Polyglot Text Classification with Neural Document Models

Suchin Gururangan

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2018

Committee:

Noah A. Smith, Chair

Ryan A. Georgi

Program Authorized to Offer Degree:
Linguistics

University of Washington

Abstract

Polyglot Text Classification with Neural Document Models

Suchin Gururangan

Chair of the Supervisory Committee:
Noah A. Smith
Computer Science

Sometimes, annotating data for text classification is expensive, so one must rely on techniques like parameter sharing and semi-supervised learning to improve classification performance in low-resource environments. In this thesis, I combine a generative, neural document model (Card et al., 2018) and multilingual word vectors (Ammar et al., 2016) to perform text classification on documents in eight languages. The model I propose jointly trains on labeled and unlabeled data from multiple languages, and incorporates additional document-level metadata, such as language ID, in its generative story. Through a series of experiments, I show that the model significantly outperforms monolingual baselines in low-resource environments.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I'd like to express sincere gratitude to my advisor Noah Smith for introducing me to NLP research, giving me the opportunity to work in the ARK group, and advising me throughout my graduate studies. I would also like to thank Ryan Georgi for being my thesis reader. His extensive feedback was crucial in helping me refine this thesis. Thanks to Roy Schwartz, Phoebe Mulcaire, and Dallas Card for useful discussions throughout the project.

Mom, Dad, and Sunjeev – thank you for your love and support. I would not be where I am today without you all.

Swabha – thank you for your advice, patience, and partnership. Your faith in me is the foundation of this work.

Chapter 1

INTRODUCTION

Text classification is a central problem in natural language processing (NLP) in which a model is presented with a piece of text and must categorize it into one of n classes. This task is useful in a variety of real-world applications such as sentiment analysis (Pang et al., 2008), spam detection (Ntoulas et al., 2006), and topic classification (Lee et al., 2011).

As more languages around the world come online, it is increasingly critical to build text classifiers that perform strongly in multiple languages. However, because annotated data is often expensive to produce, many target languages remain low-resource, without adequate amounts of labeled data for supervision. The low-resource problem is widely studied (Georgi, 2016; Agić et al., 2016; Tsvetkov and Dyer, 2016; Gormley et al., 2014; Bender, 2011).

Potential solutions to the low-resource problem arise from crosslingual studies and semi-supervised learning. Many languages share syntactic and morphological characteristics, as a result of common membership in *language families* and *shared ancestry* (Bender, 2011). Shared traits can also be a result of *borrowing*, in which linguistic units are transferred as a result of contacts between communities speaking different languages (Thomason and Kaufman, 2001). In fact, a rich body of work has shown that crosslingual distributed representations of words improve the downstream accuracy of models on a variety of tasks (Ruder et al., 2017).

Recent studies have also shown that models can achieve state-of-the-art performance on classification tasks with far less supervision by exploiting unlabeled data (Howard and Ruder, 2018; Radford et al., 2018). These findings suggest that one may be able to improve classification performance in low-resource settings by 1) incorporating data from other languages and 2) leveraging unlabeled data during training. Previous works have corroborated this hy-

pothesis for tasks like semantic role labeling (Gormley et al., 2014) and language modeling (Adams et al., 2017).

Generative models of text are a promising fit for crosslingual, semi-supervised text classification, as they directly enable the inclusion of informative priors and unlabeled data (Nigam et al., 2000; Kingma et al., 2014), which are helpful in the low-resource setting when there may be language-specific knowledge available. Advances in variational inference and deep learning (Kingma and Welling, 2013) have enabled efficient Bayesian models of text that can approximate complex posterior distributions in a black-box manner (Miao et al., 2015; Srivastava and Sutton, 2017).

In this thesis, I apply a generative document model (Card et al., 2018) across eight languages in the RCV2 document classification task (Lewis et al., 2004). I explore crosslingual and semi-supervised learning under simulated low-resource environments, in which training data of a target language is subsampled. Crucially, this thesis uses the same model architecture for all experimental settings. The approach is out-of-the-box, effective under varying amounts of supervision, and requires a limited amount of pre-processing to model documents in multiple languages.

1.1 Contributions

This thesis contributes the following:

- I apply an existing *generative document model* (SCHOLAR; Card et al. (2018)) to the RCV2 document classification task.

- I use *multilingual word vectors* to enforce the semantic coherence of topics across languages, unlike previous multilingual topic models that use parallel text or word matching priors (Ni et al., 2011; Mimno et al., 2009; Boyd-Graber and Blei, 2009).

- I show that *crosslingual parameter sharing* improves classification performance on target languages under low-resource settings, outperforming monolingual baselines.

- I show that incorporating *unlabeled data* improves classification performance on target languages under low-resource settings, outperforming supervised baselines.

- I release a *Python package* for polyglot text classification, with code for all models and experiments, on Github: `https://github.com/kernelmachine/multilingual-vae`

The remainder of this thesis is organized as follows:

- **Chapter 2** introduces related work for this thesis.

- **Chapter 3** introduces the data used for these experiments.

- **Chapter 4** explores supervised, monolingual text classification.

- **Chapter 5** explores supervised, crosslingual text classification with parameter sharing.

- **Chapter 6** explores semi-supervised, crosslingual text classification.

- **Chapter 7** presents concluding remarks.

Chapter 2

RELATED WORK

In this chapter, I describe related work that inspires this thesis. First, I describe related work on generative document models in a single language. Then, I discuss related work on multilingual document models, representation learning, and text classification. Finally, I describe related work on semi-supervised text classification.

2.1 Monolingual Document Models

Early approaches to probabilistic document models include the popular latent Dirichlet allocation (LDA) (Blei, 2012) and its variants (e.g., Mcauliffe and Blei (2008); Blei and Lafferty (2006)). These models are based upon the idea that there exist latent topics that determine how words in documents have been generated. While LDA has enjoyed successful application in a variety of fields (Boyd-Graber et al., 2017), it is hampered by the difficulty of deriving inference algorithms for computing posterior distributions when modeling assumptions are changed (Srivastava and Sutton, 2017).

Auto-encoding variational Bayes (AEVB; Kingma and Welling (2013)) is a novel inference technique for probabilistic models that circumvents these issues by approximating the posterior distribution with a neural network. Auto-encoding variational Bayes for topic models (AVITM; Srivastava and Sutton (2017)) is an AEVB-based inference method tailored to topic modeling. It is computationally efficient and produces coherent topics, while absolving the need to derive problem-specific inference algorithms.

SCHOLAR (Card et al., 2018) builds upon AVITM to incorporate both metadata and labels as inputs to the generative story and inference procedure. I leverage SCHOLAR in this work, and explore it in detail in Chapter 4.

2.2 Multilingual Document Models

Topic models capture global, document-level structure as opposed to fine-grain syntactic or semantic details of a language, like word order (Griffiths et al., 2007). This feature makes topic models apt for crosslingual studies (Vulić et al., 2015). However, most topic models rely heavily on word-document co-occurrences to generate coherent representations of documents. This is not an issue in monolingual corpora, since analogous words tend to be used in similar contexts. However, in the multilingual setting, generative models must be enhanced with different priors to bridge languages together, since translationally equivalent words will usually never appear in the same document (Boyd-Graber and Blei, 2009).

Multilingual topic models have been used in crosslingual tasks such as event clustering (De Smet and Moens, 2009), document classification (Ni et al., 2011), and word translation (Vulić et al., 2011; Vulić and Moens, 2012). Many existing multilingual topic models connect the languages by assuming parallelism at either the sentence or document level (Zhao and Xing, 2006; Ni et al., 2009). Others assume comparable corpora (Mimno et al., 2009). Boyd-Graber and Blei (2009) build multilingual topic models for unaligned text, using a dictionary and crosslingual, word-level matching prior to align topics across languages. An alternative means of inducing word-level priors into a topic model is through the use of word vectors (Nguyen et al., 2015a; Das et al., 2015b). I leverage these insights to induce crosslingual parameter sharing using multilingual word vectors (Faruqui and Dyer, 2014). The model I describe in this thesis only requires bilingual dictionaries to create the multilingual word vectors, and does not require any other parallel text to align latent topics across languages. The dictionaries do not need to have complete coverage of the language – the word-pair alignment distribution is displayed in Table 3.1.

2.3 Multilingual Representation Learning

Multilingual word vectors (Faruqui and Dyer, 2014) are used to share model parameters across languages. The use of shared representations is effective for tasks like dependency

parsing and semantic role labeling (Ammar et al., 2016; Mulcaire et al., 2018). Additional related work uses variational autoencoders for inducing crosslingual word embeddings (Deng, 2017).

