Multilingual Hierarchical Attention Networks for Document Classification

Nikolaos Pappas[1]
Andrei Popescu-Belis[1,2]

[1]Idiap Research Institute, Martigny, Switzerland
[2]School of Management and Engineering Vaud (HEIG-VD)

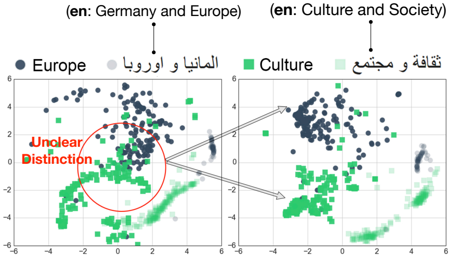
November 30, 2017

Document Representation Learning

"Learning representations which capture the underlying structural and semantic properties of a document."

Why is it important?

- Distills information
- Models linguistic structure
- Helps solving various tasks: *classification*, *summarization*



(a) Monolingual models (b) Multilingual model

Objectives of this study

→ Effectively transfer task knowledge across languages
→ Efficiently scale to many languages

Document Classification in Multiple Languages

Given $D^{(l)} = \{(x_i^{(l)}, y_i^{(l)}) \mid i = 1, \ldots, N_l\}$, a multilingual document collection with $l = 1, \ldots, M$ languages

- Documents: $x_i^{(l)} = \{w_{11}^{(l)}, w_{12}^{(l)}, \ldots w_{ST}^{(l)}\}$
- Labels: $y_i^{(l)} \in \{0, 1\}^{k_l}$

Goal:

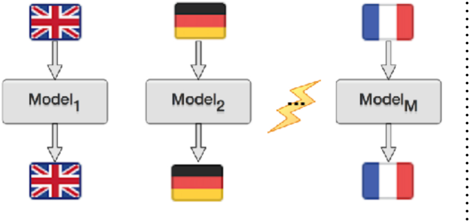
- Estimate conditional probability $p(y^{(l)}|x^{(l)})$ for any language l

Challenges:

- No document or label alignment is available

Document Classification: Monolingual Approach

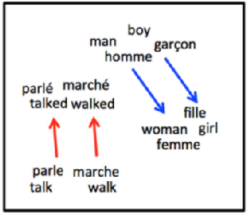
- Learn separate models $f^{(l)} : X^{(l)} \rightarrow Y^{(l)}$
  - Hierarchical document modeling ✓
  - No cross-language transfer ✗
  - Does not scale well to many languages ✗

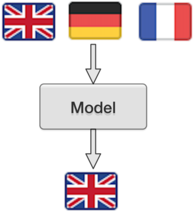


(Kim, 2014)
(Tang et al., 2015)
(Lin et al., 2015)
(Yang et al., 2016)

Document Classification: Multilingual Approach

- Learn one model $f : X \rightarrow Y$ with an aligned input and label space
  - No hierarchy, simple composition ✗
  - Cross-language transfer ✓ only with label alignment ✗
  - Scales well to many languages ✓ only with label alignment ✗





(Klementiev et al., 2012)

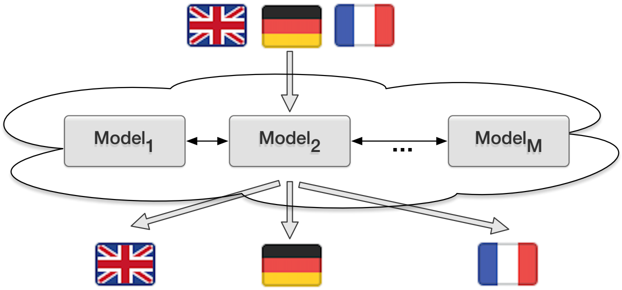
(Herman and Blunsom, 2014)

(Gouws et al., 2015)

(Ammar et al., 2016)

Document Classification: Our Approach

- Learn a multilingual model $f : X \rightarrow Y^{(l)}$ trained with multi-task learning and an aligned input space across languages
    - Hierarchical document modeling ✓
    - Cross-language transfer ✓
    - Scales well to many languages ✓



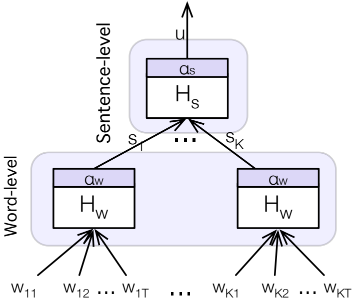
Introduction and Background
Motivation
The Problem
Previous Studies
Our Contribution

Multilingual Hierarchical Attention Networks
Hierarchical Document Modeling
Component Sharing Schemes
Training Strategy

