Expert Systems With Applications 190 (2022) 116194

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Cross-lingual embeddings with auxiliary topic models

Dong Zhou [a,*], Xiaoya Peng [a], Lin Li [b], Jun-mei Han [c]

[a] School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan, Hunan 411201, China
[b] School of Computer Science and Technology, Wuhan University of Technology, Wuhan, Hubei 430070, China
[c] National Key Laboratory for Complex Systems Simulation, Department of Systems General Design, Institute of Systems Engineering, Beijing 100101, China

ARTICLE INFO

Keywords:
Cross-lingual embeddings
Topical models
Word embedding models
Projection-based methods
Seed dictionaries

ABSTRACT

Projection-based methods for generating high-quality Cross-Lingual Embeddings (CLEs) have shown state-of-the-art performance in many multilingual applications. Supervised methods that rely on character-level information or unsupervised methods that need only monolingual information are both popular and have their pros and cons. However, there are still problems in terms of the quality of monolingual word embedding spaces and the generation of the seed dictionaries. In this work, we aim to generate effective CLEs with auxiliary Topic Models. We utilize both monolingual and bilingual topic models in the procedure of generating monolingual embedding spaces and seed dictionaries for projection. We present a comprehensive evaluation of our proposed model through the means of bilingual lexicon extraction, cross-lingual semantic word similarity and cross-lingual document classification tasks. We show that our proposed model outperforms existing supervised and unsupervised CLE models built on basic monolingual embedding spaces and seed dictionaries. It also exceeds CLE models generated from representative monolingual topical word embeddings.

1. Introduction

In recent years, reasoning about and applying Natural Language Processing (NLP) techniques to real-world scenarios are receiving increasing interest (Choudhury & Deshpande, 2021; Dabre, Chu, & Kunchukuttan, 2020; Zhou, Zhao, Wu, Lawless, & Liu, 2018). The need to represent meaning and transfer knowledge in multilingual applications has given rise to various fundamental cross-lingual methods and applications (Ruder, Vulić, & Søgaard, 2019). Based on the great success of monolingual word embedding models (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013), an exciting development in NLP is to learn Cross-Lingual Embeddings (CLEs) (Ruder, et al., 2019). Earlier work induces CLEs by leveraging bilingual supervision from multilingual corpora aligned at different levels (Dinu & Baroni, 2015; Mikolov, Le, & Sutskever, 2013; Mogadala & Rettinger, 2016). A recent trend is so-called projection-based CLE models (Glavaš, Litschko, Ruder, & Vulić, 2019). This group of approaches is to project two sets of monolingual vectors in different languages to a same shared cross-lingual vector space. In this way, cross-lingual semantics can be enabled, and words with similar meanings can be accurately captured in different languages.

Cross-Lingual Embeddings are appealing due to two reasons. One is that they enable us to compare the meaning of words across languages, which is key to bilingual lexicon extraction (Søgaard, Ruder, & Vulić, 2018), machine translation (Dabre, et al., 2020), or cross-lingual information retrieval (Vulić & Moens, 2015). Another is that CLEs facilitate knowledge transfer between languages, most notably between resource-rich and resource-lean languages (Adams, Makarucha, Neubig, Bird, & Cohn, 2017).

The usefulness of CLEs has been proved in many cross-lingual NLP tasks (Guo, Che, Yarowsky, Wang, & Liu, 2016; Heyman, Vulić, & Moens, 2017; Mikolov, Le, et al., 2013; Mogadala & Rettinger, 2016). From early models by exploiting heavy bilingual supervision in the form of parallel corpora aligned at word, sentence or document levels (Dinu & Baroni, 2015; Mikolov, Le, et al., 2013; Mogadala & Rettinger, 2016), to more recent models that require much less bilingual signals(Artetxe, Labaka, & Agirre, 2017; Smith, Turban, Hamblin, & Hammerla, 2017) or even no supervisions needed at all (Artetxe, Labaka, & Agirre, 2018; Conneau, Lample, Ranzato, Denoyer, & Jegou, 2018). Research community has witnessed the increased interests in inventing more effective CLE models.

However, most of the existing CLE models, if not all of them, are mainly focused on how to develop adequate supervised or unsupervised projection or alignment algorithms. In this paper, we make two assumptions that are equally important for the CLEs generation. One is

* Corresponding author.
E-mail addresses: dongzhou1979@hotmail.com (D. Zhou), 302834020@qq.com (X. Peng), cathylilin@whut.edu.cn (L. Li), hjm_han@163.com (J.-m. Han).

https://doi.org/10.1016/j.eswa.2021.116194
Received 15 May 2020; Received in revised form 22 July 2021; Accepted 2 November 2021
Available online 14 November 2021

that if high-quality monolingual word embeddings cannot be obtained in the first place, regardless of the complex projection process, the mapping results still tend to be unsatisfactory (Vulić, Ruder, & Søgaard, 2020). Another is that a good seed dictionary (or bilingual lexicon) will result in good CLEs, which is already verified by many previous researchers (Glavaš, et al., 2019; Søgaard, et al., 2018).

