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
HEIG-VD / HES-SO
Route de Cheseaux 1
CH-1401 Yverdon, Switzerland
andrei.popescu-belis@heig-vd.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 shared 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 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.

Previous studies in document classification attempted to address 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 doc-

Figure 1: Vectors of documents labeled with ‘Europe’, ‘Culture’ and their Arabic counterparts. The multilingual hierarchical attention network separates topics better than monolingual ones.
ument structures across languages for document classification with disjoint label sets, as opposed to training language-specific training of hierarchical attention networks (HANs) (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), in 8 languages from Deutsche Welle’s news website.1 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, which are costly to produce and are often unavailable. Here, we remove this constraint, and 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) proposed a hierarchical attention network (HAN) with bi-directional gated encoders which 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, e.g. for optical character recognition (Larochelle and Hinton, 2010), image tracking (Denil et al., 2012), or image classification (Mnih et al., 2014). For text classification, studies which aimed to learn the importance of sentences included those by Yessenalina et al. (2010); Pappas and Popescu-Belis (2014); Yang et al. (2016) and more recently those by Pappas and Popescu-Belis (2017); Ji and Smith (2017). For NMT, Bahdanau et al. (2015) proposed an attention-based encoder-decoder network, while Luong et al. (2015) proposed a local and ensemble attention model. 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 an-

swering. 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 here to share attention across languages, at one or more levels of hierarchical document models, 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, displayed in Figure 2, which is derived from the one proposed by Yang et al. (2016). Our architecture is general in the sense that it defines only the hierarchical structure, but accommodates different types of individual components, i.e. encoders and attention models. We consider a dataset $D = \{(x_i, y_i), i = 1, \ldots, N\}$ made of $N$ documents $x_i$ with labels $y_i \in \{0, 1\}^K$. Each document is represented by the sequence of $d$-dimensional embeddings of their words grouped into sentences, $x_i = \{w_{i1}, w_{i2}, \ldots, w_{iKT}\}$, $T$ 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 word level is made of an encoder $g_w$ with parameters $H_w$ and an attention model $\alpha_w$ with parameters $A_w$, while the sentence level similarly includes an encoder and an attention model ($g_s, H_s$ and $\alpha_s, A_s$). The output $u_i$ is used by the classification layer to determine $y_i$.

3.1 Encoder Layers

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

$$h_w^{(it)} = \{g_w(w_{it}) | t = 1, \ldots, K\}$$  

At the sentence level, after combining the intermediate word vectors $\{h_w^{(it)} | t = 1, \ldots, T\}$ to a sentence vector $s_i$ (as explained in 3.2), the function $g_s$ encodes the sequence of sentence vectors $\{s_i | i = 1, \ldots, K\}$, 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$^2$, 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, $\vec{g}_w$:

$$h_w^{(it)} = [\vec{g}_w(h_w^{(it)}); \vec{g}_w(h_w^{(it)})]$$

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 $\alpha_w$ and $\alpha_s$) 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:

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

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:

$$\frac{1}{K} \sum_{i=1}^{K} \alpha_s^{(i)} h_s^{(i)} = \frac{1}{K} \sum_{i=1}^{K} \sum_j \exp(v_{ij}^\top u_s) h_s^{(i)}$$

$^2$GRU is a simplified version of Long-Short Term Memory, LSTM (Hochreiter and Schmidhuber, 1997).
where \( v_i = f_x(h_s^{(i)}) \) 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 \( A_s = \{ W_s, u_s \} \).

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 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 may have more than one likely labels independent of each other, as we verified empirically in our preliminary experiments. Hence, we replace the softmax with a sigmoid function, so that for each document \( i \) represented by the vector \( u_i \) we model the probability of the \( k \) labels as follows:

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

(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:

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

(6)

where \( \theta \) 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 (SGD) (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 in this case the needed parameters grow linearly with the number of languages. Moreover, this does not exploit common knowledge across languages or to transfer it from one to another. We propose here a HAN with shared components across languages, which has slower parameter growth (hence sublinear) compared to monolingual ones and enables knowledge transfer across languages. We now consider \( M \) languages noted \( L = \{L_l | l = 1, \ldots, M\} \), and a multilingual set of topic-labeled documents \( D_l = \{(x^{(l)}_i, y^{(l)}_i) | i = 1, \ldots, N_l, l = 1, \ldots, M\} \) defined as above (Section 3).