Evaluation
Dataset and Settings
Quantitative Results
Qualitative Analysis

Conclusion

Hierarchical Attention Network



Input: Sequence of words $w_{ij} \in R^d$ (aligned)
Output: Document vector

Word-level Abstraction

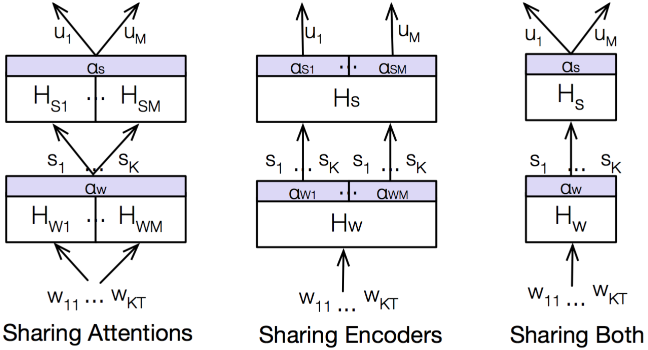
- Encoder layer
  $h_w^{(it)} = \{g_w(w_{it}) | \ t = 1, \ldots, T\}$
- Attention layer
  $s_i = \frac{1}{T} \sum_{t=1}^{T} a_w^{(it)} h_w^{(it)} \in R^{d_w}$

Sentence-level Abstraction

- Encoder layer
  $h_s^{(i)} = \{g_s(s_i) \mid i = 1, \ldots, K\}$
- Attention layer
  $u = \frac{1}{K} \sum_{i=1}^{K} a_s^{(i)} h_s^{(i)} \in R^{d_s}$

Sharing Components across Languages

1. Sharing attention layers (MHAN-att)
   - Enforces universal attention and language-specific encoders
2. Sharing encoder layers (MHAN-enc)
   - Enforces universal encodings and language-specific attention
3. Sharing both (MHAN-both)



Multilingual Output Layer and Training

For multi-label classification each vector u is input to a sigmoid layer:

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

Training objective

- Minimize the sum of cross-entropy errors

$$\mathcal{L}(\theta_1, \ldots, \theta_M) = -\frac{1}{NM}\sum_i^N \sum_l^M \mathcal{H}(y_i^{(l)}, \hat{y}_i^{(l)})$$

- Mix languages at each iteration i by sampling document-label pairs $t_i^l = (x_*^{(l)}, y_*^{(l)})$ for each language l

$$(t_1^1, \ldots, t_1^M) \rightarrow (t_2^1, \ldots, t_2^M) \rightarrow \ldots$$

Optimization

Stochastic gradient descent

Impact on Parameters

- The set of parameters for each model from $L = 1, \ldots, M$ are:
  - $\theta_{mono} = \{H_w^{(L)}, A_w^{(L)}, H_s^{(L)}, A_s^{(L)}, W_c^{(L)}\}$
  - $\theta_{enc} = \{H_w, A_w^{(L)}, H_s, A_s^{(L)}, W_c^{(L)}\}$
  - $\theta_{att} = \{H_w^{(L)}, A_w, H_s^{(L)}, A_s, W_c^{(L)}\}$
  - $\theta_{both} = \{H_w, A_w, H_s, A_s, W_c^{(L)}\}$
- Assuming fully-connected networks, we have the following:

$$|\theta_{mono}| > |\theta_{enc}| > |\theta_{att}| > |\theta_{both}|$$

e.g. Classification with 8 languages (average #params and F1 score)

HAN (1 language, aligned)	50K –	77.41 –
MHAN-att (2 languages, aligned)	40K ↓	78.30 ↑
MHAN-att (8 languages, aligned)	32K ↓	77.91 ↑
MHAN-att (8 languages, non-aligned)	32K ↓	71.23 ↓

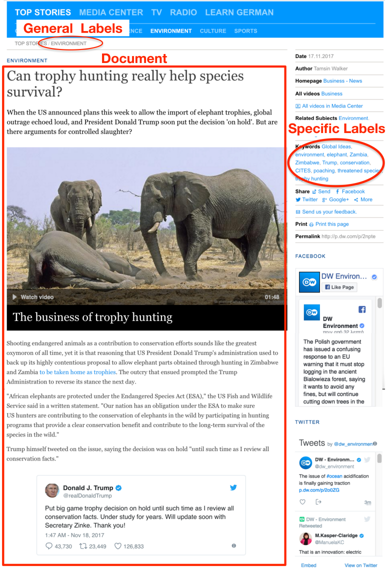
Introduction and Background
- Motivation
- The Problem
- Previous Studies
- Our Contribution

Multilingual Hierarchical Attention Networks
- Hierarchical Document Modeling
- Component Sharing Schemes
- Training Strategy

Evaluation
- Dataset and Settings
- Quantitative Results
- Qualitative Analysis

Conclusion

Deutsche Welle: A Large Multilingual Dataset



→ News articles from `dw.com`

- 600k documents, 8 languages
- Labels assigned by journalists
- Evaluation splits 80-10-10 (%)

		Labels	
Language	Documents	General	Specific
English	112,816	327	1,058
German	132,709	367	809
Spanish	75,827	159	684
Portuguese	39,474	95	301
Ukrainian	35,423	28	260
Russian	108,076	102	814
Arabic	57,697	91	344
Persian	36,282	71	127

