IDIAP RESEARCH REPORT

MULTILINGUAL HIERARCHICAL ATTENTION NETWORKS FOR DOCUMENT CLASSIFICATION

Nikolaos Pappas Andrei Popescu-Belis

Idiap-RR-17-2017

Version of SEPTEMBER 08, 2017

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Multilingual Hierarchical Attention Networks for Document Classification

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

Andrei Popescu-Belis
Idiap Research Institute
Rue Marconi 19
CH-1920 Martigny, Switzerland
andrei.popescu-belis@idiap.ch

Abstract

Hierarchical attention networks have recently achieved remarkable performance for document classification in a given language. However, when multilingual document collections are considered, training such models separately for each language entails linear parameter growth and lack of cross-language transfer. Learning a single multilingual model with fewer parameters is therefore a challenging but potentially beneficial objective. To this end, we propose multilingual hierarchical attention networks for learning document structures, with shared encoders and/or attention mechanisms across languages, using multi-task learning and an aligned semantic space as input. We evaluate the proposed models on multilingual document classification with disjoint label sets, on a large dataset which we provide, with 600k news documents in 8 languages, and 5k labels. The multilingual models outperform strong monolingual ones in low-resource as well as full-resource settings, and use fewer parameters, thus confirming their computational efficiency and the utility of cross-language transfer.

1 Introduction

Learning word sequence representations has become increasingly useful for a variety of NLP tasks such as document classification (Tang et al., 2015; Yang et al., 2016), neural machine translation (NMT) (Cho et al., 2014; Luong et al., 2015), question answering (Chen et al., 2015; Kumar et al., 2015) and summarization (Rush et al., 2015). However, when data are available in multiple languages, representation learning must address two main challenges. Firstly, the computational cost of training separate models for each language which grows linearly with their number, or even quadratically in the case of multi-way multilingual NMT (Firat et al., 2016a). Secondly, the models should be capable of cross-language transfer, which is an important component in human language learning (Ringbom, 2007). For instance, (Johnson et al., 2016) attempted to use a single sequence-to-sequence neural network model for NMT across multiple language pairs.

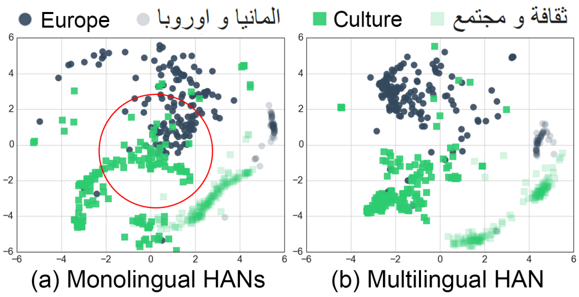


Figure 1: Vectors of documents labeled with *‘Europe’*, *‘Culture’* and their Arabic counterparts. The multilingual hierarchical attention network separates topics better than monolingual ones.

Previous studies in document classification attempted to overcome these issues by employing multilingual word embeddings which allow direct comparisons and groupings across languages (Klementiev et al., 2012; Hermann and Blunsom, 2014; Ferreira et al., 2016). However, they are only applicable when common label sets are available across languages which is often not the case (e.g. Wikipedia or news). Moreover, despite recent advances in monolingual document modeling (Tang et al., 2015; Yang et al., 2016), multilingual models are still based on shallow approaches.

In this paper, we propose *Multilingual Hierarchical Attention Networks* to learn shared document structures across languages for document classification with disjoint label sets, as opposed to training hierarchical attention networks (HANs) in a language-specific manner (Yang et al., 2016). Our networks have a hierarchical structure with word and sentence encoders, along with attention mechanisms. Each of these can either be shared across languages or kept language-specific. To enable cross-language transfer, the networks are trained with multi-task learning across languages using an aligned semantic space as input. Fig. 1 displays document vectors, projected with t-SNE (van der Maaten, 2009), for two topics and two languages, either learned by monolingual HANs (a) or by our multilingual HAN (b). The multilingual HAN achieves better separation between 'Europe' and 'Culture' topics in English as a result of the knowledge transfer from Arabic.

We evaluate our model against strong monolingual baselines, in low-resource and full-resource scenarios, on a large multilingual document collection with 600k documents, labeled with general (1.2k) and specific topics (4.4k), from Deutsche Welle – which we will make available. Our multilingual models outperform monolingual ones in both scenarios, thus confirming the utility of cross-language transfer and the computational efficiency of the proposed architecture. To encourage further research in multilingual representation learning our code and dataset are made available at `https://github.com/idiap/mhan`.

2 Related Work

Research on *learning multilingual word representations* is based on early work on word embeddings (Turian et al., 2010; Mikolov et al., 2013; Pennington et al., 2014). The goal is to learn an aligned word embedding space for multiple languages by leveraging bilingual dictionaries (Faruqui and Dyer, 2014; Ammar et al., 2016), parallel sentences (Gouws et al., 2015) or comparable documents such as Wikipedia pages (Yih et al., 2011; Al-Rfou et al., 2013). Bilingual embeddings were learned from word alignments using neural language models (Klementiev et al., 2012; Zou et al., 2013), including auto-encoders (Chandar et al., 2014). Despite progress at the word level, the document level remains comparatively less explored: the approaches proposed by Hermann and Blunsom (2014) or Ferreira et al. (2016) are based on shallow modeling and are applicable only to classification tasks with label sets shared across languages. These sets are costly to produce, and are often unavailable; here, we remove this constraint. We also develop deeper multilingual document models with hierarchical structure based on prior art at the word level.