Topic models such as probabilistic Latent Semantic Analysis (Hofmann, 1999) and Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003) have been extensively studied for a great many years. They have been proved to be another compelling way to capture semantic information hidden in texts. Topic models and word embedding models are actually considered to be interrelated (Liu, Liu, Chua, & Sun, 2015; Zhou, Wu, Zhao, Lawless, & Liu, 2017). Learning in topic models can be considered a discrete process while learning in word embedding models is mainly regarded as a continuous procedure for producing semantic representations. For example, the LDA model conceives a document as a mixture of a small number of discrete topics and topics as a (relatively sparse) distribution over words. On the contrary, in models like word2vec (Mikolov, Sutskever, et al., 2013), a word is used to input a log-linear classifier with a continuous projection layer. In this way, word embeddings are typically generated by considering context words around center words or vice versa. At the same time, topic models usually take a global approach to consider all words equally in a document.

Previous attempts have been made to combine topic models and word embedding models in order to produce more accurate word meaning representations. Word embedding models can be utilized to enhance topic models or vice versa (Li, Wang, Zhang, Sun, & Ma, 2016). Despite the attractive theoretical foundation, most previous attempts to use topic models for monolingual and cross-lingual semantic relatedness measuring are less successful (Vulić, De Smet, & Moens, 2013). In contrast, using word embeddings is far more reliable. However, most of the approaches train the two models separately or use topics as pseudo words as in (Liu, et al., 2015), which ignores some critical information during the training process (Fu, Wang, Li, Yu, & Liu, 2016; Nguyen, Billingsley, Du, & Johnson, 2015).

The effectiveness of generating CLEs with topic models in cross-lingual settings is not yet exploited. This motivates us to create an efficient Cross-lingual embedding model with Auxiliary monolingual and bilingual Topic Models (**C_ATM**). We try to use topic models in two ways. One is to generate good CLEs by inducing high quality monolingual word embedding spaces through iteratively refining topic models and word embedding models. Another is that we try to create a better bilingual seed dictionary used for both supervised and unsupervised methods to generate CLEs. We introduce a novel Refined Topic Model (**RTM**) together with a Revised Topical Skip-gram (**RTS**) model to iteratively generate high-quality monolingual word embeddings. We then employ projection-based methods (Ruder, et al., 2019) to map embedding spaces in two different languages to a same shared cross-lingual vector space. We use an augmented seed dictionary based on bilingual semantic similarities produced by bilingual topic models.

We perform two sets of extrinsic evaluations. We focus on evaluating the proposed C_ATM model in bilingual lexicon extraction, cross-lingual semantic word similarity as well as cross-lingual document classification tasks in the first set. Results suggest that our model outperforms state-of-the-art supervised and unsupervised CLE models using basic vector-based semantics only. We focus on evaluating our model in the second set of experiments by replacing the basic monolingual embeddings with the representative monolingual topical word embeddings (TWEs). In this evaluation, our proposed model again performs better, leading to a good solution in all cases tested. Finally, we compare our augmented seed dictionary with the commonly used dictionary for generating CLEs. The results prove that using bilingual topic models for enriching the seed dictionary can further improve the performance.

Our contribution in this paper can be summarized as follows:

i. We propose a novel way to learn effective cross-lingual embeddings by inducing high quality monolingual word embedding spaces through iteratively refining topic models and word embedding models. To the best of our knowledge, our method is the first to allow the construction of a CLE model by jointly considering topic models and word embedding models.
ii. We use bilingual topic models to capture semantic similarities between words in two languages. We further use this information to enrich a bilingual seed lexicon, with or without supervision, for use in the process of generating good CLEs.
iii. We find that unsupervised models do not match the performance of their supervised competitors for the NLP tasks evaluated in this paper despite recent claims.

2. Related work

We firstly review recent advances on projection-based cross-lingual embedding models. Then we focus on the word meaning representation models, which combining topic models and word embedding models.

Cross-lingual embedding models (Ruder, et al., 2019) have displayed broader applicability and versatility for multilingual applications than other paradigms such as machine translation (Dolamic & Savoy, 2010; Karimzadehgan & Zhai, 2010) or multilingual knowledge sources (Ni, Sun, Hu, & Chen, 2009). Earlier models induced CLEs by exploiting bilingual supervision in the form of bilingual corpora, aligned at the level of words, sentences, or documents (Dinu & Baroni, 2015; Mikolov, Le, et al., 2013; Mogadala & Rettinger, 2016). Recently, the focus has been put on aligning embedding spaces instead of just aligning vocabularies. Projection-based CLE models (Artetxe, et al., 2018; Conneau, et al., 2018) have gained great attention. This type of methods learns a projection or mapping between two pre-trained monolingual embeddings (Mikolov, Sutskever, et al., 2013; Zou, Socher, Cer, & Manning, 2013). Only limited supervision (usually only requires dictionaries containing few hundreds or thousands word translation pairs) or no supervision is needed for training CLEs (Artetxe, et al., 2018; Søgaard, et al., 2018). Faruqui and Dyer (2014) applied canonical correlation analysis to learn the data representations that maximize the correlations between two embedding spaces. Xing, Wang, Liu, and Lin (2015) discovered the important orthogonality constraint. Then the problem of CLEs induction boils down to solve the Procrustes problem. Smith, et al. (2017) leveraged identical character strings in different languages to create a training dictionary with limited supervision, on which they solve the Procrustes algorithm. Similar to their approach, Artetxe, et al. (2017) used a very small seed dictionary and an iterative algorithm to generate CLEs. However, their method is not suitable for languages that do not share a common alphabet. Joulin, Bojanowski, Mikolov, Jégou and Grave (2018) learned the projection by maximizing the ranking-based measures called Cross-domain Similarity Local Scaling (CSLS).

