_s will have a value of 1 in the cells corresponding to the words the sentence contains and 0 in all the other cells.

We also experiment with TF-IDF weights: in a similar way, for sentence s with vector v_s and containing words $w_1, w_2, \ldots w_n$ with TF-IDF weights $t_1, t_2, \ldots t_n$, the vector v_s will have the corresponding TF-IDF weight in cells of the words that the sentence contains and 0 in all the other cells. The TF-IDF weights[1] have been computed for each document (news article), with the IDF part taking into account all of the documents in the training set, using Python's Gensim package (Rehurek and Sojka, 2010).

Finally, the binary model using both bigrams and unigrams is defined exactly like the binary unigram model. However, prior to forming the vectors, words that appear frequently together are joined together to form a single entity. Incorporating bigrams in a binary bag-of-words model has previously been shown to increase performance in text classification (Bekkerman and Allan, 2004; Tan et al., 2002). We formed bigrams with Gensim after frequency thresholding. We used the default settings, with a minimum of 5 occurrences before a word can be considered to form a bigram and a threshold of 10. In other words, a word *a* and a word *b* in a vocabulary of size *N* will form a bigram if and only if $(|a, b| - 5) * N / (|a| * |b|) > 10$.

We ran 3 batches of 8 experiments for monolingual training for text classification, thereby having one experiment per batch for each of the 8 languages studied, and one batch per bag-of-word model. The results, displayed in Table 1, show that the TF-IDF weights model comes out with the lowest performances in 5 out of 8 cases, mak-

[1] Term Frequency multiplied by Inverse Document Frequency (Salton and McGill, 1986).

Bag-of-Words Model	en	de	es	pt	uk	ru	ar	fa
Binary Unigram	**74.7**	70.1	80.6	71.1	**89.5**	76.5	**80.8**	75.5
TF-IDF Weights	74.6	70.5	80.5	71.1	89.2	76.5	80.7	75.8
Binary Bigrams + Unigrams	74.6	**70.8**	**80.8**	**71.8**	89.4	**77.1**	80.3	**76.2**

Table 1: F1 scores of the document classification resulting from our monolingual training of three bag-of-words models using top 10% of the most frequent words.

Bag-of-Words Model	en	de	es	pt	uk	ru	ar	fa
Binary Bigrams + Unigrams	31,497	59,430	33,236	10,881	22,465	34,362	15,804	9,938
The other two models	17,293	37,039	22,006	7,133	12,681	13,502	6,360	7,121

Table 2: Vocabulary size of the binary model using both bigrams and unigrams in comparison with the TF-IDF-weighted model and the Binary Unigram model using top 10% of the most frequent words.

ing it the most frequent lowest performer, followed by the binary unigram model in 4 out of 8 cases. Moreover, the binary model using both bigrams and unigrams outperforms the other ones in 5 out of 8 cases, making it the top performer.

The most likely explanation for these differences is based on the number of parameters. Whereas the TF-IDF-weighted model and the Binary Unigram model have the same vocabulary size, the binary model using both bigrams and unigrams has a higher number of parameters. This difference is shown in Table 2.

Therefore a larger vocabulary size entails higher text classification performance in nearly all of the languages studied. However, given the large increase in vocabulary size that the bigram formation ensues, an experiment on full vocabulary is very expensive computationally with that model and this is why we will try the full vocabulary on the binary unigram model. Before doing so, we will see what kind of influence the stop words have on text classification to evaluate whether to include or exclude them.

2.2 The Influence of Stop Words

Scott and Matwin (1999) have defined stop words as "*functional or connective words that are assumed to have no information content*". Under the hypothesis that they present no information gain, we ran the following experiment to see what is their effect on text classification in our case.

We ran one additional batch of 8 experiments for monolingual training for text classification. This batch did not go through the stop words removal step. The Bag-of-Words model in these experiments is a binary unigram one. The results are displayed in Table 3, along with the ones from the binary unigram model from the previous subsection for comparison.

We see that including the stop words outperforms excluding them by a slight margin, coming out first in 4 pairs of experiments out of 8, and in a tie for 1 pair of experiments. Therefore, in our next experiment, we will include the stop words.