Early work on *neural document classification* was based on shallow feed-forward networks, which required unsupervised pre-training (Le and Mikolov, 2014). Later studies focused on neural networks with hierarchical structure. Kim (2014) proposed a convolutional neural network (CNN) for sentence classification. Johnson and Zhang (2015) proposed a CNN for high-dimensional data classification, while Zhang et al. (2015) adopted a character-level CNN for text classification. Lai et al. (2015) proposed a recurrent CNN to capture sequential information, which outperformed simpler CNNs. Lin et al. (2015) and Tang et al. (2015) proposed hierarchical recurrent NNs and showed that they were superior to CNN-based models. Recently, Yang et al. (2016) demonstrated that a hierarchical attention network with bi-directional gated encoders outperforms traditional and neural baselines. Using such networks in multilingual settings has two drawbacks: the computational complexity increases linearly with the number of languages, and knowledge is acquired separately for each language. We address these issues by proposing a new multilingual model based on HANs, which learns shared document structures and to transfer knowledge across languages.

Early examples of *attention mechanisms* appeared in computer vision tasks such as optical character recognition (Larochelle and Hinton, 2010) and image tracking (Denil et al., 2012) or classification (Mnih et al., 2014). In text classification, studies which aimed to learn the importance of sentences included Yessenalina et al. (2010); Pappas and Popescu-Belis (2014); Yang et al. (2016) and more recently (Pappas and Popescu-Belis, 2017; Ji and Smith, 2017). In NMT, Bahdanau et al. (2015) proposed an attention-based encoder-decoder network, while Luong et al. (2015) proposed a local and ensemble attention model for NMT. Firat et al. (2016a) proposed a single encoder-decoder model with shared attention across language pairs for multi-way, multilingual NMT. Hermann et al. (2015) developed attention-based document readers for question answering. Chen et al. (2015) proposed a recurrent attention model over an external memory. Similarly, Kumar et al. (2015) introduced a dynamic memory network for question answering and other tasks. We propose hierarchical models with shared multi-level attention across languages, which, to our knowledge, has not been attempted before.

3 Background: Hierarchical Attention Networks for Document Classification

We adopt a general hierarchical attention architecture for document representation derived from a specific one proposed by Yang et al. (2016), displayed in Figure 2. The goal of this generality is to be able to represent a large number of architectures rather than individual ones. We consider a dataset $D = \{(x_i, y_i), i = 1, \dots, N\}$ made of N documents x_i with labels $y_i \in \{0,1\}^k$. Each document $x_i = \{w_{11}, w_{12}, \dots, w_{KT}\}$ is represented by the sequence of d-dimensional embeddings of their words grouped into sentences, T being the maximum number of words in a sentence, and K the maximum number of sentences in a document.

The network takes as input a document x_i and outputs a document vector u_i. In particular, it has two levels of abstraction, word vs. sentence. The former is made of an encoder g_w with parameters H_w and an attention a_w with parameters A_w, while the latter similarly includes an encoder and an attention (g_s, H_s and a_s, A_s). The output u_i is used by the classification layer to determine y_i.

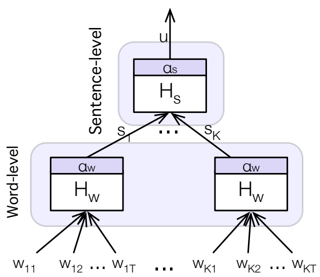


Figure 2: General architecture of hierarchical attention neural networks for modeling documents.

3.1 Encoder Layers

At the word level, the function g_w encodes the sequence of input words $\{w_{it} \mid t = 1, \dots, T\}$ for each sentence i of the document, noted as:

$$h_w^{(it)} = g_w(w_{it}), t \in [1, T], \quad (1)$$

and at the sentence level, after combining the intermediate word vectors $\{h_w^{(it)} \mid t = 1, \dots, K\}$ to a sentence vector s_i (see Section 3.2), the function g_s encodes the sequence of sentence vectors $\{s_i \mid i = 1, \dots, S\}$, noted as $h_s^{(i)}$.

The g_w and g_s functions can be any feed-forward or recurrent networks with parameters H_w and H_s respectively. We consider the following networks: a fully-connected one, noted as *Dense*, a Gated Recurrent Unit network (Cho et al., 2014) noted as *GRU*[1], and a bi-directional GRU which captures temporal information forward or backward in time, noted as *biGRU*. The latter is defined as a concatenation of the hidden states for each input vector obtained from the forward GRU, $\vec{g_w}$, and the backward GRU, $\overleftarrow{g_w}$:

$$h_w^{(it)} = \left[\vec{g_w}(h_w^{(it)}); \overleftarrow{g_w}(h_w^{(it)})\right] \quad (2)$$

The same concatenation is applied for the hidden-state representation of a sentence $h_s^{(i)}$.