Evaluation Settings

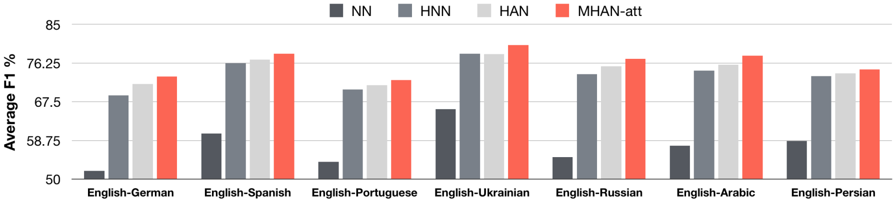
1. Full-resource scenario (English + other $\rightarrow$ both)
   - Train on the full set of documents for every language
2. Low-resource scenario (English + target $\rightarrow$ target)
   - Train on a subset of documents for the target language

Baselines

- *NN*: Logistic Regression + averaging (Klementiev et al. 2012)
- *HNN*: Hierarchical Network + averaging (Tang et al. 2015)
- *HAN*: Hierarchical Network + attention (Yang et al. 2016)

Input: 40-dim (Ammar et al. 2016), Encoders: Dense|GRU|biGRU 100-dim, Attention: Dense 100-dim, Activation: ReLU, Optimizer: ADAM, Batch size: 16, Epoch size: 25,000, Maximum epoch: 200 x $|L|$, Metric: F1-score

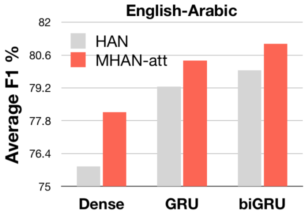
Full-Resource Scenario



→ Multilingual models outperform monolingual ones

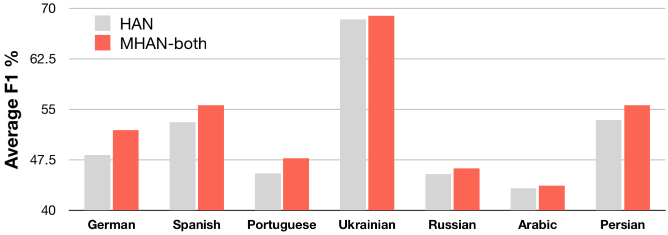
→ Sharing attention mechanisms is the optimal sharing scheme

→ Improvement holds for various encoders

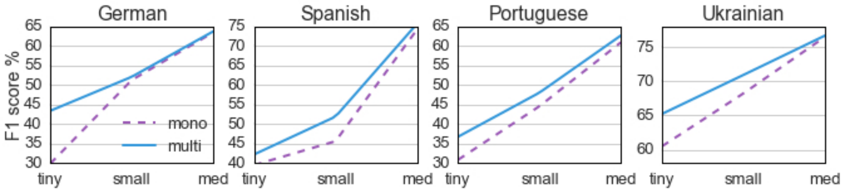


Low-Resource Scenario

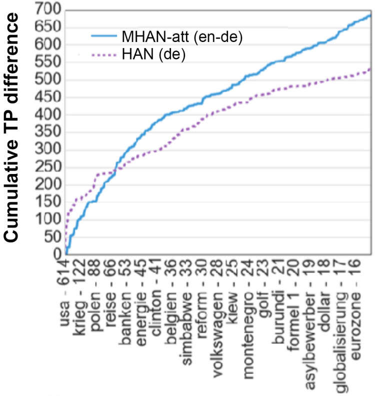
→ Sharing both attention and encoders is the best configuration



→ Multilingual models are most helpful on a very low-resource setting



Where does the improvement come from?



→ Gains across the full label frequency spectrum

→ Most improved German labels
- *russland*, *irak* and *nato*

→ Most improved English labels
- *germany*, *football* and *merkel*

Introduction and Background
Motivation
The Problem
Previous Studies
Our Contribution

Multilingual Hierarchical Attention Networks
Hierarchical Document Modeling
Component Sharing Schemes
Training Strategy

Evaluation
Dataset and Settings
Quantitative Results
Qualitative Analysis

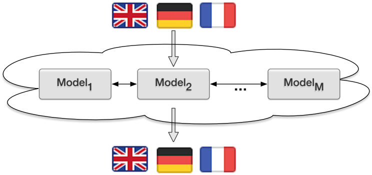
Conclusion

Conclusion

- Multilingual hierarchical models are able to learn robust representations for classification
  - → Competitive against monolingual models (full, low)
  - → Require fewer parameters than them
- New large dataset for multilingual representation learning

Future work:

- Leverage the aligned space in the output layer
- Investigate more powerful configurations and apply to other tasks



Code and data are available through Idiap's Github repository:

`http://github.com/idiap/mhan`

Thank you! Any questions?

Acknowledgments