2.3 Binary Unigram Model on Full Vocabulary

In this subsection, we evaluate the binary unigram model on the full vocabulary including stop words for the 8 languages studied. The F1 scores are shown in Table 4, along with the results from Pappas and Popescu-Belis (2017b) for their Monolingual Hierarchical Attention Networks (Mono-HAN) experiment. The Binary Unigram model trained on the full vocabulary outperforms the Mono-HAN model in 6 out of 8 languages by utilizing many more parameters than the latter.

In particular, the number of parameters per language is shown in Table 5. The number of parameters of the binary unigram model for a given language is computed by multiplying the number of words in the corpus in that language by the latter's vocabulary size (which is the dimension of the sentence vector). For the Mono-HAN model, it is computed for a given language by the sum of the word encoder parameters i.e. number of word dimensions (40) times hidden dimension (100), word attention parameters, i.e. hidden dimension times hidden dimension plus hidden dimension, sentence encoder and attention parameters, and classification layer parameters, i.e. hidden dimensions times the number of words in the corpus. The ratio in Table 5 shows that the binary unigram model uses on average 779 times more parameters than the Mono-HAN one, yet the latter remains competitive in two of the languages.

Stop Words	en	de	es	pt	uk	ru	ar	fa
Excluded	74.7	70.1	**80.6**	71.1	**89.5**	76.5	80.8	**75.5**
Included	**74.8**	**70.5**	80.3	**71.2**	89.4	**76.6**	80.8	75.3

Table 3: F1 scores of the document classification resulting from our monolingual binary unigram bag-of-words experiment with and without stop words using top 10% of the most frequent words.

Model	en	de	es	pt	uk	ru	ar	fa	#params per lang.
Binary Unigram	**75.8**	**72.9**	81.4	**74.3**	**91.0**	79.2	**82.0**	**77.0**	26,850,575
Mono-HAN (DENSE)	71.2	71.8	**82.8**	71.3	85.3	**79.8**	80.5	76.6	50,257

Table 4: F1 scores of the document classification resulting from our monolingual binary unigram bag-of-words experiment in comparison with the monolingual hierarchical attention networks experiment with DENSE encoders, Mono-HAN (DENSE), from Pappas and Popescu-Belis (2017b).

Given that the Mono-HAN model was always outperformed by one of the configurations of the Multi-HAN model, we can compare the best performance of the Multi-HAN model with the performance of the binary unigram model. The comparison in Table 6 shows that there is a tie as both models come up first in 4 out of 8 languages. When computing the average of the performance, we see that the bag-of-words model with its 79.2 average F1 score outperforms by a slight margin the Multi-HAN model with its 78.6 average F1 score. This can again be explained by the fact Binary Unigram uses many more parameters.

However, Pappas and Popescu-Belis (2017b) have also explored Multi- and Mono-HAN models with bi-directional Gated Recurrent Units (biGRU) with 100 hidden dimensions and 40-dimensional word embeddings, for Arabic and English. The results for testing on English give **77.7** for Mono-HAN, which outperforms by more than 1 point our Binary Unigram model. Likewise, testing on Arabic with a Multi-HAN model sharing attention mechanisms gives **84.0**, which outperforms by 2 points our Binary Unigram model.[2] We will therefore attempt to see if other kinds of configurations can give better results than those obtained by them.

3 Neural Network Model: MHAN

Pappas and Popescu-Belis (2017b) designed the hierarchical attention networks for document representation building upon a proposal by Yang et al. (2016). Pappas and Popescu-Belis (2017b) considered two levels of aggregation to construct a document representation: words to sentence (word level) and then sentences to document (sentence level). The network they conceived uses encoders and attention mechanisms, that are conceived separately for each level. The functions used by the encoders have parameters H_w for the word level and H_s for sentence level. Likewise, attention mechanisms are defined as α_w for the word level and α_s for the sentence level. The word embeddings that were used were 40-dimensional pre-trained ones provided by Ammar et al. (2016). They are multilingual and aligned embeddings, therefore enabling transfer of knowledge from one language to another.