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, noted \( \theta_{enc} = \{H_w, W_w^{(l)}, H_s, W_s^{(l)}, W_c\} \); (b) sharing the parameters of word and sentence attention models, noted \( \theta_{att} = \{H_w^{(l)}, W_w^{(l)}, H_s^{(l)}, W_s^{(l)}\} \); and (c) sharing both previous sets of parameters, noted \( \theta_{both} = \{H_w^{(l)}, W_w^{(l)}, H_s^{(l)}, W_s^{(l)}, W_c^{(l)}\} \). 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:

\[
|\theta_{mono}| > |\theta_{enc}| > |\theta_{att}| > |\theta_{both}|
\]

(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 necessarily language-specific, 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. The word embeddings are not considered as parameters in our setup because

Figure 3: Multilingual hierarchical attention networks for modeling documents and classifying them over disjoint label sets.
they are fixed during training. For learned word embeddings, the argument still holds if we consider their parameters as part of the word-level encoder.

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 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 lack of parallel information, as we show in Section 6.2, Table 4.

### 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 $\theta_1, \ldots, \theta_M$, across different language networks:

$$
L(\theta_1, \ldots, \theta_M) = -\frac{1}{Z} \sum_{i}^{N_e} \sum_{l}^{M} \mathcal{H}(y_i^{(l)}, \hat{y}_i^{(l)}) \quad (8)
$$

where $Z = M \times N_e$ and $N_e$ is the epoch size.3

The joint objective $L$ can be minimized with respect to the parameters $\theta_1, \ldots, \theta_M$ using SGD as before. However, when training on examples from different languages consecutively it is difficult to learn a shared space that works well across languages. This is because updates for each language apply only on a subset of parameters and may bias the model away from other languages. To address this issue, we employ the training strategy proposed by (Firat et al., 2016a), who sampled parallel sentences for multi-way machine translation from different language pairs in a cyclic fashion at each iteration.4 Here, we sample a document-label pair from each language at iteration. For mini-batch SGD, the number of samples per language is equal to the batch size divided by $M$.

### Table 2: Statistics of the Deutsche Welle corpus: $\bar{s}$ and $\bar{w}$ are the average numbers of sentences and words per document.

<table>
<thead>
<tr>
<th>Languages</th>
<th>Documents</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L$</td>
<td>$</td>
<td>X</td>
</tr>
<tr>
<td>English</td>
<td>112,816</td>
<td>17.9</td>
</tr>
<tr>
<td>German</td>
<td>132,709</td>
<td>22.3</td>
</tr>
<tr>
<td>Spanish</td>
<td>75,827</td>
<td>13.8</td>
</tr>
<tr>
<td>Portuguese</td>
<td>39,474</td>
<td>20.2</td>
</tr>
<tr>
<td>Ukrainian</td>
<td>35,423</td>
<td>17.6</td>
</tr>
<tr>
<td>Russian</td>
<td>108,076</td>
<td>16.4</td>
</tr>
<tr>
<td>Arabic</td>
<td>57,697</td>
<td>13.3</td>
</tr>
<tr>
<td>Persian</td>
<td>36,282</td>
<td>18.7</td>
</tr>
<tr>
<td>All</td>
<td>598,304</td>
<td>17.32</td>
</tr>
</tbody>
</table>

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 multiple categories may apply to each document. For instance, this is the case for online multilingual news agencies, which must keep track of news topics across languages.

Two datasets for multilingual document classification have been used in previous studies: Reuters RCV1/RCV2 (6,000 documents, 2 languages and 4 labels), introduced by (Klementiev et al., 2012), 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, shown in Table 2. The DW dataset contains nearly 600,000 documents, in 8 languages, annotated by journalists with several topic labels. 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 ($Y_s$) both described by one or more words. We consider (and count in Table 2) only those specific labels that appear at least 100 times, to avoid sparsity issues.

The number of labels varies greatly across the
6 Evaluation

6.1 Settings

We evaluate our multilingual models on full-resource and low-resource scenarios of multilingual document classification on the Deutsche Welle corpus. 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 topic labels (\(Y_g\)) on a low-resource scenario. We report the micro-averaged F1 scores for each test set, as in previous work (e.g., Hermann and Blunsom, 2014).