3.2 Attention Layers

A typical way to obtain a representation for a given word sequence at each level is by taking the last hidden-state vector that is output by the encoder. However, it is hard to encode all the relevant input information needed in a fixed-length vector. This problem is addressed by introducing an attention mechanism at each level (noted α_w and α_s) that estimates the importance of each hidden state vector to the representation of the sentence or document meaning respectively. The sentence vector $s_i \in R^{d_w}$, where d_w is the dimension of the word encoder, is thus obtained as follows:

$$\frac{1}{T}\sum_{t=1}^{T} \alpha_w^{(it)} h_w^{(it)} = \frac{1}{T}\sum_{t=1}^{T} \frac{\exp(v_{it}^\top u_w)}{\sum_j \exp(v_{ij}^\top u_w)} h_w^{(it)} \quad (3)$$

where $v_{it} = f_w(h_w^{(it)})$ is a fully-connected neural network with W_w parameters. Similarly, the document vector $u \in R^{d_s}$, where d_s is the dimension of the sentence encoder, is obtained as follows:

$$\frac{1}{K}\sum_{i=1}^{K} \alpha_s^{(i)} h_s^{(i)} = \frac{1}{K}\sum_{i=1}^{K} \frac{\exp(v_i^\top u_s)}{\sum_j \exp(v_j^\top u_s)} h_s^{(i)} \quad (4)$$

where $v_i = f_s(h_w^{(i)})$ is a fully-connected neural network with W_s parameters. The vectors u_w and u_s are parameters which encode the word context and sentence context respectively, and are learned jointly with the rest of the parameters. The total set of parameters for a_w is $A_w = \{W_w, u_w\}$ and for a_s is $A_s = \{W_s, u_s\}$.

[1]GRU is a simplified version of Long-Short Term Memory, LSTM (Hochreiter and Schmidhuber, 1997).

3.3 Classification Layers

The output of such a network is typically fed to a softmax layer for classification, with a loss based on the cross-entropy between gold and predicted labels (Tang et al., 2015) or a loss based on the negative log-likelihood of the correct labels (Yang et al., 2016). However, softmax overemphasizes the probability of the most likely label, which may not be ideal for multi-label classification because each document will have very few labels with likely values (e.g. 0.2–0.9). Instead, it is more appropriate to have independent predictions for each label; as we also verified empirically in our preliminary experiments. Hence, we replace the softmax with a sigmoid function, for each document i represented by the vector u_i we model the probability of the k labels as follows:

$$\hat{y}_i = p(y|u_i) = \frac{1}{1 + e^{-(W_c u_i + b_c)}} \in [0,1]^k \quad (5)$$

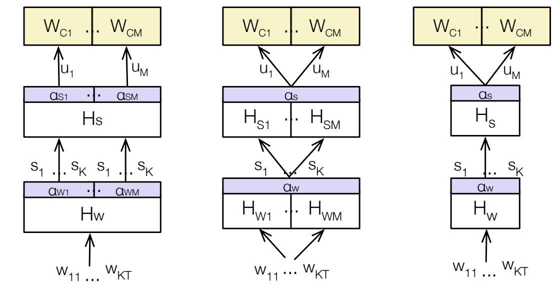
where W_c is a $d_s \times k$ matrix and b_c is the bias term for the classification layer. The training loss based on cross-entropy is computed as follows:

$$\mathcal{L}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\mathcal{H}(y_i, \hat{y}_i) \quad (6)$$

where θ is a notation for all the parameters of the model (i.e. H_w, A_w, H_s, A_s, W_c), and $\mathcal{H}$ is the binary cross-entropy of the gold labels y_i and predicted labels $\hat{y}_i$ for a document i. The above objective is differentiable and can be minimized with stochastic gradient descent (Bottou, 1998) or variants such as Adam (Kingma and Ba, 2014), to maximize classification performance.

4 Multilingual Hierarchical Attention Networks: MHANs

When multilingual data is available, the above network can be trained on each language separately but the needed parameters grow linearly with the number of languages. Moreover, they do not exploit common knowledge across languages or to transfer it from one to another. We propose here a hierarchical attention network with shared components across languages which have slower parameter growth compared to monolingual ones (hence sublinear), which enables knowledge transfer across languages. We now consider M languages noted $L = \{L_l \mid l = 1, \dots, M\}$, and a multilingual set of topic-labeled documents $D = \{(x_i^{(l)}, y_i^{(l)}) \mid i = 1, \dots, N_l\}$ defined as above.



(a) Sharing Encoders (b) Sharing Attentions (c) Sharing Both

Figure 3: Multilingual attention networks for modeling documents over disjoint label spaces.

4.1 Sharing Components across Languages

To enable multilingual learning, we propose three distinct ways for sharing components between networks in a multi-task learning setting, depicted in Figure 3, namely: (a) sharing the parameters of word and sentence encoders, hence $\theta_{enc} = \{H_w, W_w^{(l)}, H_s, W_s^{(l)}, W_c^{(l)}\}$; (b) sharing the parameters of word and sentence attention models, hence $\theta_{att} = \{H_w^{(l)}, W_w, H_s^{(l)}, W_s, W_c^{(l)}\}$; and (c) sharing both previous sets of parameters, hence $\theta_{both} = \{H_w, W_w, H_s, W_s, W_c^{(l)}\}$. For instance, the document representation of a text for language l based on a shared sentence-level attention would be computed based on Eq.4 by using the same parameters W_s and u_s across languages.

Let $\theta_{mono} = \{H_w^{(l)}, W_w^{(l)}, H_s^{(l)}, W_s^{(l)}, W_c^{(l)}\}$ be the parameters of multiple independent monolingual models with Dense encoders, then we have:

$$|\theta_{mono}| > |\theta_{enc}| > |\theta_{att}| > |\theta_{both}| \quad (7)$$

where $|\cdot|$ is the number of parameters in a set. For GRU and biGRU encoders, the inequalities still hold, but swapping $|\theta_{enc}|$ and $|\theta_{att}|$. Excluding the classification layer which is nessesarily language-specific[2], the (a) and (b) networks have sublinear numbers of parameters and the (c) network has a constant number of parameters with respect to the number of languages.