The supervised CLE models still need bilingual supervision signals (often a seed dictionary or bilingual lexicon) to induce the cross-lingual semantic spaces. However, there are models that do not require any bilingual signals (Artetxe, et al., 2018; Bai, Cao, Chen, & Zhao, 2019; Conneau, et al., 2018; Hoshen & Wolf, 2018). These methods induce seed translations automatically and learn vector space transformations to align different semantic spaces by considering inherent structural similarities of monolingual word embedding spaces. Artetxe, et al. (2018) used a self-learning bootstrapping procedure to expand the initially aligned word pairs. Conneau, et al. (2018) utilized an adversarial learning-based model to learn a projection until a discriminator cannot distinguish whether a vector comes from source or target embedding spaces. Hoshen and Wolf (2018) designed an iterative closest point algorithm to learn a shared embedding space by solving the Procrustes problem. Bai, et al. (2019) trained two autoencoders jointly to transform the source and the target monolingual word embeddings into a shared embedding space. However, as point out by Søgaard, et al. (2018), unsupervised models strongly rely on isomorphism of

monolingual embedding spaces, often leading to poor performance in particular for distant languages.

None of the above-mentioned CLE models considers using topic models to enhance word embeddings. Instead, they primarily focus on how to develop a mapping strategy. However, if high-quality word embeddings cannot be obtained in the first place, the complex projection process will not ensure good mapping results can be acquired.

Integrating topic models and word embedding models to produce better monolingual embeddings has attracted a significant amount of attention since the appearance of word embedding models. Several approaches have been proposed to improve word embedding models and topic models or both. As a pioneer work, Liu, et al. (2015) employed latent topic models while learning word embeddings by considering words and topics together. They learn word and topic embeddings by regarding each topic as a pseudo word. Their model uses the topic of the target word to predict context words in the same way as using a word. The basic idea behind this approach is that they regard each topic as a pseudo word that appears in all positions of words assigned to the topic. Then they trained topic embeddings and word embeddings separately and concatenated them together to get a final output. Using one model to enhance another as a pipeline is not the only option. Li, et al. (2016) proposed a model to enrich topic models for short texts by using semantically related words promoted by auxiliary word embeddings. Nguyen, et al. (2015) modeled topics as mixtures of the multinomial and a word embedding component.

There are also studies to learn word embeddings and latent topics jointly (Shi, Lam, Jameel, Schockaert, & Lai, 2017). They used a generation function to predict surrounding words given a target word and its topic. An EM-negative sampling method is employed to obtain the embedding vectors. Their model is an analogy to Bi-gram topical model (BTM) (Cheng, Yan, Lan, & Guo, 2014), in which word embeddings replace one group of the words. In fact, their model can still be regarded as utilizing word embedding models to enhance topic models. Fu, et al. (2016) proposed a word-topic mixture model to learn word embeddings by training topic models and then learning topic models by the revised embeddings.

Our approach is different from the methods described above. Shi, et al. (2017) 's work updates word embeddings in each iteration of training process. On the contrary, we develop a joint loss to train the topic models and word embedding models in a more aggregate way. Our model avoids inherent noises that may be introduced by the naïve word2vec models such as skip-gram. Our work is also different from topical word embeddings (Liu, et al., 2015) and Fu, et al. (2016)'s work, in which they use pivot topics to predict context words and train different embeddings separately. We also use topic models to enrich the seed dictionary for the projection process, which has not been explored before.

As a final overview, we list all representative projection-based CLE models, monolingual topical embedding models with their main approaches and key components in Table 1. The table is meant to reveal the high-level overview of different models. The minor differences and implementation details can be found in the original papers.