This transfer can be used to share encoders and attention mechanisms across languages. Pappas and Popescu-Belis (2017b) have considered three options: sharing encoders, sharing attention mechanisms and sharing both of them. These options can be visualised in Figure 1. They ran these experiments in a multilingual setting: pairs of distinct languages with one of them being English. They have concluded that sharing attention mechanisms gives the best results for text categorisation for English, but the experiments give mixed results for the seven other languages, as can be seen in Table 7. To explore more options and evaluate the best performing one, as suggested by Pappas and Popescu-Belis (2017b), we will run two experiments with encoders using a fully-connected network called DENSE that they have used. First, we will share the attention mechanisms at both levels and the encoders at sentence level only. Then, we will share attention mechanisms at sentence level only and not share the encoders. These two additional experiments can be visualised in Figure 2.

[2]Note that these GRU and BiGRU models perform well despite their relatively low capacity in terms of parameters, namely higher than DENSE encoders, but much lower than Binary Unigram. In fact, the performance of neural models can be further improved by increasing their parameters, for example, by using 200 hidden dimensions for the encoders and 300-dimensions for the word embeddings.

Model	en	de	es	pt	uk	ru	ar	fa
Binary Uni.	34,923,200	109,615,000	2,3027,600	5,688,000	7,906,400	23,931,900	4,881,100	4,831,400
Mono-HAN	67,629	71,669	50,661	44,197	37,430	44,904	43,793	41,773
Ratio	859	2294	843	400	620	615	235	369

Table 5: Number of parameters of the monolingual binary unigram model with full vocabulary in comparison with the monolingual hierarchical attention networks (Mono-HAN) model with DENSE encoders from Pappas and Popescu-Belis (2017b).

Model	en	de	es	pt	uk	ru	ar	fa	#params per lang.
Binary Unigram	**75.8**	**72.9**	81.4	**74.3**	**91.0**	79.2	82.0	77.0	26,850,575
Multi-HAN (DENSE)	74.2	72.5	**82.9**	71.6	87.7	**80.8**	**82.1**	**77.1**	40,128

Table 6: F1 scores of the document classification resulting from our monolingual binary unigram bag-of-words experiment with full vocabulary in comparison with the performance of the best multilingual hierarchical attention network (Multi-HAN) with DENSE encoders from Pappas and Popescu-Belis (2017b).

Since we are trying to compare our results with the ones obtained in Pappas and Popescu-Belis (2017b), we will run the experiments under the same conditions. That means that we will use a maximum of 30 words per sentence, and a maximum of 30 sentences per document. If there is a surplus, we will cut it short, and if there is less than the maximum, we will use zero-padding. We also use all the documents of the DW Corpus. Settings, such as the 100 dimensions for encoders and attention embeddings, the batch size of 16 and the epoch size of 25,000, remain the same. These experiments use the *general* labels. The decision threshold will be set at 0.4.

	w	s
α		
H	✓	✓

(a) Sharing Encoders

	w	s
α	✓	✓
H		

(b) Sharing Attention Mechanisms

	w	s
α	✓	✓
H	✓	✓

(c) Sharing Both

Figure 1: Visualisation of the configurations evaluated in multilingual experiments in Pappas and Popescu-Belis (2017b).

	w	s
α	✓	✓
H		✓

(a) Sharing Encoders at Sentence Level Only and Sharing Attention Mechanisms

	w	s
α		✓
H		

(b) Sharing Attention Mechanisms at Sentence Level Only

Figure 2: Visualisation of the configurations evaluated in multilingual experiments in this paper.

First, in Table 7, we can observe that the best model bilingually on average is the one with shared attention across languages. However, when looking at the performance on languages other than English: the model with shared attention mechanisms at both levels comes first only in 3 out of 7 experiments, and the one with shared encoders and attention mechanisms comes first in the same number of experiments. Therefore, we decided to try a hybrid configuration which is a combination between these two: sharing attention mechanisms at both levels and sharing encoders at sentence level only. The results obtained are displayed in Table 8.