2.4 Semi-supervised Text Classification

Xu et al. (2017) use semi-supervised variational autoencoders (VAEs) for monolingual text classification. Other approaches (Dai and Le, 2015; Howard and Ruder, 2018; Radford et al., 2018) pre-train a language model and then fine-tune it to a target task with supervision. Peters et al. (2018) and McCann et al. (2017) leverage hidden representations from a pre-trained language or machine translation model when training a supervised model on a target classification task.

Chapter 3

DATA

This project focuses on the Reuters RCV2 corpus (Lewis et al., 2004; Klementiev et al., 2012), one of the very few polyglot datasets available for document classification. RCV2 covers over 487,000 newswire stories in thirteen languages. In this work, I study text classification in eight of those thirteen languages: English (en), German (de), French (fr), Spanish (es), Italian (it), Russian (ru), Japanese (ja) and Chinese (zh). These languages come from Indo-European, Japonic, and Sino-Tibetan language families, which enables the study of model generalization across typological disparities, with the caveat that these languages are only moderately morphologically complex (Bender, 2011). Following Klementiev et al. (2012), I use the top level categories for evaluation: GCAT (government social), ECAT (economics), MCAT (markets), and CCAT (corporate industrial).

I generate multiple data splits to analyze classifier performance under a different levels of annotation. The training data contains either 1000 or 200 documents, and the development data is 20% of the training data size. To ensure that all subsampled datasets maintain a uniform distribution of labels, I use a script from MLDoc (Schwenk and Li, 2018) to subsample according to balanced class priors.

3.1 Preprocessing

Prior to training, I perform pre-processing steps on each corpus that are aimed to be as normalized between languages as possible. In particular, I lowercase all characters, and perform language-specific tokenization using SpaCy[1]. I create the vocabulary using only the training data. As a proxy for removing stopwords, I remove all words that appear

[1]https://spacy.io

	DE	ES	FR	IT	JA	RU	ZH
Word pairs	101997	112580	113286	103612	25969	48714	21597
Word translation accuracy	68.4	78.2	64.4	69.4	35.0	50.5	37.4

Table 3.1: Alignment between English and target languages. First row: number of word-pairs in MUSE (Conneau et al., 2017b) English $\leftrightarrow$ target crosslingual dictionaries. Second row: Accuracy of word translation performed via k-nearest neighbors (k=5)

in greater than 99% of documents. Finally, I restrict the vocabulary to the 5000 most frequently occurring words in the training data. Experiments show that this vocabulary size achieves a good balance between classification accuracy and speed of training. Lightweight pre-processing assures that this framework can be tested on a variety of languages in an out-of-the-box manner. Lightweight pre-processing works for this use case since fine-grained syntactic or semantic details of each language matter less for document-level tasks such as topic classification (Griffiths et al., 2007). These pre-processing decisions may not be appropriate for other classification tasks.

3.2 Monolingual and Multilingual word vectors

For monolingual word vectors, I use fastText word embeddings (Joulin et al., 2016; Conneau et al., 2017b), which are trained on Wikipedia. I project monolingual fastText word embeddings into the same space using multiCCA (Ammar et al., 2016). For each non-English language, I use a small crosslingual dictionary and canonical correlation analysis (CCA) to find a transformation of the non-English vectors into the English vector space (Faruqui and Dyer, 2014). The dictionaries are provided by MUSE (Conneau et al., 2017b).

I evaluate the crosslingual word embeddings with a word translation task, following the same setup as Conneau et al. (2017b). I measure how many times one of the correct translations of a source word in a target dictionary is retrieved using a k-nearest neighbors algorithm, where $k = 5$. Table 3.1 shows that Germanic and Romantic languages have sub-

stantially higher word translation accuracy than Japanese, Russian, and Chinese, which are linguistically distant to English. Word embedding quality can be improved with crosslingual dictionaries that have a larger coverage of source-target word pairs.

Chapter 4

MONOLINGUAL DOCUMENT CLASSIFICATION

In this chapter, I describe a framework for monolingual document classification with SCHOLAR (Card et al., 2018). First, I describe generation and inference procedures in the monolingual setting, as well as methods for evaluation. Then, I perform experiments to identify strong continuous representations of documents, which I incorporate as additional prior information into the model. These results set the stage for extending the model to the multilingual context.

4.1 Latent Variable Models on Text

Understanding large collections of documents is non-trivial, since they contain a complex weave of themes, entities, and concepts. It is difficult to build good features to capture these document-level characteristics since they are not readily observed in the text. Latent variable modeling enables one to represent this rich data in terms of a simpler latent representation, which I will call z. One might use this latent representation as an explanatory tool for exploring the larger corpus, or one may leverage it in a downstream discriminative task.

In latent variable modeling, given some observed documents x, one estimates the joint probability $P(x, z)$. An effective latent variable model is the *topic model*, where z represents latent vectors of topics or themes across the corpus. Arguably the most popular topic model is latent Dirichlet allocation (LDA; Blei (2012)), which I briefly describe next.

Since $P(x, z) = \sum_z P(z)P(x|z)$, to estimate the joint probability, one can incorporate priors over the generative process. For example, in LDA, documents are treated as a bag of words from the vocabulary generated by one or more of K topics. A Dirichlet prior is imposed over the topic distribution, and each topic is in turn characterized by a multinomial

distribution over observed words. The Dirichlet prior turns out to be important for producing coherent topic assignments (Wallach et al., 2009).

4.2 SCHOLAR

One cannot use LDA for document classification, since it offers no way to incorporate labels into its generative process. This thesis instead leverages SCHOLAR (Card et al., 2018), a latent variable model that additionally uses document metadata and labels to guide latent representation learning.

Metadata in the document might be something like language ID, author, or publication date. In models that use crosslingual sharing (Ammar et al., 2016; Mulcaire et al., 2018), language ID is especially important. I explore the use of language ID as metadata in the generative story in Chapter 5.

In addition to being unable to incorporate information other than the documents themselves, any variation on LDA necessitates re-deriving a new inference algorithm, which can be time-consuming and require technical expertise (Srivastava and Sutton, 2017). SCHOLAR aims to be a flexible framework for document modeling that can incorporate different priors and architectural choices for topic modeling, while leveraging black-box algorithms for efficient posterior inference.

SCHOLAR specifically uses Auto-encoding Variational Bayes (AEVB) methods (Kingma and Welling, 2013) for approximate inference of the posterior distribution of topics across the corpus. AEVB replaces intractable inference algorithms with a neural inference network, which directly maps documents to a posterior distribution, and whose weights can be updated with stochastic gradient descent. With AEVB, the same model can be adapted to a variety of training procedures without needing to modify the inference algorithm.

SCHOLAR also allows users to leverage informative priors in the generative process. Here, I use monolingual word vectors to enforce the semantic coherence of words in a language. In the next chapter, I extend this to the multilingual context by using multilingual word vectors. Word vector priors allow us to leverage large external corpora, which is especially

useful under low-resource settings, when one may not have enough training data to establish word-level semantic relationships.

4.2.1 Generative Story

To perform document classification with SCHOLAR, one must estimate the joint distribution $P(w, y|c)$ in a generative manner, using document words w, labels y, and metadata c. Formally, this joint distribution is parameterized with the latent variable z as $P(w, y|c) = \sum_z P(y|z, c)P(w|z, c)P(z)$. To perform classification, one must set a prior on z, generate words and predict labels given z and any metadata.

$P(z)$ is set as an isotropic Gaussian, with parameters μ and σ^2. Both μ and σ are drawn from logistic normal priors (with Dirichlet parameter α) to aid in inference (Srivastava and Sutton, 2017):

$$u_0 = \log(\alpha) - \alpha$$
$$\sigma_0^2 = \frac{1}{\alpha}(1 - \frac{2}{K}) + \frac{\alpha}{K}$$

A multilayer neural network is used to estimate $P(w|z, c)$ (which I will call f_g, a generative network). f_g takes as input a sample of the latent variable, as well an embedding of metadata, if available. Each word is sampled over a categorical softmax of this generative network's output.

Finally, $P(y|z, c)$ is estimated with a separate multilayer network f_y.

Formally, the generative story is described below:

for document d in corpus D **do**
  Draw $z_d \sim \mathcal{N}(z|\mu_0(\alpha), diag(\sigma_0^2(\alpha)))$ # sample latent variable from Gaussian prior
  $\theta_d = \text{softmax}(z_d)$ # normalize
  $\rho_d = f_g(\theta_d, c_d)$ # feed latent variable, metadata, and labels into generative network
  for word i in document d **do**
    Draw $\mathbf{w}_{di} \sim p(w|\rho_d, \mu_{z_d}, \sigma_{z_d}^2)$ # sample words from output of generative network

7: **end for**

8: $y_d = f_y(\theta_d, c_d)$ # generate labels with another multilayer network

9: **end for**

where α is a fixed Dirichlet parameter.