Depending on the label sets, several types of document classification problems can be solved with such architectures. First, label sets can be common or disjoint across languages. Second, considering labels as k-hot vectors, $k = 1$ corresponds to a multi-class task, while $k > 1$ is a multi-label task. We focus here on the multi-label

[2]Note that the word embeddings are not considered as parameters in our setup because they are fixed during training. For learned word embeddings, the analysis still applies if we consider their parameters as part of the word-level encoder.

problem with disjoint label sets. Moreover, we assume an aligned input space i.e with multilingual word embeddings that have aligned meanings across languages (Ammar et al., 2016). With non-aligned word embeddings, the multilingual transfer is harder due to the complete lack of parallel information, as we show in Section 6.

4.2 Training over Disjoint Label Sets

For training, we replace the monolingual training objective (Eq. 6) with a joint multilingual objective that facilitates the sharing of components, i.e. a subset of parameters for each language $\theta_1, \theta_2, ..., \theta_M$, across different language networks:

$$\mathcal{L}(\theta_{1,...,}\theta_M) = -\frac{1}{Z}\sum_{i}^{N_e}\sum_{l}^{M}\mathcal{H}(y_i^{(l)}, \hat{y}_i^{(l)}) \quad (8)$$

where $Z = M \times N_e$ with N_e being the epoch size. Here it may also be beneficial to add a term γ_l for each language objective which encodes our prior knowledge about its importance; we keep this as future work. The joint objective $\mathcal{L}$ can be minimized with respect to the parameters $\theta_1, \theta_2, ..., \theta_M$ using SGD as before. However, when training on examples from different languages consecutively (I iterations) it is difficult to learn a shared space which works well across languages.

This is because each language updates apply only on a subset of parameters and may bias the model away from other languages. To address this, we employ the training strategy from (Firat et al., 2016a), who sampled parallel sentences for multi-way machine translation from different language pairs in a cyclic fashion at each iteration.[3] Here, we sample instead a document-label pair j from each language i at iteration j, $L_i^{(j)}$, as follows:

$$\left(L_1^{(1)} \ldots L_M^{(1)}\right) \rightarrow, \ldots, \rightarrow \left(L_1^{(I)} \ldots L_M^{(I)}\right) \quad (9)$$

For minibatch SGD, the number of samples per language are equal to the batch size divided by M.

5 A New Corpus for Multilingual Document Classification: DW

Multilingual document classification datasets are usually limited in size, have target categories aligned across languages, and assign documents to only one category. However, classification is often necessary in cases where the categories are not strictly aligned and can be multiple per document. For instance, this is the case for online news agencies keeping track of multilingual news written and annotated by different journalists, a process which is expensive and time-consuming.

[3]We verified this empirically in our preliminary experiments and found out that mixing languages in a single batch performed better than keeping them in separate batches.

Languages	Documents			Labels	
L	$\|X\|$	$\bar{s}$	$\bar{w}$	$\|Y_g\|$	$\|Y_s\|$
English	112,816	17.9	516.2	327	1,058
German	132,709	22.3	424.1	367	809
Spanish	75,827	13.8	412.9	159	684
Portuguese	39,474	20.2	571.9	95	301
Ukrainian	35,423	17.6	342.9	28	260
Russian	108,076	16.4	330.1	102	814
Arabic	57,697	13.3	357.7	91	344
Persian	36,282	18.7	538.4	71	127
All	598,304	17.52	436.7	1,240	4,397

Table 1: Statistics of the Deutsche Welle corpus: $\bar{s}$ and $\bar{w}$ are the average numbers of sentences and words per document.

Two datasets for multilingual document classification have been used in previous studies: Reuters RCV1/RCV2 (6,000 documents, 2 languages and 4 labels) and TED talk transcripts (12,078 documents, 12 languages and 15 labels) introduced by Hermann and Blunsom (2014). The former is tailored for evaluating word embeddings aligned across languages, rather than complex multilingual document models. The latter is twice as large and covers more languages, in a multi-label setting, but biases evaluation by including translations of talks in all languages.

Here, we present and use a much larger dataset collected from Deutsche Welle, Germany's public international broadcaster. The DW dataset contains nearly 600,000 documents annotated by journalists with several topic labels in 8 languages, shown in Table 1. Documents are on average 2.6 times longer than in Yang et al.'s (2016) monolingual dataset (436 vs. 163 words). There are two types of labels, namely *general topics* (Y_g) and *specific* ones (Y_s) described by one or more words. We consider (and count in Table 1) only those specific labels that appear at least 100 times, to avoid sparsity issues.

The number of labels varies greatly across DW's language services. Moreover, we found for instance that only 25-30% of the labels could be manually aligned between English and German. The commonalities are mainly concentrated on the most frequent labels, reflecting a shared top-level division of the news domain, but the long tail exhibits significant independence across languages.