Algorithm 1. Generating an augmented seed dictionary

1. $\mathscr{X}^s, \mathscr{X}^t \leftarrow$monolingual embeddings of source and target languages
2. $D \leftarrow$ initial real dictionary or dictionary generated by unsupervised methods
3. **for** $iteration = 1, 2, \cdots, N_{iteration}$ **do**
4. $\widetilde{\mathscr{X}^s}, \widetilde{\mathscr{X}^t} \leftarrow$lookups for D in $\mathscr{X}^s, \mathscr{X}^t$
5. $\mathscr{X}^{s'} \leftarrow \mathscr{X}^s - \widetilde{\mathscr{X}^s}, \mathscr{X}^{t'} \leftarrow \mathscr{X}^t - \widetilde{\mathscr{X}^t}$
6. compute $nn\left(x_i^{s'}, x_j^{t'}\right)$ for all words in $\mathscr{X}^{s'}, \mathscr{X}^{t'}$
7. $D \leftarrow D \cup nn\left(x_i^{s'}, x_j^{t'}\right)$
8. **end for**

3. CLEs learning

Our research aims to develop an effective Cross-lingual embedding model with Auxiliary Topic Models (**C_ATM**). As with previous studies, our model is built upon the projection-based framework. In this framework, we try to learn a projection between independently trained monolingual embedding spaces. We defer the monolingual embedding space generation process to the next section, which is another key component of our **C_ATM** model. We will introduce an iterative refinement procedure to jointly train the Refined Topic Model (**RTM**) with the Revised Topical Skip-gram (**RTS**) model to obtain high-quality monolingual word embeddings. This section focuses on the principal projection (or mapping) procedure for CLEs learning in our Cross-lingual embedding model with Auxiliary Topic Models (**C_ATM**). We introduce the use of bilingual topic models for augmenting a seed dictionary that is utilized for CLEs generation.

Table 1
A summary of key components of representative projection-based CLE models, monolingual topical embedding models and our proposed methods.

Models	Supervision required	Approaches	Topics	Enhanced mono Embs	CLEs	Topical CLEs
Representative CLE models						
Faruqui and Dyer (2014)	**Yes**	Canonical correlation analysis	×	×	√	×
Xing, et al. (2015)	**Yes**	Normalization, orthogonality	×	×	√	×
Smith, et al. (2017)	**Yes**	Orthogonality, inverted softmax identical character strings	×	×	√	×
Artetxe, et al. (2017)	**Yes**	Normalization, orthogonality, mean centering, bootstrapping	×	×	√	×
Joulin, et al. (2018)	**Yes**	Ranking-based optimization	×	×	√	×
Artetxe, et al. (2018)	**No**	Heuristic alignment	×	×	√	×
Conneau, et al. (2018)	**No**	Adversarial learning	×	×	√	×
Hoshen and Wolf (2018)	**No**	Iterative closest point algorithm	×	×	√	×
Bai, et al. (2019)	**No**	Bilingual adversarial autoencoder	×	×	√	×
Representative monolingual topical embedding models						
Liu, et al. (2015)	–	Learning topics and embeddings together by treating topics as pseudo words	√	√	×	×
Li, et al. (2016)	–	Learning topics by auxiliary word embeddings	√	√	×	×
Nguyen, et al. (2016)	–	Learning topics as mixtures of the multinomial and a word embedding component	√	√	×	×
Fu, et al. (2016)	–	Word-topic mixture model	√	√	×	×
Shi, et al. (2017)	–	A generation function with EM-negative sampling	√	√	×	×
Ours (Supervised version)	**Yes**	Refined topic models with Revised topical CLE model, orthogonality, inverted softmax identical character strings	√	√	√	√
Ours (Unsupervised version)	**No**	Refined topic models with Revised topical CLE model, adversarial learning	√	√	√	√

According to (Glavaš, et al., 2019), let the monolingual word embedding spaces be $\mathscr{X}^s$ and $\mathscr{X}^t$, all projection-based approaches encompass three steps:

i. Construct a seed translation dictionary or bilingual lexicon $D = \{w_i^s, w_j^t\}_{i,j=1}^{U}$ containing U word pairs. We use $\boldsymbol{x}_i^s$ and $\boldsymbol{x}_j^t$ to denote the vectors for w_i^s and w_j^t, respectively.
ii. Align monolingual embedding space $\mathscr{X}^s$ and $\mathscr{X}^t$ using D. This procedure retrieves vectors $\{\boldsymbol{x}_i^s\}_{i=1}^{U}$ and $\{\boldsymbol{x}_j^t\}_{j=1}^{U}$ to form $\widetilde{\mathscr{X}^s}$ and $\widetilde{\mathscr{X}^t}$.
iii. Learn to project $\widetilde{\mathscr{X}^s}$ and $\widetilde{\mathscr{X}^t}$ to a shared cross-lingual embedding space $\widetilde{\mathscr{X}^{st}}$. The way we use one single projection matrix $\widetilde{\mathscr{X}^{st}}$ instead of two matrices for each embedding space is a simplified strategy used by most of the previous work to learn a single-direction projection.