4.2.2 Inference

Each document d is assumed to have a latent representation z_d. As is conventional in AEVB, for each document, one assumes a variational approximation to the posterior, $q_\phi(z_d|w_d, c_d)$, and minimizes the KL divergence between it and the true posterior, $p(z_d|w_d, c_d)$. To this end, one obtains the evidence lower bound (ELBO) for a single document:

$$\begin{aligned}\mathcal{L}(w_d) = \mathbb{E}_{q_\phi(z_d|w_d,c_d,y_d)}[\Sigma_i \log p(w_{di}|z_d, \mu_{z_d}, \sigma^2_{z_d})] \\ -D_{KL}[q_\phi(z_d|w_d, c_d)||p(z_d)]\end{aligned} \tag{4.1}$$

Intuitively, the first term can be thought of as a *reconstruction loss*, ensuring that generated words stay true to the original document distributions, and the second term attempts to reduce the entropy between the variational approximation to the posterior distribution of the latent space and the real posterior.

As it stands, the optimization of the ELBO is unsupervised. One can additionally use labels to optimize the latent representation of documents, by adding in a cross-entropy loss term to $\mathcal{L}$:

$$\begin{aligned}\mathcal{L}(w_d) = \mathbb{E}_{q_\phi(z_d|w_d)}[\Sigma_i \log p(w_{di}|z_d, \mu_{z_d}, \sigma^2_{z_d})] \\ -D_{KL}[q_\phi(z_d|w_d, c_d)||p(z_d)] \\ +\mathbb{E}_{q_\phi(z_d|w_d,c_d)}(\log p(y_d|z_d, c_d))\end{aligned} \tag{4.2}$$

As in the original formulation of the variational autoencoder, $q_\phi(z_d|w_d, c_d)$ is estimated with a multivariate normal distribution such that $q_\phi(z_d|w_d, c_d) = \mathcal{N}(\mu_d, \sigma^2_d)$. μ and σ^2 are

parameterized by shared multilayer neural networks f_μ and f_{σ^2}. Breaking these networks down,

$$\pi_d = f_e([W_x x_d; W_c c_d])$$
$$\mu_d = f_\mu(\pi_d)$$
$$\log(\sigma_d^2) = f_{\sigma^2}(\pi_d)$$

where x_d is a vector of word frequency counts, W_x, W_c are weights, and f_e, f_μ and f_{σ^2} are multilayer feedforward networks. In section 4.5, I augment the input to f_e with a word vector-based representation of the document, such that:

$$E_d = f_v(V_d)$$

$$\pi_d = f_e([W_x x_d; E_d; W_c c_d])$$

where V_d is a set of word vectors that make up the document d and f_v is an arbitrary neural network.

Finally, the reparameterization trick is used (Kingma and Welling, 2013) to sample z_d:

$$\epsilon \sim \mathcal{N}(0, \mathbf{I})$$
$$z_d = \mu_d + \sigma_d \cdot \epsilon$$

This sample of z_d is used in a Monte Carlo approximation of $\mathcal{L}(w_d)$.

4.3 Evaluating variational autoencoders

The use of both supervised (cross-entropy) and unsupervised (reconstruction) loss in the overall learning procedure allows for multiple modes of evaluation. One can evaluate the intrinsic quality of the latent representation z, as well as the accuracy of SCHOLAR on a downstream classification task, to measure how well the model learns document representations.

4.3.1 Unsupervised evaluations

The primary methods of unsupervised evaluation are *perplexity*, *coherence*, and *qualitative evaluation*. In document modeling, perplexity is computed by

$$\exp(-\frac{1}{D}\sum^{N_d}\frac{1}{N_d}\log p(w_d)) \tag{4.3}$$

where D is the number of documents, N_d represents the length of the dth document and $p(w_d)$ is the log probability of words in the document.

Coherence is measured in non-negative point-wise mutual information (NPMI; Chang et al. (2009)). I evaluate NPMI using the top 10 words assigned to each topic dimension.

Additionally, I qualitatively evaluate the semantic information learned by the model by rendering the weights of the decoder applied to z to reconstruct the text.

4.3.2 Supervised evaluation

To evaluate the learned latent representation in a supervised manner, I use it in downstream document classification tasks. In all experiments, I use raw accuracy on the test data as the evaluation method.

4.4 Architecture and Training Hyperparameters

In the following experiments, I set the latent dimension of z at 128. I use a 2 layer feed-forward network with 512 hidden dimensions for f_e, f_g, f_μ and f_{σ^2}. I use a batch size of 64, and Adam (Kingma and Ba (2014); with β=0.99 and learning rate of 0.001) for optimization. I train all models until development accuracy does not increase for 100 epochs.

4.5 Choosing f_v

Documents are commonly represented as a bag of words that make up the document, as is the case in the original implementation of SCHOLAR. In this section, I augment this bag

of words representation with word vectors, using the neural network f_v described in Section 4.2.

Given a document d and word vectors W of dimension k for all n_d words in the document, we experiment with a variety of architectures for f_v:

- **Average Word Vector**: f_v is a linear averaging of all word vectors in d. This approach prioritizes computational efficiency while leveraging the semantically meaningful traits of word vector averaging.

- **Hierarchical ConvNet**: Convolutional neural networks are used in high performing text classifiers (Zhao et al., 2015; Kim, 2014). AdaSent (Zhao et al., 2015) aggregates convolutional representations of sentences at different levels of abstraction. Following (Conneau et al., 2017a), I use a simplified (and faster) version of this hierarchical architecture with a series of convolutional neural networks.

- **BiLSTM with Maxpool**: Each document is represented by a biLSTM network with a maxpool. Conneau et al. (2017a) find that this architecture achieves state-of-the-art performance on downstream sentence classification tasks.

Because the rest of the model consists of feedforward networks, and the vocabulary size is restricted, the input document encoder is the primary speed bottleneck, and there exists a tradeoff between runtime and quality of document representation. By evaluating each of these encoders across eight languages in RCV2 (Table 4.1), I find that setting f_v as the BiLSTM with MaxPool performed on average 0.2 points better than the hierarchical ConvNet, and 1.3 points better than word vector averaging. I find that using a hierarchical ConvNet and word vector averaging produced the most coherent latent representations, and that word vector averaging resulted in significantly higher perplexities than the other architectures. Given these results, and taking into account the fact that the biLSTM with Maxpool architecture takes 3-5 times longer to train, I use the hierarchical ConvNet for f_v in all future experiments.

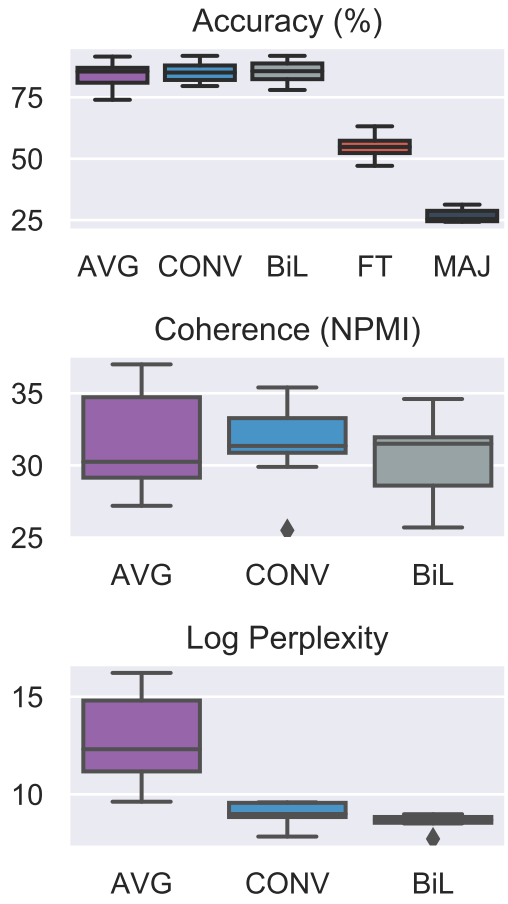

Figure 4.1: Summary effect of continuous document representations on downstream accuracy, perplexity, and coherence. Key: *AVG* - word vector average, *CONV* - hierarchical ConvNet, *BiL* - BiLSTM with maxpool, *FT* - FastText baseline, *MAJ* - Majority baseline

The modeling decisions made here are based on a smaller set of experiments, and may not necessarily generalize well to other tasks, datasets, or languages.

4.6 Qualitative evaluations

In Tables 4.2, 4.3, and 4.4, I present words collocating in the same dimensions of the latent space. The model learns interpretable topics from each monolingual corpus.