			English + Auxiliary → English							English + Auxiliary → Auxiliary						
		Models	de	es	pt	uk	ru	ar	fa	de	es	pt	uk	ru	ar	fa
$Y_{general}$	Mono	NN (Avg)				50.7				53.1	70.0	57.2	80.9	59.3	64.4	66.6
		HNN (Avg)				70.0				67.9	82.5	70.5	86.8	77.4	79.0	76.6
		HAN (Att)				71.2				71.8	82.8	71.3	85.3	79.8	80.5	76.6
	Multi	MHAN-Enc	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
		MHAN-Att	**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
		MHAN-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**
$Y_{specific}$	Mono	NN (Avg)				24.4				21.8	22.1	24.3	33.0	26.0	24.1	32.1
		HNN (Avg)				39.3				39.6	37.9	33.6	42.2	39.3	34.6	43.1
		HAN (Att)				43.4				44.8	46.3	41.9	46.4	45.8	41.2	49.4
	Multi	MHAN-Enc	45.4	45.9	44.3	41.1	42.1	44.9	41.0	43.9	46.2	39.3	47.4	45.0	37.9	48.6
		MHAN-Att	**46.3**	**46.0**	**45.9**	**45.6**	**46.4**	**46.4**	**46.1**	**46.5**	**46.7**	**43.3**	**47.9**	45.8	**41.3**	48.0
		MHAN-Both	45.7	45.6	41.5	41.2	45.6	44.6	43.0	45.9	46.4	40.3	46.3	**46.1**	40.7	**50.3**

Table 2: Full-resource classification performance (F_1) on general (top) and specific (bottom) topic categories using bilingual training with English as target (left) and the auxiliary language as target (right).

6 Evaluation

We evaluate our multilingual models on full-resource and low-resource scenarios of multilingual document classification over disjoint label sets on the Deutsche Welle corpus.

6.1 Settings

Following the typical evaluation protocol in the field, the corpus is split per language into 80% for training, 10% for validation and 10% for testing. We evaluate both type of labels (Y_g, Y_s) on a *full-resource scenario* and only the general topics (Y_g) on a *low-resource scenario*. We report the micro-averaged F1 scores for each test set, as in previous work (Hermann and Blunsom, 2014).

Model configuration. For all models, we use the aligned pre-trained 40-dimensional multilingual embeddings trained on the Leipzig corpus using multi-CCA from Ammar et al. (2016). The non-aligned embeddings are trained with the same method and data. We zero-pad documents up to a maximum of 30 words per sentence and 30 sentences per document. The hyper-parameters were selected on the validation sets. We made the following settings: 100-dimensional encoder and attention embeddings (at every level), relu activation function for all intermediate layers, batch size of 16, epoch size of 25k, and optimization using SGD with Adam until convergence.

All the hierarchical models have *Dense* encoders in both scenarios (Tables 2, 4, and 5), and *GRU* and *biGRU* in the full-resource scenario for English+Arabic (Table 3). For the low-resource scenario, we define three levels of data availability: *tiny* from 0.1% to 0.5%, *small* from 1% to 5% and *medium* from 10% to 50% of the original training set. We report the average F_1 scores on the test set for each level based on discrete increments of 0.1, 1 and 10 respectively. The decision threshold for the full-resource scenario is set to 0.4 for $|Y_s| < 400$ and 0.2 for $|Y_s| \geq 400$, and for the low-resource scenario it is 0.3 for all sets.

Baselines. We compare against the following monolingual neural networks, with shallow or hierarchical structure. These networks are based on the state of the art in the field, reviewed in Section 2, and thus represent strong baselines.

- **NN** : A neural network which feeds the average vector of the input words directly to a classification layer, as the common baseline for multilingual document classification by Klementiev et al. (2012).
- **HNN** : A hierarchical network with encoders and average pooling at every level, followed by a classification layer.
- **HAN**: A hierarchical network with encoders and attention, followed by a classification layer. This model is the one proposed by Yang et al. (2016) adapted to our task.

When training our multilingual system with the three options described in Section 4.1 noted as *Enc*, *Att* and *Both*, we use in fact the multilingual version of the third monolingual baseline, as illustrated in Fig. 3, trained with the objective of Eq. 8.

6.2 Results

Full-resource Scenario. Table 2 displays the results of full-resource document classification using Dense encoders for general and specific labels. In particular, on the left, the performance on the English sub-corpus is shown when English and an auxiliary sub-corpus are used for training, and

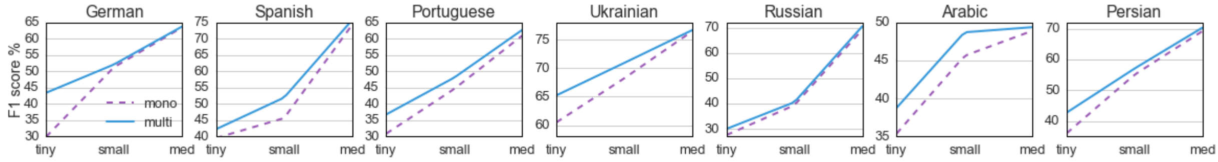


Figure 4: Low-resource document classification performance (F_1) of our *multilingual* attention network ensemble vs. a *monolingual* attention network on the DW corpus.

on the right the performance on the auxiliary sub-corpus is shown when that sub-corpus and English are used for training.

The multilingual model trained on pairs of languages outperforms on average all the examined monolingual models, from simple bag-of-words neural models to hierarchical neural models using average pooling or attention. The best-performing multilingual architecture is the one with shared attention across languages, especially when tested on English (documents and labels). Interestingly, this reveals that the transfer of knowledge across languages in a full-resource setting is maximized with language-specific word and sentence encoders, but language-independent (i.e. shared) attention for both words and sentences.