In step ii above, all methods contain a lookup procure to generate monolingual subspaces (i.e. $\widetilde{\mathscr{X}^s}$, $\widetilde{\mathscr{X}^t}$←lookups for D in $\mathscr{X}^s$, $\mathscr{X}^t$). However, we argue that the dictionary typically contains minimal translation pairs, which is the main burden of building high-quality cross-lingual embeddings. Similar to the self-learning procedures that supervised (Glavaš, et al., 2019) and unsupervised (Artetxe, et al., 2018) methods used for augmenting the initially lexicon D, in this paper, we propose a simple bootstrapping method using the bilingual topic models for generating an enriched dictionary D. The procedure is summarized in Algorithm 1. For supervised method, it is shown that a fix-size bilingual dictionary can reach the state-of-the-art performance (Artetxe, et al., 2018; Glavaš, et al., 2019). However, if this dictionary can be further improved, we believe that the performance can be further enhanced. For unsupervised method, the seed dictionary is created automatically by assuming approximate isomorphism between $\mathscr{X}^s$ and $\mathscr{X}^t$. As most unsupervised method just uses this seed dictionary for final embedding generation, we use the extended bilingual dictionary to increase the performance further. For those words that do not capture by the seed dictionary, we calculate their cross similarity using bilingual topic models and then choose their nearest neighbors (nn) to form a new dictionary.

The cross-lingual similarities between words are calculated according to the procedure described in (Vulić & Moens, 2014). We choose to use the Late-Fusion model in their paper for this purpose. The model performs best in their paper. It is defined as:

$$sim\left(w_i^s, w_j^t\right) = \lambda cos\left(\boldsymbol{x}_i^{s^c}, \boldsymbol{x}_j^t\right) + (1-\lambda) cos\left(\boldsymbol{x}_i^s, \boldsymbol{x}_j^t\right) \tag{1}$$

where cos is a similarity function (cosine similarity used here), $\boldsymbol{x}_i^s$ here represents the topical vector representation of the word w_i^s. The topics are calculated based on the multilingual probabilistic topic modeling (MuPTM) framework (Vulić, De Smet, Tang, & Moens, 2015) (i.e., by bilingual topic models). $\boldsymbol{x}_i^{s^c}$ denotes the vector representations for the co-occurrence-based context set for word w_i^s and λ is the interpolation parameter. We omit the model details here. Readers should refer to their original paper (Vulić & Moens, 2014) for technical details.

3.1. Supervised method

With the augmented dictionary, we now describe the projection details. Our goal is to learn a mapping between the two sets such that the translations are close in the shared space. For this purpose, we need to learn a linear mapping $\boldsymbol{W}$ between the source space and the target space to minimize Euclidean distance

$$\boldsymbol{W}^* = argmin_w \|\boldsymbol{W}\widetilde{\mathscr{X}^s} - \widetilde{\mathscr{X}^t}\|_F \tag{2}$$

where $\boldsymbol{W}$ belongs to the space of $\tau \times \tau$ matrices of real numbers. $\widetilde{\mathscr{X}^s}$ and $\widetilde{\mathscr{X}^t}$ are two aligned matrices of size $\tau \times U$, where U is the size of the augmented dictionary containing pairs of words $\{w_i^s, w_j^t\}_{i,j=1}^{U}$. Xing, et al. (2015) find that the aligning performance can be greatly improved by constraining $\boldsymbol{W}$ to be an orthogonal matrix. Then the above equation can be thought of the famous Procrustes problem, and a closed-form solution can be obtained from the singular value decomposition of $\widetilde{\mathscr{X}^t}\widetilde{\mathscr{X}^{s^T}}$:

$$\boldsymbol{W}^* = argmin_w \|\boldsymbol{W}\widetilde{\mathscr{X}^s} - \widetilde{\mathscr{X}^t}\|_F = \boldsymbol{U}\boldsymbol{V}^T, \text{ subject to } \boldsymbol{W}\boldsymbol{W}^T = \boldsymbol{I}$$

$$\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T = \boldsymbol{SVD}\left(\widetilde{\mathscr{X}^t}\widetilde{\mathscr{X}^{s^T}}\right) \tag{3}$$

When we use readily available bilingual dictionaries, the method is referred to as the supervised method. According to previous studies (Litschko, Glavas, Ponzetto, & Vulić, 2018), we choose the model used in (Smith, et al., 2017) for supervised training of our **C_ATM** model. Their later work (Litschko, Glavas, Vulić, & Dietz, 2019) used two dictionaries to train the CLEs. We discard this attempt because our goal is to test the CLEs as a whole rather than test the effectiveness of our model in resource-lean settings. We retain this for our future work.

3.2. Unsupervised method

In unsupervised settings, dictionaries are iteratively generated by self-learning. In Conneau, et al. (2018)'s work, an adversarial learning-based model is proposed to create word translation pairs in a fully unsupervised way. They learn a projection $\boldsymbol{W}$ from $\widetilde{\mathscr{X}^s}$ to $\widetilde{\mathscr{X}^t}$ by a discriminator with loss:

$$\boldsymbol{L}(\boldsymbol{W}) = \log p\left(source = 1 \middle| \boldsymbol{W}\widetilde{\mathscr{X}^s}\right) - \log p\left(source = 0 \middle| \widetilde{\mathscr{X}^t}\right) \tag{4}$$

$\boldsymbol{W}$ is trained so that the discriminator is unable to distinguish whether a vector originally from $\boldsymbol{W}\widetilde{\mathscr{X}^s}$ produced by a generator or from $\widetilde{\mathscr{X}^t}$. The discriminator is usually a multi-layer perceptron network. The model is trained following the standard training procedure of Generative Adversarial Network (GAN). After training, the projection can be further improved by an iterative bootstrapping procedure similar to supervised methods described above. We choose the model used in (Conneau, et al., 2018) and their open-source implementation for unsupervised training of our unsupervised version of the **C_ATM** model.