Accuracy (%)

f_v	EN	ES	DE	IT	FR	RU	JA	ZH	μ
Mean Word Vector	86.2	91.6	86.3	84.8	89.2	74.0	79.9	81.3	84.2
Hierarchical ConvNet	86.5	91.9	87.6	82.2	89.1	79.6	83.7	81.7	85.3
BiLSTM + MaxPool	87.2	91.9	88.4	82.7	90.0	78.0	81.5	84.1	85.5
FastText (Baseline)	54.6	47.1	49.6	55.0	58.3	57.0	63.2	53.2	54.8
Majority (Baseline)	24.8	31.1	24.6	24.4	24.3	28.0	26.1	31.3	26.8

Coherence (NPMI)

f_v	EN	ES	DE	IT	FR	RU	JA	ZH	μ
Mean Word Vector	27.2	35.4	37	30.7	29.8	34.5	27.2	29.8	31.5
Hierarchical ConvNet	25.5	35.4	31.2	31.5	34.4	29.9	32.9	31.2	31.5
BiLSTM + MaxPool	25.7	34.6	31.8	31.5	27.4	32.4	29.0	31.5	30.5

Log Perplexity

f_v	EN	ES	DE	IT	FR	RU	JA	ZH	μ
Mean Word Vector	5.5	4.3	5.2	4.2	7.0	5.0	6.9	6.3	5.6
Hierarchical ConvNet	4.2	3.8	4.2	4.2	3.9	3.9	3.4	3.9	3.9
BiLSTM + MaxPool	3.8	3.8	3.8	3.9	3.7	3.7	3.4	3.7	3.7

Table 4.1: Accuracy (top; higher is better), Coherence (middle; higher is better), and Log Perplexity (bottom; lower is better), on RCV2 document classification with different continuous document representations. FastText and Majority baselines shown for downstream accuracy comparisons. Refer to Figure 4.1 for summary.

English		Spanish	
0.37	0.35	0.63	0.13
explosion	port	ataque (*attack*)	unidos (*united*)
blasts	vessels	ayudó (*helped*)	tokio (*Tokyo*)
station	loading	detención (*detention*)	Nikkei (*Nikkei*)
exploded	views	intento (*attempt*)	estados (*states*)
spate	vessel	soldado (*soldier*)	económicos (*economic*)
wounding	delays	ocurre (*occur*)	ocde (*OECD*)
sitting	bulk	árabes (*Arab*)	junta (*board*)
bomb	orleans	permanecían (*stay*)	cerró (*closed*)
antonio	emerge	desafío (*challenge*)	accionistas (*shareholders*)
hizbollah	financing	presuntamente (*presumably*)	tenemos (*we have*)

German	
0.54	0.29
finanzdienst (*financial services*)	staaten (*states*)
ös (*ancestor*)	israel (*Israel*)
maculan (*Maculan*)	verstümmelt (*mutilated*)
wertpapieren (*securities*)	schutztruppe (*security forces*)
vorgemerkt (*noted*)	vereinten (*united*)
stamm (*regulars*)	uno (*U.N.*)
ausgesetzt (*exposed*)	gemeinsamen (*collective*)
österreich (*Austria*)	nationen (*nations*)
änderungen (*changes*)	begrüßte (*welcome*)
umsatzdaten (*sales data*)	palästinensischen (*palestinians*)

Table 4.2: Randomly picked topics derived from latent decoder weights of SCHOLAR trained independently in each language. Coherence measures displayed on top.

Italian		French	
0.46	0.47	0.50	0.29
missione (*mission*)	scalfaro (*Scalfaro*)	boris (*Boris*)	titre (*title*)
albania (*Albania*)	oscar (*Oscar*)	tchoubaïs (*Tchoubaïs*)	pathé (*Pathé*)
prodi (*warriors*)	luigi (*Luigi*)	anatoli (*Anatoli*)	roland (*Roland*)
protezione (*security*)	repubblica (*republic*)	eltsine (*Yeltsin*)	garros (*Garros*)
albanese (*albanese*)	quirinale (*Quirinal*)	compté (*account*)	bskyb (*BskyB*)
multinazionale (*multinational*)	ricevuto (*receive*)	descendre (*to descend*)	lvmh (*LVMH*)
balcano (*Balkan*)	bossi (*Bossi*)	russes (*Russian*)	audiovisuel (*audio-visual*)
bashkim (*unity - in Albanian*)	bloccato (*stuck*)	habitants (*inhabitants*)	chelem (*slam*)
apr (*apr*)	d'alema (*D'Alema*)	nordouest (*northwest*)	australie (*Australia*)
rimasto (*remain*)	capo (*chief*)	évolué (*evolved*)	vig (*VIG*)

Russian	
0.36	0.54
облигаций (*bonds*)	народное (*popular*)
ВСМ (*BCM*)	слово (*word*)
облигационного (*bonded*)	узбекской (*Uzbek*)
высокоскоростные (*high-speed*)	ташкентский (*Tashkentskiy*)
образования (*of education*)	востока (*east*)
смирнова (*Smirnov*)	ташкенте (*Tashkent*)
выплат (*payments*)	узбекистана (*of Uzbekistan*)
магистрали (*highways*)	правда (*true*)
дохода (*income*)	кыргызстана (*of Kyrgyzstan*)
анна (*Anna*)	киргизской (*Kyrgyz*)

Table 4.3: Randomly picked topics derived from latent decoder weights of SCHOLAR trained independently each language. Coherence measures displayed on top.

Japanese		Chinese	
0.65	0.35	0.68	0.57
翌日 (*next day*)	マルク (*Deutsche Mark*)	訊 (*news*)	公告 (*announcement*)
積み (*loading*)	下げ (*lowering*)	位于 (*located*)	股東 (*shareholder*)
インデックス (*index*)	帯 (*band*)	審 (*review*)	上午 (*morning*)
かい離 (*detachment*)	達夫 (*Tatsuo*)	利潤 (*profit*)	臨時 (*temporary*)
担保 (*collateral*)	付け (*bill*)	省 (*province*)	刊登 (*publish*)
無 (*nothing*)	インター (*Inter*)	部 (*unit*)	決議 (*resolution*)
通告 (*notice*)	スパン (*span*)	官 (*official*)	審議 (*review*)
調節 (*adjustment*)	とどまる (*stay*)	快 (*fast*)	提請 (*draw*)
日銀 (*Bank of Japan*)	みえ (*appearance*)	推薦 (*recommend*)	分配 (*distribution*)
件 (*item*)	パターン (*pattern*)	萬元 (*1000 yuan*)	方案 (*plan*)

Table 4.4: Randomly picked topics derived from latent decoder weights of SCHOLAR trained independently each language. Coherence measures displayed on top.

Chapter 5

CROSSLINGUAL JOINT TRAINING

In this chapter, I extend the monolingual SCHOLAR to a multilingual setting. I explore *crosslingual training*, a procedure in which I pool training data from multiple languages together and use multilingual word vectors as the input to f_v. I first describe an experiment that demonstrates the usefulness of augmenting crosslingual training data with document metadata (i.e., language ID). Then I explore the effect of crosslingual training under high- and low-resource settings.

5.1 Crosslingual learning methods

5.1.1 Simple sharing

Under *simple sharing*, I concatenate data from multiple languages and train a single model jointly across languages. I rely on pre-trained multilingual word vectors (Ammar et al., 2016) to represent distinct vocabularies belonging to each language in a shared space, enabling crosslingual learning (Klementiev et al., 2012).

5.1.2 Simple sharing + Language Identification

Previous work has shown that including language ID as input to crosslingual models improves performance on crosslingual tasks (Ammar et al., 2016; Mulcaire et al., 2018). Here, I include language ID as a covariate for each document. The metadata is represented as a one-hot embedding indicating which language the document came from.

5.2 Experimental settings

Mulcaire et al. (2018) shows that shared distributional semantic representations tends to benefit models applied to low-resource languages, and I test a similar hypothesis here. In these experiments, I evaluate our model under different annotation levels: when the target language is under a high-resource setting (1000 annotated documents) or a low-resource setting (200 annotated documents). For each target language $L_i \in \mathcal{L}$, the space of eight languages, I train a crosslingual model on a concatenation of data in the target language with 1000 annotated documents from the seven other languages of the corpus. I also compare each cross-lingual model to a monolingual baseline in which I independently train SCHOLAR, with no parameters shared, on the target language. I additionally train a baseline monolingual FastText (Joulin et al., 2016) model on each target language.