However, when transferring from English to Portuguese (en→pt), Russian (en→ru) and Persian (en→fa) on general categories, it is more effective to have only language-independent components. We hypothesize that this is due to the underlying commonness between the label sets rather than the relationship between languages. The consistent gain for English as target could be attributed to the alignment of the word embeddings to English and to the large number of English labels, which makes it easier to find multilingual labels from which to transfer knowledge.

We quantify now the impact of three important model choices on the performance: encoder type, word embeddings, and number of languages used for training. In Table 3, we observe that when we replace the *Dense* encoder layers with *GRU* or *biGRU* layers, the improvement from the multilingual training is still present. In particular, the multilingual models with shared attention are superior to alternatives regardless of the employed encoders. For reference, using simply logistic regression with bag-of-words (counts) for classification leads to F_1 scores of 75.8% in English and 81.9% in Arabic, using many more parameters than biGRU: 56.5M vs. 410k in English and 5.8M vs. 364k in Arabic.

In Table 4, we observe that when we train

	Encoders	Mono	Multi		
	$Y_{general}$	HAN	Enc	Att	Both
ar→en	Dense	71.2	70.0	**73.8**	68.9
	GRU	77.0	74.8	**77.5**	75.4
	biGRU	**77.7**	77.1	77.5	76.7
en→ar	Dense	80.5	79.0	**82.1**	79.1
	GRU	81.5	81.2	**83.4**	83.1
	biGRU	82.2	82.7	**84.0**	83.0

Table 3: Full-resource classification performance (F_1) for English-Arabic with various encoders.

Word emb.	$\|L\|$	$Y_{general}$ (n_l,	f_l)	$Y_{specific}$ (n_l,	f_l)
aligned	1	50K –	77.41 –	90K –	44.90 –
	2	40K ↓	78.30 ↑	80K ↓	45.72 ↑
	8	32K ↓	77.91 ↑	72K↓	45.82 ↑
non-aligned	8	32K ↓	71.23 ↓	72K ↓	33.41 ↓

Table 4: Average number of parameters per language (n_l), average F_1 per language (f_l), and their variation (arrows) with the number of languages $|L|$ and the word embeddings used for training.

our multilingual model (MHAN-att) on eight languages at the same time, the F_1 scores improve across languages – for both types of labels, general or specific – while the number of parameters per language decreases, by 36% for $Y_{general}$ and 20% for $Y_{specific}$. Lastly, when we train the same model with word embeddings that are not aligned across languages, the performance of the multilingual model drops significantly. An input space that is aligned across languages is thus crucial.

Low-resource Scenario. We assess the ability of the multilingual attention networks to transfer knowledge across languages in a low-resource scenario. The results for seven languages when trained jointly with English are displayed in detail in Table 5 and summarized in Figure 4. In all cases, at least one of the multilingual models outperforms the monolingual one, which demonstrates the usefulness of multilingual training for low-resource document classification.

Moreover, the improvements obtained from our multilingual models for lower levels of availability (*tiny* and *small*) are larger than in higher levels (*medium*). This is also clearly observed in Figure 4

	Size	Mono	Multi			
	$Y_{general}$	HAN	Enc	Att	Both	$\Delta\%$
en→de	0.1-0.5%	29.9	**41.0**	37.0	39.4	+37.2
	1-5%	51.3	51.7	49.7	**52.6**	+2.6
	10-50%	63.5	63.0	63.8	**63.8**	+0.5
en→es	0.1-0.5%	39.5	38.7	33.3	**41.5**	+4.9
	1-5%	45.6	45.5	**50.8**	50.1	+11.6
	10-50%	74.2	**75.7**	74.2	75.2	+2.0
en→pt	0.1-0.5%	30.9	25.3	31.6	**33.8**	+9.6
	1-5%	44.6	44.3	37.5	**47.3**	+6.0
	10-50%	60.9	61.9	62.1	**62.1**	+1.9
en→uk	0.1-0.5%	60.4	**62.4**	59.8	60.9	+3.1
	1-5%	68.2	67.7	**70.6**	69.0	+3.4
	10-50%	76.4	76.2	76.3	**76.7**	+0.3
en→ru	0.1-0.5%	27.6	26.6	27.0	**29.1**	+5.4
	1-5%	39.3	38.2	39.6	**40.2**	+2.2
	10-50%	69.2	**70.5**	70.4	69.4	+1.9
en→ar	0.1-0.5%	35.4	35.5	**39.5**	36.6	+11.7
	1-5%	45.6	**48.7**	47.2	46.6	+6.9
	10-50%	48.9	**52.2**	46.8	47.8	+6.8
en→fa	0.1-0.5%	36.0	35.6	33.6	**41.3**	+14.6
	1-5%	55.0	**55.6**	51.9	55.5	+1.0
	10-50%	69.2	**70.3**	70.1	70.0	+1.5

Table 5: Low-resource classification performance (F_1) with various sizes of training data.