It should be noted that other options for both supervised and unsupervised methods certainly exist (Glavaš, et al., 2019), such as ranking-based optimization (Joulin, et al., 2018) and joint training (Wang, et al., 2020), etc. We focus on the current paper the projection-based methods for the following three reasons. Firstly, the primary goal of the current paper is to develop an effective cross-lingual embeddings model with auxiliary topic models rather than specific alignment procure. Thus, we try to study the cross-lingual embedding problem from a different angle. The second point is that the projection-based methods are the main research stream for generating CLEs as mentioned above in the related work. Other methods relax the orthogonality condition imposed on the projection, allow for distortions of the source embedding space after projection. However, the exact nature of these distortions and their impact on different tasks require further investigation. Third, the actual correlation between intrinsic and extrinsic evaluation tasks is not always consistent. For example, rank-based algorithms may work well for bilingual lexicon extraction, but they may fail in other tasks such as cross-lingual information retrieval (Glavaš, et al., 2019). Examining different alignment methods in different evaluation tasks is beyond the

scope of the current paper. We will leave it for future work.

4. Monolingual word embedding spaces

In this section, our motivation is to produce high-quality monolingual word embedding spaces that will be used as a solid foundation for the CLEs generation process. To do this, we firstly introduce a novel topic model refined by integrating word embeddings. Then a novel word embedding model is presented by incorporating the learned topics.

4.1. Refined topic model using word embeddings

Latent Dirichlet Allocation (LDA) is a generative probabilistic model assuming that a document is represented as a multinomial distribution of topics. In this section, we introduce a Refined Topic Model (**RTM**) using word embeddings based on LDA and subsequently describe a procedure of a Revised Topical Skip-gram (**RTS**) model with a loss function to learn topics and embeddings jointly.

Algorithm 2. Generative process for **RT** model

1. **for** each topic $k \in [1, K]$ **do**
2. sample the mixture of words $\phi \sim$ *Dirichlet*(β)
3. **end for**
4. **for** each document d **do**
5. sample the mixture of topics $\theta_d \sim$ *Dirichlet*(α)
6. **for** each word $w_{d,i}$ indexed by $i = 1, \cdots, N_d$ **do**
7. sample a topic $z_{d,i} \sim$ *Multnomial*(θ_d)
8. sample a word $w_{d,i} \sim$ *Multnomial*$\left(\phi_{z_{d,i}}\right)$
9. **for** each dimension r of the embedding of $w_{d,i}$, **do**
10. sample an embedding score $f^r_{d,i} \sim$ *Normal*$\left(\mu_{z_{d,i}}, \sigma_{z_{d,i}}\right)$
11. **end for**
12. **end for**
13. **end for**

Unlike previous studies by incorporating word embedding models into the topic models, our **RTM** model makes two significant modifications. One is that, unlike previous studies that are using fixed word embeddings, we update embeddings during inference to make a full interconnection and interaction between topic models and word embedding models. Another improvement is that our model is not drawn from multivariate Normal distribution. We draw each dimension from a univariate Normal distribution. There are two reasons for this choice. First, the model with univariate Normal distribution can obtain better performance than that with multivariate version. This is caused by the relative independence between each dimension of embedding representation. We assume that the dimensions between a word embedding are relatively independent, which means there is no cross influence. Therefore, it is much better to use the univariate Normal distribution to describe the distribution of each dimension. Second, the model with univariate Normal distribution is more computationally efficient. A univariate normal distribution is described using just the two parameters. For a multivariate distribution, we need an additional parameter (i.e., a covariance matrix) to model the correlation between each pair of random variables, which is more computational complex. It is worth noting that the Normal distribution is not the only choice in our model. Other methods like Softmax or von Mises-Fisher distribution can also fulfill the requirement (Batmanghelich, Saeedi, Narasimhan, & Gershman, 2016). We leave it for our future work.

The probability distribution from the **RTM** model describes the statistical relationships of occurrences in the corpus and includes the dimension information of the pre-trained word embeddings. The generative process is summarized in Algorithm 2.

In the generative process, the embedding scores of words are computed by a word embedding model (skip-gram in our case), which is dynamically updated during the inference process. We use a fixed number of latent topics K, although an automatic version is undoubtedly possible. The posterior distribution of topics depends on both words and word embeddings.