5.3 Results

5.3.1 Effect of Language ID

Language ID increases downstream performance over monolingual baselines under both high- and low-resource settings (Figures 5.1, 5.2, 5.3, 5.4). Under a low-resource setting, I observe that the inclusion of language ID increases performance of a crosslingual model 14.8 ± 9.4 points over a model that lacks this covariate (Figures 5.2, 5.4).

5.3.2 Effect of Annotation levels on Crosslingual Performance

Figure 5.1 shows results of crosslingual training under a high-resource setting. I observe that crosslingual training does not significantly improve performance relative to monolingual baselines under this setting. I observe that in some languages (Figure 5.3), there is slight performance decrease when using crosslingual training, suggesting that the inclusion of other languages in the training data may introduce noise that makes learning slightly harder.

On the other hand, I observe a 5.5 ± 6.5 point improvement over monolingual baselines when I perform crosslingual training under a low-resource setting (Figure 5.2). In Russian

0.10	0.40
生産 (*production*)	總統 (*president*)
rifondazione (*Refoundation*)	периодики (*periodicals*)
party	изложенной (*outlined*)
兌 (*exchange*)	ней (*her*)
convegno (*convention*)	覚悟 (*resolution*)
被害 (*damage*)	かりに (*instead*)
democratic	込み (*included*)
相殺 (*offset*)	表ざた (*expressed*)
魚 (*fish*)	示談 (*settlement*)
頭数 (*head count*)	使い (*use*)

Table 5.1: Randomly picked topics derived from latent decoder weights in crosslingual VAEs trained jointly on all languages and tuned to Spanish. Multilingual word vectors enable topic alignment across languages without the use of parallel data.

and Chinese (Figure 5.4), I observe that performance drops slightly compared to monolingual baselines.

5.4 Qualitative alignment of Topics across Languages

A result of crosslingual sharing is that some dimensions of the latent space assign words in multiple languages to the same topic (Table 5.1). While I do not observe this phenomenon in every dimension, the existence of these shared dimensions suggests that the model is able to align topics across languages without parallel data.

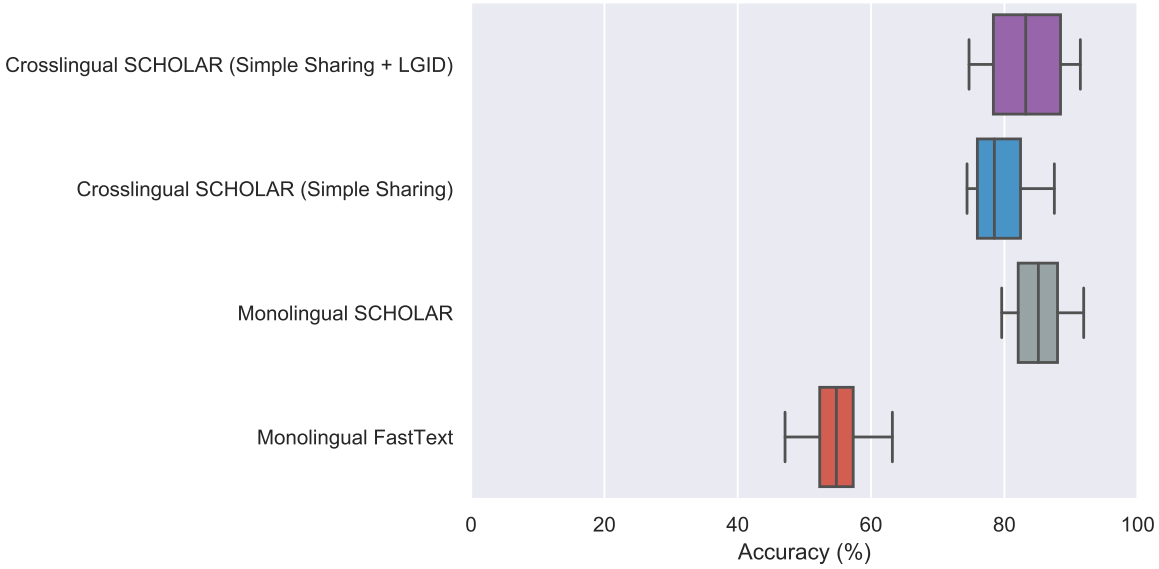

Figure 5.1: Summary effect of crosslingual training under a high-resource setting (1000 documents in target language and every other language).

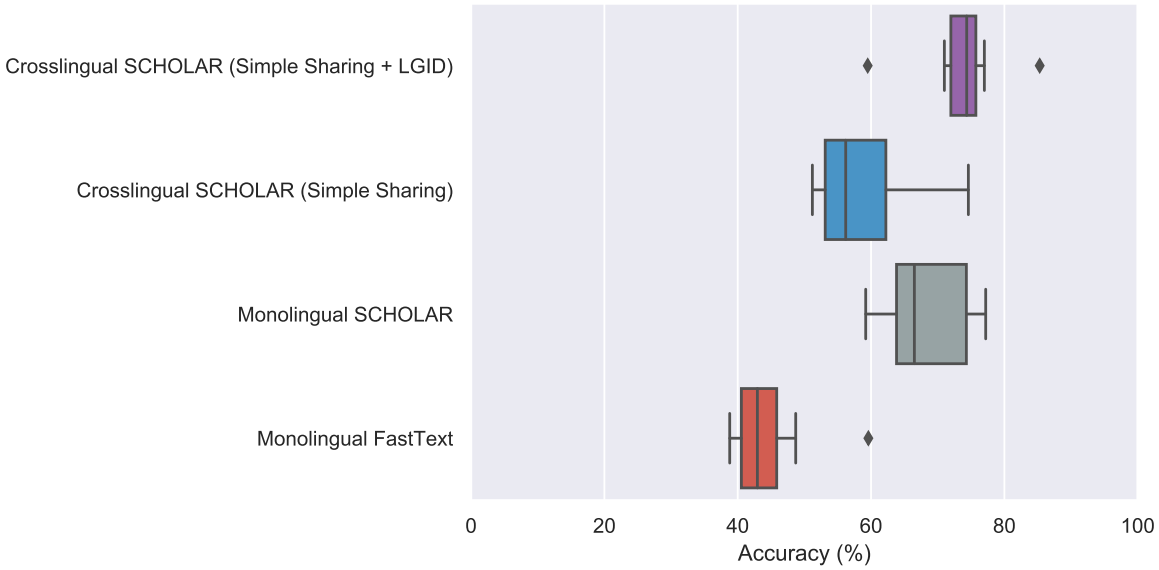

Figure 5.2: Summary effect of crosslingual training under a low-resource setting (200 documents in target language and 1000 documents in every other language).

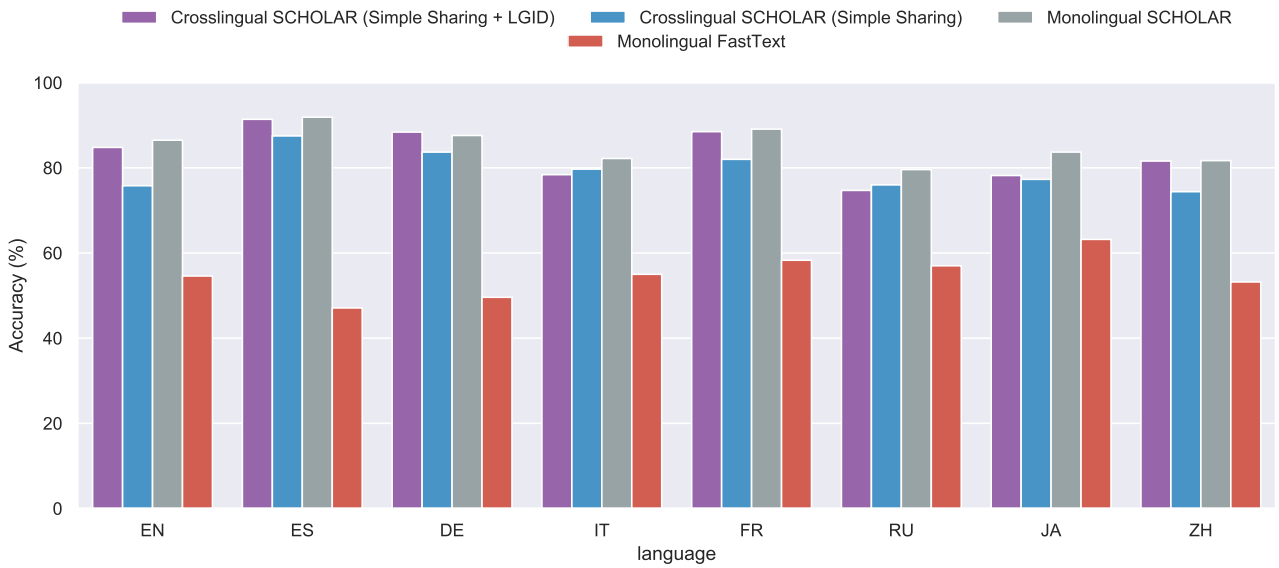

Figure 5.3: Detailed effect of crosslingual training under a high-resource setting (1000 documents in target language and every other language).