To compare the results from this configuration with the ones obtained in Pappas and Popescu-Belis (2017b), we computed the differences to be able to see how our configuration performs with regards to the other ones. The resulting comparisons for testing on English are in Figure 3 and the ones for testing on the other languages are in Figure 4. Figure 4 shows that our configuration comes out as the most frequent lowest performer with lowest performances in 3 out of 7 experiments. However, Figure 3 shows that, for testing on English, although our configuration is always outperformed by the configuration with shared attention mechanisms at both levels, it outperforms both the configurations with shared encoders and shared encoders and attention mechanisms in 5 out of 7 experiments. This mixed performance can be explained by the fact that this configuration has fewer parameters than the configuration with shared attention mechanisms at both levels. Although the configuration with shared encoders and attention mechanisms is the only one with fewer

	Testing on English							Testing on the other language						
English and ...	de	es	pt	uk	ru	ar	fa	de	es	pt	uk	ru	ar	fa
Sharing H	71.0	69.9	69.2	70.8	71.5	70.0	71.3	69.7	**82.9**	69.7	86.8	80.3	79.0	76.0
Sharing α	**74.0**	**74.2**	**74.1**	**72.9**	**73.9**	**73.8**	**73.3**	**72.5**	82.5	70.8	**87.7**	80.5	**82.1**	76.3
Sharing Both	72.8	71.2	70.5	65.6	71.1	68.9	69.2	70.4	82.8	**71.6**	87.5	**80.8**	79.1	**77.1**

Table 7: F1 scores of the document classification resulting from the multilingual hierarchial attention networks configurations with DENSE encoders in Pappas and Popescu-Belis (2017b).

	Testing on English							Testing on the other language						
English and ...	de	es	pt	uk	ru	ar	fa	de	es	pt	uk	ru	ar	fa
Results	73.8	72.9	70.0	70.9	69.5	73.2	72.3	69.8	82.9	69.2	86.7	79.3	81.6	76.6

Table 8: F1 scores of the document classification resulting from our multilingual hierarchial attention networks configuration with DENSE encoders shared at sentence level only and attention mechanisms shared at both word and sentence levels.

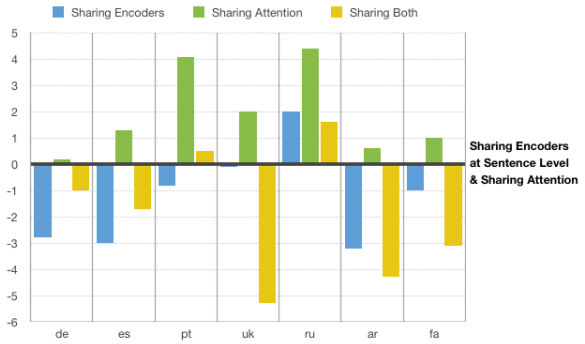

Figure 3: Differences between the results when testing for English between multilingual hierarchical attention networks configurations in Pappas and Popescu-Belis (2017b) and our configuration with shared attention mechanisms at both word and sentence levels and shared encoders at sentence level only.

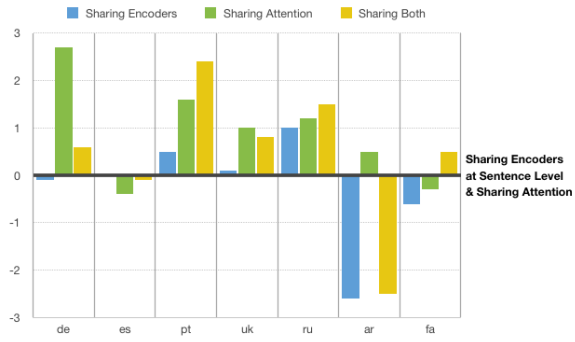

Figure 4: Differences between the results when testing for the other languages between multilingual hierarchical attention networks configurations in Pappas and Popescu-Belis (2017b) and our configuration with shared attention mechanisms at both word and sentence levels and shared encoders at sentence level only.

parameters than this configuration, the latter was outperformed by the former most of the time when testing on the other languages, but the trend was reversed when testing for English.

Second, seeing that our first configuration presented mixed results, we decide to test another configuration closer to the one with shared attention mechanisms at both levels (the most successful one): a configuration where encoders are not shared and attention mechanisms are shared at sentence level only. The results obtained are displayed in Table 9. We again computed the differences and the comparisons of the results for testing for English are in Figure 5 and the ones for testing for the other languages are in Figure 6.