We begin with the joint distribution $p(\boldsymbol{w}, \boldsymbol{f}, \boldsymbol{z} | \alpha, \beta, \mu, \sigma, \mathscr{X}, \mathscr{C})$ where $\mathscr{C}$ refers to the document corpus containing a fixed number of M words $\{w_1, \cdots, w_M\}$, $\mathscr{X}$ refers to pre-trained monolingual word embeddings and N_d denotes the number of the word in document d.

$$
\begin{aligned}
p(\boldsymbol{w}, \boldsymbol{f}, \boldsymbol{z} | \alpha, \beta, \mu, \sigma, \mathscr{X}, \mathscr{C}) &= p(\boldsymbol{w} | \boldsymbol{z}, \beta) p(\boldsymbol{f} | \mu, \sigma, \boldsymbol{z}, \mathscr{X}, \mathscr{C}) p(\boldsymbol{z} | \alpha) \\
&= \int \prod_{d=1}^{|\mathscr{C}|} \prod_{i=1}^{N_d} p\left(w_{d,i} \middle| \phi_{z_{d,i}}\right) \prod_{z=1}^{K} p\left(\phi_z | \beta\right) d\Phi \times \int \prod_{d=1}^{|\mathscr{C}|} \\
&\times \prod_{i=1}^{N_d} p\left(f_{d,i} \middle| \mu_{z_{d,i}}, \sigma_{z_{d,i}}, \mathscr{X}, \mathscr{C}\right) \\
&\times \int \prod_{d=1}^{|\mathscr{C}|} \left(\prod_{i=1}^{N_d} p\left(z_{d,i} | \theta_d\right) p\left(\theta_d | \alpha\right) \right) d\Theta \\
&= \left(\frac{\Gamma\left(\sum_{w=1}^{M} \beta_w\right)}{\prod_{w=1}^{M} \Gamma\left(\beta_w\right)} \right)^{K} \left(\frac{\Gamma\left(\sum_{z=1}^{K} \alpha_z\right)}{\prod_{z=1}^{K} \Gamma\left(\alpha_z\right)} \right)^{|\mathscr{C}|} \times \prod_{d=1}^{|\mathscr{C}|} \\
&\times \prod_{i=1}^{N_d} p\left(f_{d,i} \middle| \mu_{z_{d,i}}, \sigma_{z_{d,i}}, \mathscr{X}, \mathscr{C}\right) \\
&\times \prod_{z=1}^{K} \frac{\prod_{w=1}^{M} \Gamma\left(n_{z_w} + \beta_w\right)}{\Gamma\left(\sum_{w=1}^{M} \left(n_{z_w} + \beta_w\right)\right)} \prod_{d=1}^{|\mathscr{C}|} \frac{\prod_{z=1}^{K} \Gamma\left(n_{d_z} + \alpha_z\right)}{\Gamma\left(\sum_{z=1}^{K} \left(n_{z_d} + \alpha_z\right)\right)}
\end{aligned} \tag{5}
$$

Inference is intractable in our model. We employ collapsed Gibbs sampling (Porteous, et al., 2008) to perform approximate inference. For the sake of simplicity and speed, we estimate the parameters μ and σ by the method of moments, once per iteration of Gibbs sampling. We use conjugate priors to simplify the integrals and obtain the conditional probability by using the chain rule:

$$
\begin{aligned}
p\left(z_{d,i} | \boldsymbol{w}, \boldsymbol{f}, \boldsymbol{z}_{\neg d,i}, \alpha, \beta, \mu, \sigma, \mathscr{X}, \mathscr{C}\right) &= \frac{p\left(z_{d,i}, w_{d,i}, f_{d,i} | \boldsymbol{w}_{\neg d,i}, \boldsymbol{f}_{\neg d,i}, \boldsymbol{z}_{\neg d,i}, \alpha, \beta, \mu, \sigma, \mathscr{X}, \mathscr{C}\right)}{p\left(w_{d,i}, f_{d,i} | \boldsymbol{w}_{\neg d,i}, \boldsymbol{f}_{\neg d,i}, \boldsymbol{z}_{\neg d,i}, \alpha, \beta, \mu, \sigma, \mathscr{X}, \mathscr{C}\right)} \\
&= \frac{p(\boldsymbol{w}, \boldsymbol{f}, \boldsymbol{z} | \alpha, \beta, \mu, \sigma, \mathscr{X}, \mathscr{C})}{p\left(\boldsymbol{w}, \boldsymbol{f}, \boldsymbol{z}_{\neg d,i} | \alpha, \beta, \mu, \sigma, \mathscr{X}, \mathscr{C}\right)} \propto \frac{p(\boldsymbol{w}, \boldsymbol{f}, \boldsymbol{z} | \alpha, \beta, \mu, \sigma, \mathscr{X}, \mathscr{C})}{p\left(\boldsymbol{w}_{\neg d,i}, \boldsymbol{f}_{\neg d,i}, \boldsymbol{z}_{\neg d,i} | \alpha, \beta, \mu, \sigma, \mathscr{X}, \mathscr{C}\right)} \propto \left(n_{d,z_{d,i}} + \alpha_{z_{d,i}} - 1\right) \times \frac{n_{z_{d,i},w_{d,i}} + \beta_{w_{d,i}} - 1}{\sum_{w=1}^{M} \left(n_{z_{d,i},w} + \beta_w\right) - 1} \\
&\times \prod_{r=1}^{\tau} \frac{1}{\sqrt{2\pi} f^r_{d,i} \sigma_{z_{d,i}}} \exp\left(- \frac{\left(f^r_{d,i} - \mu_{z_{d,i}}\right)^2}{2 {\sigma_{z_{d,i}}}^2} \right)
\end{aligned} \tag{6}
$$