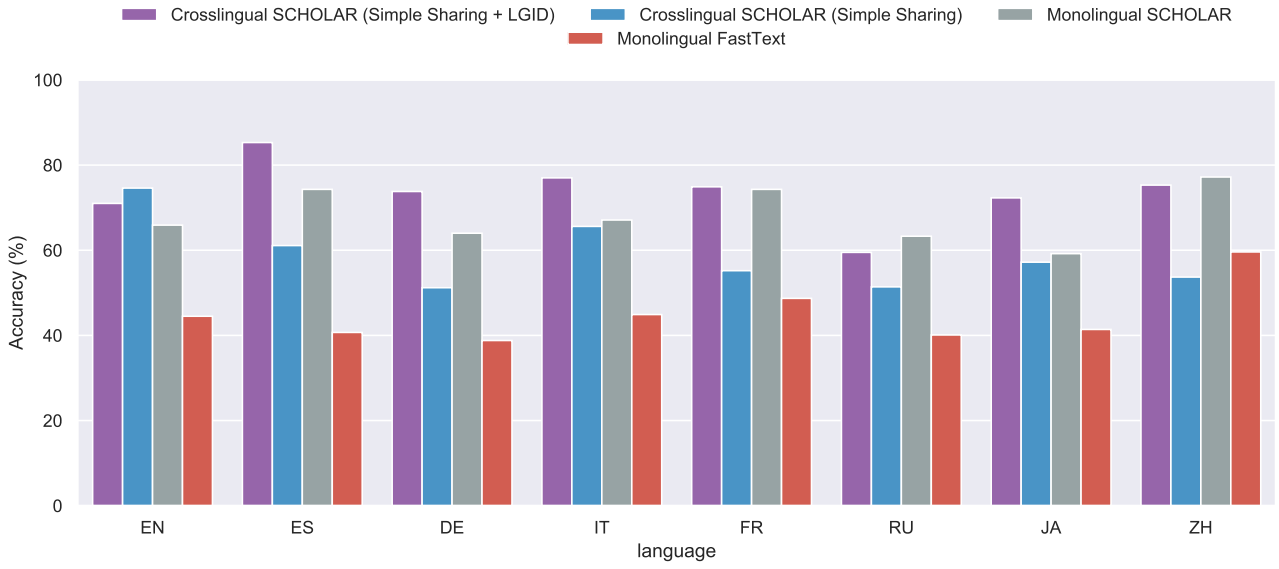

Figure 5.4: Detailed effect of crosslingual training under a low-resource setting (200 documents in target language and 1000 documents in every other language).

Chapter 6

EXPLOITING UNLABELED DATA

In this chapter, I use semi-supervised learning to leverage unlabeled data in SCHOLAR. First, I describe how to generalize the original SCHOLAR model to the semi-supervised case. Then, I describe experiments to analyze the performance of monolingual and crosslingual models under semi-supervised settings. I show that unlabeled data improves classification performance over supervised baselines.

6.1 Semi-supervised SCHOLAR

To generalize SCHOLAR to the semi-supervised setting, a binary gate δ_d is added to the classification loss in Equation 4.2:

$$\begin{aligned}\mathcal{L}(w_d) = \mathbb{E}_{q_\phi(z_d|w_d)}[\Sigma_i \log p(w_{di}|z_d, \mu_{z_d}, \sigma^2_{z_d})] \\ -D_{KL}[q_\phi(z_d|w_d, c_d)||p(z_d)] \\ +\delta_d \cdot \mathbb{E}_{q_\phi(z_d|w_d,c_d)}(\log p(y_d|z_d, c_d))\end{aligned} \tag{6.1}$$

δ_d is 1 if the document d is labeled, and δ_d is 0 if the document d is unlabeled. δ_d allows unlabeled data to influence the model's reconstruction loss only, and allows labeled data to influence both the reconstruction loss and classification loss. This method of incorporating unlabeled data is slightly different from what has been proposed for semi-supervised variational autoencoders (Kingma et al., 2014). Future work should compare these techniques.

6.2 Experiments

In line with the previous section, I explore the performance of the model under high- and low-resource settings. Under a high-resource setting, the target language has 1000 labeled

documents, and under a low-resource setting, the target language has 200 labeled documents. In these experiments, I additionally sample a disjoint set of 1000 documents in each language as unlabeled documents. I train on labeled and unlabeled data jointly, using the loss function described in the previous section. The use of unlabeled data from the same corpus minimizes out-of-domain effects that are detrimental to semi-supervised algorithms (Odena et al., 2018). Future work should experiment with using unlabeled data from other domains.

I experiment with two settings for semi-supervision: monolingual and crosslingual. For monolingual experiments, the target language's training data is concatenated with an additional 1000 unlabeled documents from the target language. The model is trained on each language separately, with no parameters shared. For crosslingual experiments, the target language's training data is concatenated with 1000 unlabeled documents from each of the seven other languages in the corpus. The model is trained on this multilingual dataset jointly. For clarity, I introduce new labels for each experimental setting under this setup (Table 6.1).

6.3 Results

I observe that under a high-resource setting, semi-supervised SCHOLAR significantly outperforms a FastText baseline, and is on-par with its supervised counterparts (Figure 6.1, 6.3).

Under a low-resource setting, I observe that semi-supervision is the most effective strategy for document classification in six out of eight languages (Figure 6.2, 6.4). Monolingual semi-supervision tends to outperform crosslingual semi-supervision, likely due to out-of-domain effects. Monolingual semi-supervision is especially effective for French, Russian, Japanese, and Chinese.

Label	Target Docs (L)	Auxillary Docs (L)	Target Docs (UL)	Auxillary Docs (UL)
Cross-S2	200/1000	0	0	7000
Cross-S1	200/1000	7000	0	0
Mono-S2	200/1000	0	1000	0
Mono-S1	200/1000	0	0	0
Mono-FT-S1	200/1000	0	0	0

Table 6.1: Number and type of documents in training data under each experimental setting described in this chapter. We train with either 200 or 1000 labeled documents in a target language (Target Docs (L)), and incorporate either labeled documents in seven other languages (Auxillary Docs (L)), unlabeled documents in the target language (Target Docs (UL)), and or unlabeled documents in seven other languages (Auxillary Docs (UL)). Rows 1-4 are experimental settings with SCHOLAR. The last row is an experimental setting with a FastText baseline.

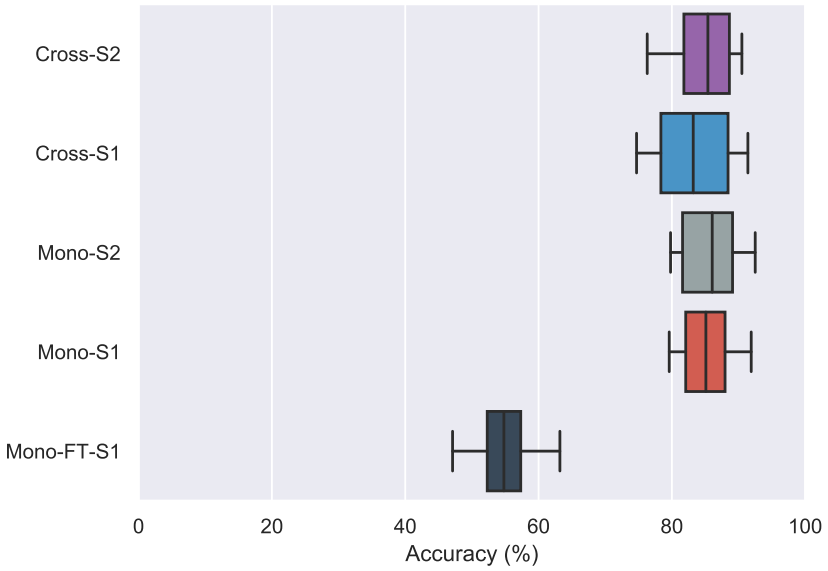

Figure 6.1: Summary effect of crosslingual training under a high-resource setting (1000 documents in target language and every other language). Refer to Table 6.1 for description of labels.