The results when testing for the other languages remain very mixed and no clear conclusion can be drawn, as conclusions are language-specific. Figure 6 shows that sharing attention mechanisms at sentence level only outperforms sharing attention mechanisms at both levels for 3 out of 7 experiments by a small margin (about 0.46 F1 on average over 3 languages). Likewise, when sharing attention at both levels outperforms our configuration, it is by a larger margin (about 0.80 F1 on average over 4 languages). We can however see that sharing encoders is the most frequent experiment to have the lowest performance out of the 4, performing the worst in 6 out of 7 experiments. The most frequent top performer remains sharing attention mechanisms at both levels, coming first in 3 out of 7 experiments. Nonetheless, when testing on English, sharing attention mechanisms at sentence level only outperforms sharing attention mechanisms at both levels on 5 out of 7 languages by a small margin (about 0.4 F1 on average), as it

	Testing on English							Testing on the other language						
English and . . .	de	es	pt	uk	ru	ar	fa	de	es	pt	uk	ru	ar	fa
Results	73.7	73.9	74.2	73.5	74.3	74.1	74.1	71.5	82.8	70.2	87.3	81.2	80.9	76.7

Table 9: F1 scores of the document classification resulting from our multilingual hierarchial attention networks configuration with DENSE encoders not shared and attention mechanisms shared at sentence level only.

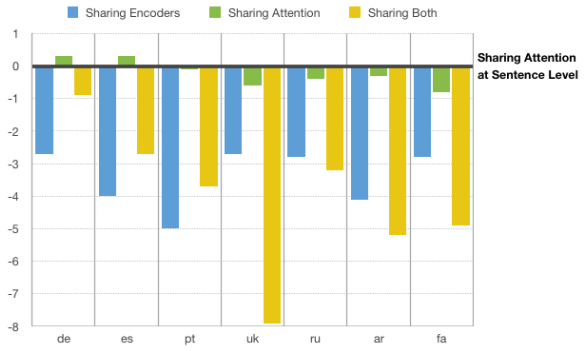

Figure 5: Differences between the results when testing for English between multilingual hierarchical attention networks configurations in Pappas and Popescu-Belis (2017b) and our configuration with shared attention mechanisms at sentence level.

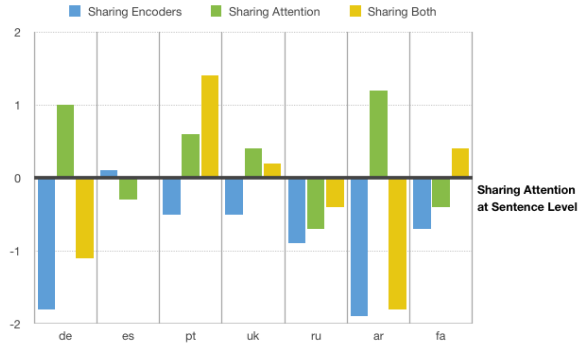

Figure 6: Differences between the results when testing for the other languages between multilingual hierarchical attention networks configurations in Pappas and Popescu-Belis (2017b) and our configuration with shared attention mechanisms at sentence level.

can be seen in Figure 5.

Apart from the 5 languages where this configuration slightly outperformed all the other ones, the performance was top when testing for Russian too. Moreover, for Persian, it outperformed sharing attention at both levels, but not sharing encoders and attention. For Portuguese, this configuration was outperformed by the ones where attention is shared at both levels, and where both encoders and attention are shared by a clear margin. The same can be observed for Ukrainian at a smaller margin. However, testing for Arabic presents mixed results, as this configuration outperforms by a clear margin the configurations with encoders shared and with both attention and encoders shared, but presents a decrease when compared with the configuration with shared attention at both levels.

4 Conclusions

In this study, we benchmarked several text classification models taking as baseline the monolingual and multilingual models investigated by Pappas and Popescu-Belis (2017b) in multilingual text classification on Deutsche-Welle corpus.

First, we tested three Bag-of-Words representations: Binary Unigram, TF-IDF weights, and Binary Bigrams mixed with Unigrams. We combined them with a Logistic Regression model, stripped the corpus of stop words, and filtered the vocabulary to keep the 10% most popular words based on term frequency for computational purposes. We found out that the most frequent top performer was the representation using Bigrams, due to the fact that it uses more parameters than its competitors. Then, we evaluated the importance of stop words and found out that is better to include them. Our results showed that the Binary Unigram model outperforms the HAN model with DENSE encoders on the full vocabulary in 6 out of 8 languages, and ties against MHAN with the DENSE encoders, when it uses the full vocabulary i.e. many more parameters than neural models. However, this is not true when we limit the number of parameters to be used and (or) we increase the sophistication of the neural encoders to GRU or biGRU.