**Model configuration.** For all models, we use the aligned 40-dimensional multilingual embeddings pre-trained on the Leipzig corpus using multi-CCA from Ammar et al. (2016). The non-aligned embeddings used for comparison purposes 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 1, 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 F1 scores on the test set for each level based on discrete increments of 0.1, 1 and 10 percentage points respectively. The decision threshold for the value of \(p\) in Eq. 5 for the full-resource scenario is set to 0.4 for labels such that \(|Y_g| < 400\) and 0.2 for \(|Y_g| \geq 400\), and for the low-resource scenario it is 0.3 for all sets. For the ensemble in the low-resource setting, we train the three proposed multilingual models and choose the optimal one based on the validation data for each language respectively (see Fig. 4).

**Baselines.** We compare against the following monolingual neural networks, with shallow or hierarchical structures. 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 (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.

Our multilingual models with the three sharing
configurations from Section 4.1, are noted as Enc, Att and Both. Their implementation amounts to, first, creating a HAN model for each language, second, sharing components across multiple languages as illustrated in Fig. 3, and, third, training them with the objective of Eq. 8.

6.2 Results

Full-resource scenario. Table 1 displays the results of full-resource document classification using DENSE encoders for general and specific labels. On the left side, the performance on the English sub-corpus is shown when English and an auxiliary sub-corpus are used for training, and on the right side, the performance on the auxiliary sub-corpus is shown when that sub-corpus and the English sub-corpus are used for training.

The multilingual model trained on pairs of languages outperforms on average all the examined monolingual models, namely a bag-of-word neural model and two hierarchical neural models which use average pooling and attention respectively. The best-performing multilingual model bilingually on average is the one with shared attention across languages, especially when tested on English. The consistent gain for English as target could be attributed to the alignment of the word embeddings to English and to the many English labels, which makes it easier to find multilingual labels from which to transfer knowledge. 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 to a relationship between languages, which is hard to identify on linguistic grounds.

We will now quantify 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, when we train our multilingual model (MHAN-att) on eight languages at the same time, the \( F_1 \) score improves on average across languages – for both types of labels, general or specific – while the number of parameters per language decrease, by 36% for \( Y_{\text{general}} \) and 20% for \( Y_{\text{specific}} \). 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, i.e. training on a fraction of the available data, as defined in 6.1 above. 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 availabili-

<table>
<thead>
<tr>
<th>Encoders</th>
<th>Mono</th>
<th>Multi</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y_{\text{general}} )</td>
<td>HAN</td>
<td>Enc</td>
</tr>
<tr>
<td>DENSE</td>
<td>71.2</td>
<td>70.0</td>
</tr>
<tr>
<td>GRU</td>
<td>77.0</td>
<td>74.8</td>
</tr>
<tr>
<td>biGRU</td>
<td>77.7</td>
<td>77.1</td>
</tr>
<tr>
<td>DENSE</td>
<td>80.5</td>
<td>79.0</td>
</tr>
<tr>
<td>GRU</td>
<td>81.5</td>
<td>81.2</td>
</tr>
<tr>
<td>biGRU</td>
<td>82.2</td>
<td>82.7</td>
</tr>
</tbody>
</table>

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

| Word embeddings \(| L |\) | \( n_l \) | \( f_l \) | \( n_L \) | \( f_L \) |
|----------|------|------|------|------|
| Aligned | 50K | 77.41 | 90K | 44.90 | ↓ |
| 40K | 78.30 | 80K | 45.72 | ↑ |
| 32K | 77.91 | 72K | 45.82 | ↑ |
| Non-aligned | 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.
we could generalize to a large number of languages, we shared word and sentence representations. Hence, language, i.e. the knowledge transfer benefits from using encoders is superior to alternatives for Arabic than the “shared both” model. For instance, sharing architectures with increased complexity perform better in a few cases, we observe that the other architectures may need to consider more dynamic models which are able to choose for each language individually which sharing scheme is the most appropriate for transferring from another language. Lastly, we did not generally observe a negative (or positive) correlation of the similarity 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 multilingual sharing scheme (Figure 3) for each setting independently of the encoder type, rather than the optimal combination of sharing scheme and encoder. Therefore, as shown in Table 3, increasing the sophistication of the encoders (from DENSE to GRU to biGRU) is expected to further improve accuracy.