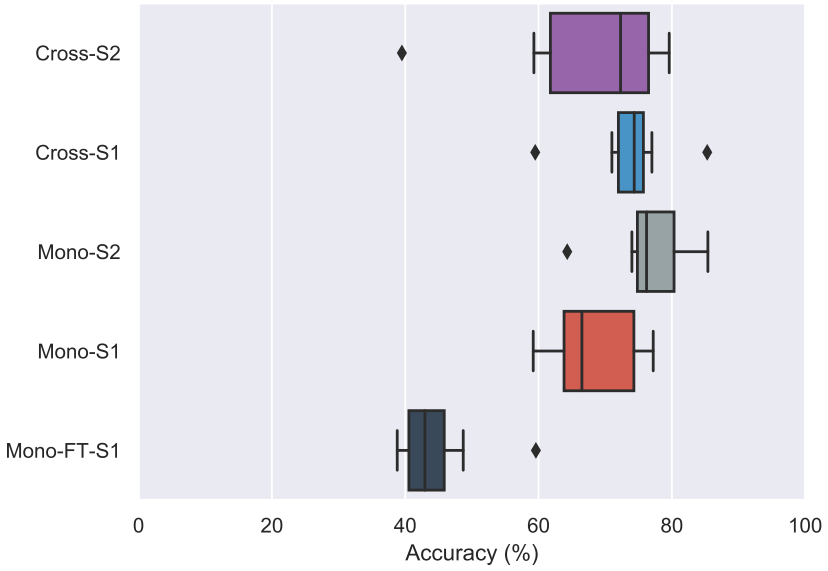

Figure 6.2: Summary effect of crosslingual training under minimal annotation (200 documents in target language and 1000 documents in every other language). Refer to Table 6.1 for description of labels.

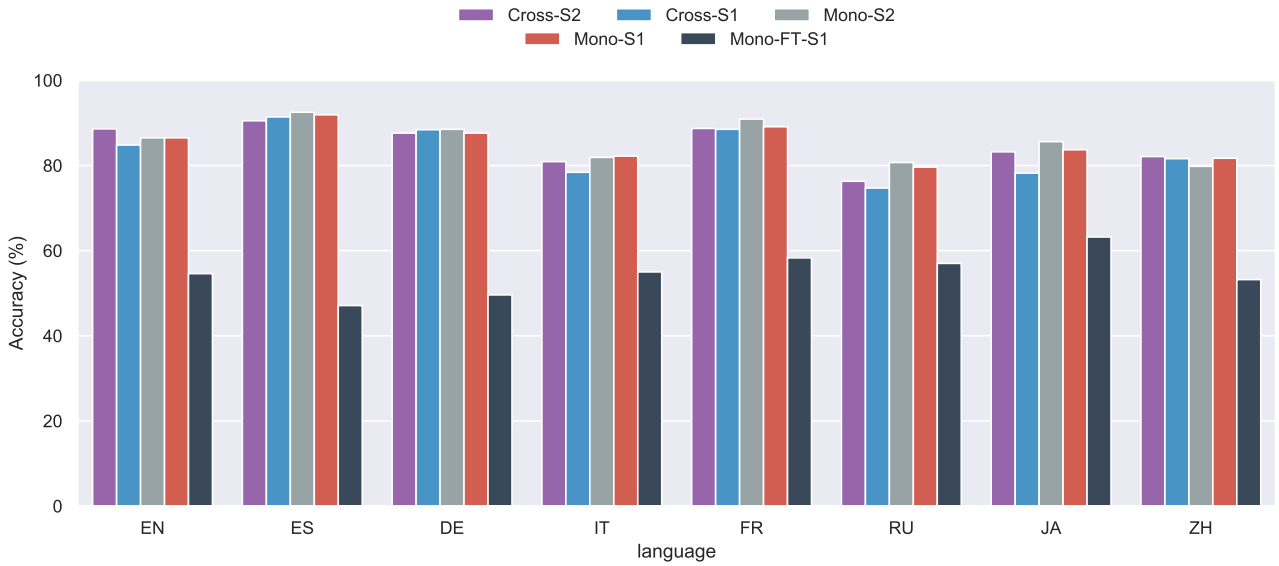

Figure 6.3: Detailed effect of crosslingual training under a high-resource setting (1000 documents in target language and every other language). Refer to Table 6.1 for description of labels.

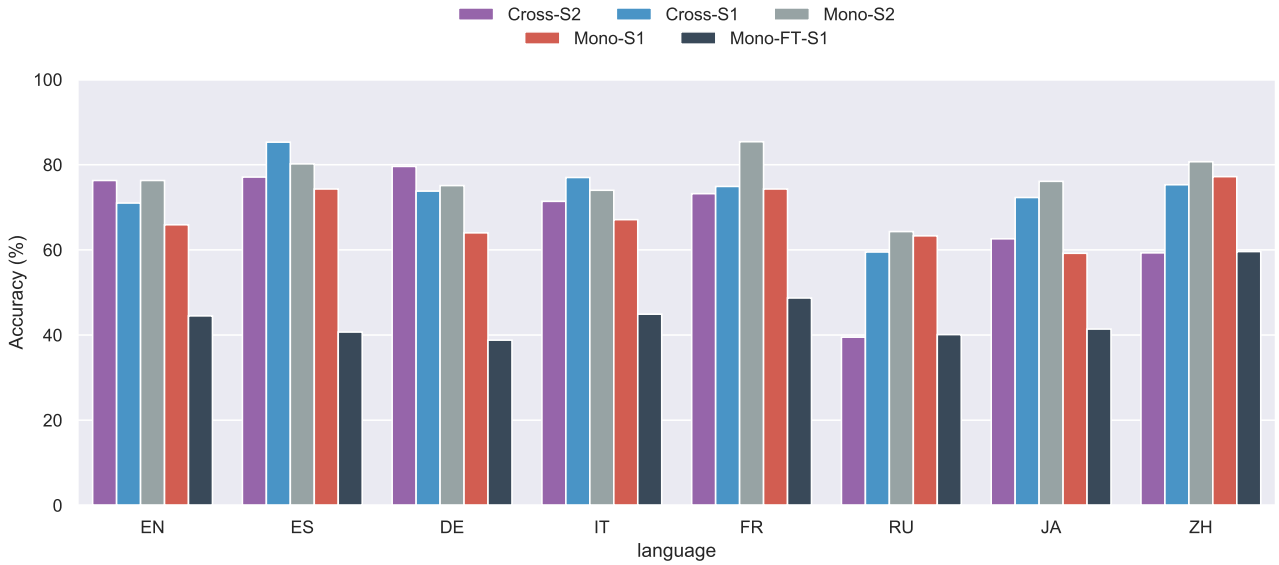

Figure 6.4: Detailed effect of crosslingual training under minimal annotation (200 documents in target language and 1000 documents in every other language). Refer to Table 6.1 for description of labels.

Chapter 7

DISCUSSION

In this thesis, I present a latent variable model for document classification. I show that the choice of input document representation can have significant effects on the coherence of word-topic assignments, perplexity, and downstream accuracy. I additionally show that using labeled data from different languages, and unlabeled data from the same language, are effective techniques for improving classification performance under low-resource settings.

7.1 Error Analysis

I generally observe that using unlabeled data from different languages gives a higher median accuracy than monolingual baselines, but larger variance in the results. Labeled and unlabeled data in other languages are actually detrimental to classification performance in Russian and Chinese. The poor quality of multilingual word vectors for these languages (Table 3.1) might explain these results, as the crosslingual semantic relationships might be noisy for these target languages in the shared space. A different pivot language for crosslingual projection or crosslingual dictionaries with larger coverage might improve the effectiveness of these techniques in languages distant from English.

It is possible that out-of-domain effects play a role in the high variance of results seen in crosslingual semi-supervision. It has been shown that incorporating unlabeled data from domains that are distant from the labeled data can be detrimental to downstream model performance (Odena et al., 2018). It would be especially pertinent to investigate the effects of the domain and size of unlabeled data during training.

The observation that semi-supervised learning becomes less useful under high-resource settings has been well documented (Odena et al., 2018). Crosslingual learning is also not as

useful under a high-resource setting, a phenomenon that has been observed in other tasks like semantic role labeling (Mulcaire et al., 2018).

7.2 Future Work

In this thesis, I explore semi-supervised learning and parameter sharing independently. Because crosslingual training and semi-supervision tend to help independently, it would be interesting to explore whether combining the techniques provides an additive benefit that improves classification performance even further.

To perform crosslingual training, the training data of all languages are pooled together. However, it is important to consider that languages group into families that delineate ancestral or typological similarities (Bender, 2011). Initial experiments suggest that sharing parameters based on language family is more effective than non-discriminant sharing. Future work might try to use language family specific encoders in the model, or include document metadata related to language family membership. The flexibility of the framework I describe in this thesis allows users to experiment with additional priors or features of documents without modifying the underlying inference procedure.