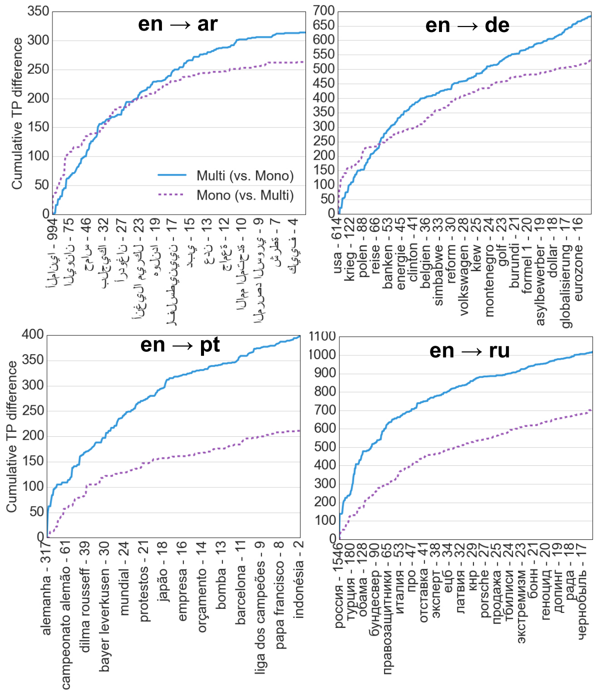


Figure 5: Cumulative true positive (TP) difference between *monolingual* and *multilingual* (ensemble) models for topic classification with *specific* labels, in the full resource scenario.

with our multilingual attention network ensemble, i.e. when we do model selection among the three multilingual variants on the development set. The best performing architecture in a majority of cases is the one which shares both the encoders and the attention mechanisms across languages. Moreover, this architecture also has the fewest number of parameters for the document modeling.

This promising finding for the low-resource scenario means that the classification performance can greatly benefit from the multilingual training (sharing encoders and attention) without increasing the parameters beyond that of a single monolingual document model. Nevertheless, in a few cases, we observe that the other two architectures, with increased complexity perform better than the "shared both" model. For instance, sharing encoders is superior to alternatives for Arabic language, i.e. the knowledge transfer benefits from shared word and sentence representations. Hence, to generalize to a large number of languages, we may need to consider more sophisticated models than sharing a monolingual document model across languages. Lastly, we did not generally observe a negative (or positive) correlation of the closeness between languages with the performance in the low-resource scenario, although the largest improvements were observed on languages more related to English (German, Spanish, Portuguese) than others (Arabic).

Overall, the above experiments pinpointed the most suitable general class of architectures for each language availiability setting rather than specific ones. Having said that, increasing the sophistication of the encoders is expected to improve accuracy across the board (as shown in Table 3).

6.3 Qualitative Analysis

We analyze the performance of the model over the full range of labels, to observe on which type of labels it performs better than the monolingual model, and provide some qualitative examples. Figure 5 shows the cumulative true positive (TP) difference between the monolingual and multilingual models on the Arabic and German, Portuguese and Russian test sets, ordered by label frequency. We can observe that the cumulative TP difference of multilingual model consistently keeps increasing as the frequency of the label decreases. This shows that labels across the entire range of frequencies benefit from joint training with English and not only a subset, e.g. only frequent labels.

For example, the top 5 labels on which the multilingual model performed better than the monolingual one for en → de were: *russland* (21), *berlin* (19), *irak* (14), *wahlen* (13) and *nato* (13), while for the opposite direction those were: *ger-*

many (259), *german* (97), *soccer* (73), *football* (47) and *merkel* (25). These topics are likely better covered in the respective auxiliary language which helps the multilingual model to better distinguish them in the target language. This is also observed in Figure 1, as the improved separation of topics using multilingual model vs. monolingual ones.

7 Conclusion

We proposed multilingual hierarchical attention networks for document classification and showed that they can benefit both full-resource and low-resource scenarios by using fewer parameters than with monolingual networks. In the former scenario, the most beneficial was to share only the attention mechanisms, while in the latter one, to share the encoders along with the attention mechanisms. These results confirm the merits of language transfer, which is also an important component of human language learning (Odlin, 1989; Ringbom, 2007). Moreover, our study broadens the applicability of multilingual document classification, since our framework is not restricted to common label sets.

There are several future directions for this study. In their current form, our models cannot generalize to languages without any example, as attempted by Firat et al. (2016b) for neural MT. This could be achieved by a label-size independent classification layer as in zero-shot classification (Qiao et al., 2016; Nam et al., 2016). Moreover, although we explored three distinct architectures, other configurations could be examined to improve document modeling, for example by sharing the attention mechanism at the sentence-level only. Lastly, the learning objective could be further constrained with sentence-level parallel information, to embed multilingual vectors of similar topics more closely together in the learned space.

Acknowledgments

We are greatful for the support from the European Union through its Horizon 2020 program (SUMMA project n. 688139, www.summa-project.eu). We would also like to thank Sebastião Miranda for crawling the news articles from Deutsche-Welle and the anonymous reviewers for their helpful suggestions.

References

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. In *Proc. of the Seventeenth Conference on Computational Natural Language Learning*, Sofia, Bulgaria.

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

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proc. of the 5th International Conference on Learning Representations*, San Diego, CA, USA.

Léon Bottou. 1998. On-line learning and stochastic approximations. In David Saad, editor, *On-line Learning in Neural Networks*, pages 9–42. Cambridge University Press.

Sarath Chandar, Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C. Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. In *Advances in Neural Information Processing Systems 27*, pages 1853–1861.