Lastly, we evaluated new configurations for parameter sharing for the Multi-HAN models. The first configuration, sharing attention at both levels and encoders at sentence level, did not yield better results than the previously proposed Multi-

HAN configurations. However, the second configuration, sharing attention at sentence level, outperformed by a small margin the best multilingual configuration on 3 out of 7 languages excluding English, and on 5 out of 7 languages when considering performance only on English. For the remaining languages except English, there is no clear winner among the multilingual configurations, since two out of the five configurations (sharing α and sharing both) outperform the rest with similar performance on average.

References

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively multilingual word embeddings. *CoRR* abs/1602.01925.

Ron Bekkerman and James Allan. 2004. Using bigrams in text categorization. Technical report, Technical Report IR-408, Center of Intelligent Information Retrieval, UMass Amherst.

Nuria Bel, Cornelis HA Koster, and Marta Villegas. 2003. Cross-lingual text categorization. In *International Conference on Theory and Practice of Digital Libraries*. Springer, pages 126–139.

James Clark, Irena Koprinska, and Josiah Poon. 2003. A neural network based approach to automated e-mail classification. In *Web Intelligence, 2003. WI 2003. Proceedings. IEEE/WIC International Conference on*. IEEE, pages 702–705.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. *arXiv preprint arXiv:1601.01073* .

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, pages 873–882.

Scott Jarvis and Scott A Crossley. 2012. *Approaching Language Transfer Through Text Classification: Explorations in the Detectionbased Approach*, volume 64. Multilingual Matters.

Rie Johnson and Tong Zhang. 2014. Effective use of word order for text categorization with convolutional neural networks. *arXiv preprint arXiv:1412.1058* .

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*. pages 1188–1196.

Edward Loper and Steven Bird. 2002. NLTK: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*. Association for Computational Linguistics, Stroudsburg, PA, USA, ETMTNLP '02, pages 63–70.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* .

Claudia Aparecida Martins, Maria Carolina Monard, and Edson Takashi Matsubara. 2003. Reducing the dimensionality of bag-of-words text representation used by learning algorithms. In *Proc of 3rd IASTED International Conference on Artificial Intelligence and Applications*. pages 228–233.

Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. 2011. Cross lingual text classification by mining multilingual topics from wikipedia. In *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, pages 375–384.

Nikolaos Pappas and Andrei Popescu-Belis. 2017a. Explicit document modeling through weighted multiple-instance learning. *Journal of Artificial Intelligence Research* pages 591–626.

Nikolaos Pappas and Andrei Popescu-Belis. 2017b. Multilingual hierarchical attention networks for document classification. In *Proceedings of the 8th International Joint Conference on Natural Language Processing (IJCNLP)*. Association for Computational Linguistics, Taipei, Taiwan.

Radim Rehurek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Citeseer.

Leonardo Rigutini, Marco Maggini, and Bing Liu. 2005. An em based training algorithm for cross-language text categorization. In *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*. IEEE Computer Society, pages 529–535.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management* 24(5):513–523.

Gerard Salton and Michael J. McGill. 1986. *Introduction to modern information retrieval*. McGraw-Hill, Inc.

Sam Scott and Stan Matwin. 1999. Feature engineering for text classification. In *ICML*. volume 99, pages 379–388.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2015. Minimum risk training for neural machine translation. *arXiv preprint arXiv:1512.02433* .

Chade-Meng Tan, Yuan-Fang Wang, and Chan-Do Lee. 2002. The use of bigrams to enhance text categorization. *Information processing & management* 38(4):529–546.

Sida Wang and Christopher D Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*. Association for Computational Linguistics, pages 90–94.

Yan Xu and Lin Chen. 2010. Term-frequency based feature selection methods for text categorization. In *Genetic and Evolutionary Computing (ICGEC), 2010 Fourth International Conference on*. IEEE, pages 280–283.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pages 1480–1489.

Jiajun Zhang, Shujie Liu, Mu Li, Ming Zhou, Chengqing Zong, et al. 2014. Bilingually-constrained phrase embeddings for machine translation. In *ACL (1)*. pages 111–121.

Will Y Zou, Richard Socher, Daniel M Cer, and Christopher D Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *EMNLP*. pages 1393–1398.