While I experiment with different encoders for this model, it would be interesting to explore alternative decoders as well. This thesis uses a categorical decoder, which ends in a softmax layer that maps to a probability distribution over a pre-defined vocabulary. For efficiency purposes, the vocabulary size must be restricted. Related work has proposed alternatives to the softmax that map to a probability distribution over subwords (Sennrich et al., 2015) or a hypersphere of word vectors (Kumar and Tsvetkov, 2018). These decoders would allow the model to have an open vocabulary, which may improve crosslingual sharing.

This thesis aligns with active areas of research in using language models on large unlabeled corpora to improve downstream performance on a variety of classification tasks (Howard and Ruder, 2018; Radford et al., 2018; Peters et al., 2018). One may think of topic models as a form of language modeling, in which words are assigned probabilities based on ambient topic distributions. It would be interesting to bridge fine-grained language modeling and the

more coarse-grained topic modeling, as these connections may prove useful for building text classifiers that can perform strongly in multiple languages and domains.

BIBLIOGRAPHY

Oliver Adams, Adam Makarucha, Graham Neubig, Steven Bird, and Trevor Cohn. 2017. Cross-lingual word embeddings for low-resource language modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. volume 1, pages 937–947.

Željko Agić, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, and Anders Søgaard. 2016. Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics* 4:301–312.

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

Emily M Bender. 2011. On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology* 6(3):1–26.

David M Blei. 2012. Probabilistic topic models. *Commun. ACM* 55(4):77–84.

David M Blei and John D Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*. ACM, pages 113–120.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL* pages 31–40.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. 2015. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349* .

Jordan Boyd-Graber and David Blei. 2012. Multilingual topic models for unaligned text .

Jordan Boyd-Graber and David M Blei. 2009. Multilingual topic models for unaligned text. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. AUAI Press, pages 75–82.

Jordan Boyd-Graber, Yuening Hu, David Mimno, et al. 2017. Applications of topic models. *Foundations and Trends® in Information Retrieval* 11(2-3):143–296.

Dallas Card, Chenhao Tan, and Noah A Smith. 2018. Neural models for documents with metadata. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. volume 1, pages 2031–2040.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-graber, and David M Blei. 2009. Reading tea leaves: How humans interpret topic models. In Y Bengio, D Schuurmans, J D Lafferty, C K I Williams, and A Culotta, editors, *Advances in Neural Information Processing Systems 22*, Curran Associates, Inc., pages 288–296.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017a. Supervised learning of universal sentence representations from natural language inference data .

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017b. Word translation without parallel data. *arXiv preprint arXiv:1710.04087* .

Andrew M Dai and Quoc V Le. 2015. Semi-supervised sequence learning .

R Das, M Zaheer, and C Dyer. 2015a. Gaussian LDA for topic models with word embeddings. *ACL (1)* .

Rajarshi Das, Manzil Zaheer, and Chris Dyer. 2015b. Gaussian lda for topic models with word embeddings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. volume 1, pages 795–804.

Wim De Smet and Marie-Francine Moens. 2009. Cross-language linking of news stories on the web using interlingual topic modelling. In *Proceedings of the 2nd ACM workshop on Social web search and mining*. ACM, pages 57–64.

Liangchen Wei Zhi-Hong Deng. 2017. A variational autoencoding approach for inducing cross-lingual word embeddings .

Chris Dyer, Jonathan Weese, Hendra Setiawan, Adam Lopez, Ferhan Ture, Vladimir Eidelman, Juri Ganitkevitch, Phil Blunsom, and Philip Resnik. 2010. Cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proceedings of the ACL 2010 System Demonstrations*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACLDemos '10, pages 7–12.

Jacob Eisenstein, Amr Ahmed, and Eric P Xing. 2011. Sparse additive generative models of text .

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of EACL*.

Ryan Alden Georgi. 2016. *From Aari to Zulu: Massively Multilingual Creation of Language Tools using Interlinear Glossed Text*. Ph.D. thesis.

Matthew R Gormley, Margaret Mitchell, Benjamin Van Durme, and Mark Dredze. 2014. Low-resource semantic role labeling. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. volume 1, pages 1177–1187.

David Graff and C Cieri. 2003. English gigaword corpus. *Linguistic Data Consortium* .

Thomas L Griffiths, Mark Steyvers, and Joshua B Tenenbaum. 2007. Topics in semantic representation. *Psychological review* 114(2):211.

M D Hoffman, D M Blei, C Wang, and J Paisley. 2013. Stochastic variational inference. *The Journal of Machine* .

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification .

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651* .

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* .

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. 2014. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*. pages 3581–3589.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* .

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *COLING*.

Sachin Kumar and Yulia Tsvetkov. 2018. Machine translation with continuous outputs. In *In ICML 2018 workshop on Theoretical Foundations and Applications of Deep Generative Models*.

Kathy Lee, Diana Palsetia, Ramanathan Narayanan, Md Mostofa Ali Patwary, Ankit Agrawal, and Alok Choudhary. 2011. Twitter trending topic classification. In *Data Min-*

ing Workshops (ICDMW), 2011 IEEE 11th International Conference on. IEEE, pages 251–258.

David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research* 5(Apr):361–397.

Jon D Mcauliffe and David M Blei. 2008. Supervised topic models. In *Advances in neural information processing systems*. pages 121–128.

David Mimno Hanna M Wallach Jason Naradowsky David A Smith Andrew McCallum. 2009. Polylingual topic models. In *EMNLP*.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors .

Yishu Miao, Lei Yu, and Phil Blunsom. 2015. Neural variational inference for text processing .

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality .

David Mimno, Hanna M Wallach, Jason Naradowsky, David A Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*. Association for Computational Linguistics, pages 880–889.

Phoebe Mulcaire, Swabha Swayamdipta, and Noah A. Smith. 2018. Polyglot semantic role labeling. *CoRR* abs/1805.11598.

Dat Quoc Nguyen, Richard Billingsley, Lan Du, and Mark Johnson. 2015a. Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics* 3:299–313.

Dat Quoc Nguyen, Richard Billingsley, Lan Du, and Mark Johnson. 2015b. Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics* 3(0):299–313.

Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. 2009. Mining multilingual topics from wikipedia. In *Proceedings of the 18th international conference on World wide web*. ACM, pages 1155–1156.

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

Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. 2000. Text classification from labeled and unlabeled documents using em. *Machine learning* 39(2-3):103–134.

Alexandros Ntoulas, Marc Najork, Mark Manasse, and Dennis Fetterly. 2006. Detecting spam web pages through content analysis. In *Proceedings of the 15th international conference on World Wide Web*. ACM, pages 83–92.

Augustus Odena, Avital Oliver, Colin Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. 2018. Realistic evaluation of semi-supervised learning algorithms .

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval* 2(1–2):1–135.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations .

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative Pre-Training .

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2017. A survey of cross-lingual word embedding models .

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Paris, France.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909* .

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*. pages 1631–1642.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models .

Sarah Grey Thomason and Terrence Kaufman. 2001. *Language contact*. Citeseer.

Yulia Tsvetkov and Chris Dyer. 2016. Cross-lingual bridges with models of lexical borrowing. *Journal of Artificial Intelligence Research* 55:63–93.

Ivan Vulić, Wim De Smet, and Marie-Francine Moens. 2011. Identifying word translations from comparable corpora using latent topic models. In *Proceedings of the 49th Annual*

Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2. Association for Computational Linguistics, pages 479–484.

Ivan Vulić, Wim De Smet, Jie Tang, and Marie-Francine Moens. 2015. Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications. *Information Processing & Management* 51(1):111–147.

Ivan Vulić and Marie-Francine Moens. 2012. Detecting highly confident word translations from comparable corpora without any prior knowledge. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 449–459.

Hanna M Wallach, David M Mimno, and Andrew McCallum. 2009. Rethinking lda: Why priors matter. In *Advances in neural information processing systems*. pages 1973–1981.

Weidi Xu, Haoze Sun, Chao Deng, and Ying Tan. 2017. Variational autoencoder for semi-supervised text classification. In *AAAI*. pages 3358–3364.

Zichao Yang, Zhiting Hu, Ruslan Salakhutdinov, and Taylor Berg-Kirkpatrick. 2017. Improved variational autoencoders for text modeling using dilated convolutions. *arXiv preprint arXiv:1702.08139* .

Bing Zhao and Eric P Xing. 2006. Bitam: Bilingual topic admixture models for word alignment. In *Proceedings of the COLING/ACL on Main conference poster sessions*. Association for Computational Linguistics, pages 969–976.

Han Zhao, Zhengdong Lu, and Pascal Poupart. 2015. Self-adaptive hierarchical sentence model. In *IJCAI*. pages 4069–4076.