Jianshu Chen, Ji He, Yelong Shen, Lin Xiao, Xiaodong He, Jianfeng Gao, Xinying Song, and Li Deng. 2015. End-to-end learning of LDA by mirror-descent back propagation over a deep architecture. In *Advances in Neural Information Processing Systems 28*, pages 1765–1773, Montreal, Canada.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734, Doha, Qatar.

Misha Denil, Loris Bazzani, Hugo Larochelle, and Nando de Freitas. 2012. Learning where to attend with deep architectures for image tracking. *Neural Computation*, 24(8):2151–2184.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proc. of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471, Gothenburg, Sweden.

Daniel C. Ferreira, André F. T. Martins, and Mariana S. C. Almeida. 2016. Jointly learning to embed and predict with multiple languages. In *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2019–2028, Berlin, Germany.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016a. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proc. of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, California.

Orhan Firat, Baskaran Sankaran, Yaser Al-Onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016b. Zero-resource translation with multi-lingual neural machine translation. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pages 268–277, Austin, Texas.

Stephan Gouws, Yoshua Bengio, and Gregory S. Corrado. 2015. BilBOWA: Fast bilingual distributed representations without word alignments. In *Proc. of the 32nd International Conference on Machine Learning*, pages 748–756, Lille, France.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. In *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 58–68, Baltimore, Maryland.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proc. of the 28th International Conference on Neural Information Processing Systems*, NIPS'15, pages 1693–1701, Montreal, Canada.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. In *Neural Computation*, volume 9 (8), pages 1735–1780. MIT Press.

Yangfeng Ji and Noah Smith. 2017. Neural discourse structure for text categorization. *CoRR*, abs/1702.01829.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. *CoRR*, abs/1611.04558.

Rie Johnson and Tong Zhang. 2015. Effective use of word order for text categorization with convolutional neural networks. In *Proc. of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 103–112, Denver, Colorado.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pages 1746–1751, Doha, Qatar.

Diederik P. Kingma and Jimmy Lei Ba. 2014. Adam: A method for stochastic optimization. In *Proc. of the International Conference on Learning Representations*, Banff, Canada.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proc. of the International Conference on Computational Linguistics*, Bombay, India.

Ankit Kumar, Ozan Irsoy, Jonathan Su, James Bradbury, Robert English, Brian Pierce, Peter Ondruska, Ishaan Gulrajani, and Richard Socher. 2015. Ask me anything: Dynamic memory networks for natural language processing. In *Proc. of the 33rd International Conference on Machine Learning*, pages 334–343, New York City, NY, USA.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *Proc. of the 29th AAAI Conference on Artificial Intelligence*, pages 2267–2273, Austin, Texas.

Hugo Larochelle and Geoffrey Hinton. 2010. Learning to combine foveal glimpses with a third-order Boltzmann machine. In *Proc. of the 23rd International Conference on Neural Information Processing Systems*, pages 1243–1251, Vancouver, Canada.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proc. of the 31st International Conference on Machine Learning*, pages 1188–1196, Beijing, China.

Rui Lin, Shujie Liu, Muyun Yang, Mu Li, Ming Zhou, and Sheng Li. 2015. Hierarchical recurrent neural network for document modeling. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pages 899–907, Lisbon, Portugal.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal.

Laurens van der Maaten. 2009. Learning a parametric embedding by preserving local structure. In *Proc. of the 12th International Conference on Artificial Intelligence and Statistics*, pages 384–391, Clearwater Beach, FL, USA.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proc. of the International Conference on Learning Representations*, Scottsdale, AZ, USA.

Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. 2014. Recurrent models of visual attention. *CoRR*, abs/1406.6247.

Jinseok Nam, Eneldo Loza Mencía, and Johannes Fürnkranz. 2016. All-in text: Learning document, label, and word representations jointly. In *Proc. of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 1948–1954, Phoenix, AR, USA.

Terence Odlin. 1989. Language transfer: Cross-linguistic influence in language learning. In *Cambridge Applied Linguistics*. Cambridge University Press.

Nikolaos Pappas and Andrei Popescu-Belis. 2014. Explaining the stars: Weighted multiple-instance learning for aspect-based sentiment analysis. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pages 455–466, Doha, Qatar.

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

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, Doha, Qatar.

Ruizhi Qiao, Lingqiao Liu, Chunhua Shen, and Anton van den Hengel. 2016. Less is more: Zero-shot learning from online textual documents with noise suppression. In *Proc. of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2249–2257, Las Vegas, NV, USA.

Hakan Ringbom. 2007. *Cross-linguistic Similarity in Foreign Language Learning*. Second language acquisition. Multilingual Matters.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal.

Duyu Tang, Bing Qin, and Ting Liu. 2015. Document modeling with gated recurrent neural network for sentiment classification. In *Empirical Methods on Natural Language Processing*, pages 1422–1432, Lisbon, Spain.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proc. of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394, Uppsala, Sweden.

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

Ainur Yessenalina, Yisong Yue, and Claire Cardie. 2010. Multi-level structured models for document-level sentiment classification. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pages 1046–1056, Cambridge, MA.

Wen-tau Yih, Kristina Toutanova, John C. Platt, and Christopher Meek. 2011. Learning discriminative projections for text similarity measures. In *Proc. of the 15th Conference on Computational Natural Language Learning*, pages 247–256, Portland, OR, USA.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28*, pages 649–657, Montreal, Canada.

Will Y. Zou, Richard Socher, Daniel Cer, and Christopher D. Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398, Seattle, Washington, USA.