### 6.3 Qualitative Analysis

We analyze the performance of the multilingual 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, German, Portuguese and Russian test sets, ordered by label frequency. We can observe that the cumulative TP difference of the multilingual model consistently increases as the frequencies of the labels decrease. This shows that labels across the entire range of frequencies benefit from joint training with English and not only a subset, for example only the highly 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: germany (259), german (97), soccer (73), football (47) and merkel (25). These topics are likely better covered in the respective auxiliary language which helps

<table>
<thead>
<tr>
<th>Size</th>
<th>Mono Enc Att Both</th>
<th>Multi Enc Att Both</th>
<th>∆%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HAN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.1-0.5</td>
<td>29.9</td>
<td>41.0</td>
<td>37.0</td>
</tr>
<tr>
<td>1-5</td>
<td>51.3</td>
<td>51.7</td>
<td>49.7</td>
</tr>
<tr>
<td>10-50%</td>
<td>63.5</td>
<td>63.0</td>
<td>63.8</td>
</tr>
<tr>
<td>0.1-0.5</td>
<td>39.5</td>
<td>38.7</td>
<td>33.3</td>
</tr>
<tr>
<td>1-5</td>
<td>45.6</td>
<td>45.5</td>
<td>50.8</td>
</tr>
<tr>
<td>10-50%</td>
<td>74.2</td>
<td>75.7</td>
<td>74.2</td>
</tr>
<tr>
<td>0.1-0.5</td>
<td>30.9</td>
<td>25.3</td>
<td>31.6</td>
</tr>
<tr>
<td>1-5</td>
<td>44.6</td>
<td>44.3</td>
<td>37.5</td>
</tr>
<tr>
<td>10-50%</td>
<td>60.9</td>
<td>61.9</td>
<td>62.1</td>
</tr>
<tr>
<td>0.1-0.5</td>
<td>60.4</td>
<td>62.4</td>
<td>59.8</td>
</tr>
<tr>
<td>1-5</td>
<td>68.2</td>
<td>67.7</td>
<td>70.6</td>
</tr>
<tr>
<td>10-50%</td>
<td>76.4</td>
<td>76.2</td>
<td>76.3</td>
</tr>
<tr>
<td>0.1-0.5</td>
<td>27.6</td>
<td>26.0</td>
<td>27.0</td>
</tr>
<tr>
<td>1-5</td>
<td>39.3</td>
<td>38.2</td>
<td>39.6</td>
</tr>
<tr>
<td>10-50%</td>
<td>69.2</td>
<td>70.5</td>
<td>70.4</td>
</tr>
<tr>
<td>0.1-0.5</td>
<td>35.4</td>
<td>35.5</td>
<td>39.5</td>
</tr>
<tr>
<td>1-5</td>
<td>45.6</td>
<td>48.7</td>
<td>47.2</td>
</tr>
<tr>
<td>10-50%</td>
<td>48.9</td>
<td>52.2</td>
<td>46.8</td>
</tr>
<tr>
<td>0.1-0.5</td>
<td>36.0</td>
<td>35.6</td>
<td>33.6</td>
</tr>
<tr>
<td>1-5</td>
<td>55.0</td>
<td>55.6</td>
<td>51.9</td>
</tr>
<tr>
<td>10-50%</td>
<td>69.2</td>
<td>70.3</td>
<td>70.1</td>
</tr>
</tbody>
</table>
The multilingual model to better distinguish them in the target language as well. This is also observed in Figure 1 presented in the introduction, through an 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 settings, while using fewer parameters than monolingual networks. In the former setting, the best option was to share only the attention mechanisms, while in the latter one, it was sharing 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 machine translation. This could be achieved by a classification layer independent of the size of the label set 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 grateful for the support from the European Union through its Horizon 2020 program in the SUMMA project n. 688139, see http://www.summa-project.eu. We would also like to thank Sebastiâo Miranda at Priberam for gathering the news articles from Deutsche Welle and the anonymous reviewers for their helpful suggestions. The second author contributed to the paper while at the Idiap Research Institute.

References


Jianshu Chen, Ji He, Yelong Shen, Lin Xiao, Xiaodong He, Jianfeng Gao, Xinying Song, and Li Deng.