We update both $\mu_{z_{d,i}}$ and $\sigma_{z_{d,i}}$ after each Gibbs sample iteration by maximum likelihood estimation:

$$
\widehat{\mu}_{z_{d,i}} = \frac{\sum_{d'=1}^{N} \sum_{i' \wedge \left(z_{d',i'} = z_{d,i}\right)}^{N_d} f_{d',i'}}{n_{z_{d,i}}} \tag{7}
$$

$$\widehat{\sigma}^2_{z_{d,i}} = \frac{\sum_{d'=1}^{N}\sum_{i'\wedge(z_{d',i'}=z_{d,i})}^{N_d}\left(f_{d',i'}-\widehat{\mu}\right)^2}{n_{z_{d,i}}} \tag{8}$$

The likelihood function can be written as:

$$LL = \prod_{i=1}^{\tau} P(f_i;\mu,\sigma) = \prod_{i=1}^{\tau}\frac{1}{\sqrt{2\pi}}\sqrt{\sigma}e^{-\frac{\sigma}{2}(f_i-\mu)^2} = \left(\frac{\sigma}{2\pi}\right)^{\tau\Big/2^{e-\frac{\sigma}{2}\sum_{i=1}^{\tau}(f_i-\mu)^2}} \tag{9}$$

Algorithm 3. **RTM** model inference procedure and embedding updating

1. **for** *iteration* $= 1, 2, \cdots, N_{iteration}$ **do**
2. **for** $d = 1, 2, \cdots, N$ **do**
3. **for** $i = 1, 2, \cdots, N_d$ **do**
4. sample $z_{d,i}$ from $p(z_{d,i}|\boldsymbol{w}, \boldsymbol{f}, \boldsymbol{z}_{\neg d,i}, \alpha, \beta, \mu, \sigma, \mathscr{X}, \mathscr{C})$
5. update $n_{z_{d,i},w_{d,i}}$, $n_{d,z_{d,i}}$
6. **end for**
7. **end for**
8. **for** $z = 1, 2, \cdots, K$ **do**
9. update μ_z, σ_z
10. **end for**
11. **if** *iteration* $>$ *burn* $-$ *in* **then**
12. update the embeddings according to Eq. (13)
13. **end for**

By taking the natural logarithm of the likelihood function, we can get the log-likelihood function as follows:

$$\begin{aligned}\ln LL &= \ln\left(\left(\frac{\sigma}{2\pi}\right)^{\frac{\tau}{2}e^{-\frac{\sigma}{2}\sum_{i=1}^{\tau}(f_i-\mu)^2}}\right) = \ln\left(\left(\frac{\sigma}{2\pi}\right)^{\frac{\tau}{2}}\right) + \ln\left(e^{-\frac{\sigma}{2}\sum_{i=1}^{\tau}(f_i-\mu)^2}\right)\\ &= -\frac{\tau}{2}\ln\left(\frac{2\pi}{\sigma}\right) + \sum_{i=1}^{\tau}\left(-\frac{\sigma}{2}\right)(f_i-\mu)^2\\ &= -\frac{\tau}{2}\ln(2\pi) + \frac{\tau}{2}\ln(\sigma) + \sum_{i=1}^{\tau}\left(-\frac{\sigma}{2}\right)(f_i-\mu)^2\end{aligned} \tag{10}$$

The first two terms for a single embedding dimension are constant numbers. After omitting these terms, the function can be written as:

$$\ln LL = \sum_{i=1}^{\tau}\left(-\frac{\sigma}{2}\right)(f_i-\mu)^2 \tag{11}$$

So that the log-likelihood of all the data given the model parameters can be defined as:

$$L = \sum_{w=1}^{M}\sum_{z=1}^{K} n_{z,w}\sum_{r=1}^{\tau}\left(-\frac{\sigma_{z,r}}{2}\right)\left(f_{w,r}-\mu_{z,r}\right)^2 \tag{12}$$

We employ gradient ascent to maximize the log-likelihood to update the original embeddings. The gradients are calculated as:

$$\frac{\partial L}{\partial f_{w,r}} = \sum_{z=1}^{K} n_{z,w}\left(-\sigma_{z,r}\right)\left(f_{w,r}-\mu_{z,r}\right) \tag{13}$$

The whole process, including inference procedure and embedding updating is summarized in Algorithm 3. First, we fix the embeddings to sample the topics and infer the model parameters (Described in Line 2–10, Algorithm 3). After a number of iterations (i.e. burn-in period), we fix the topics and parameters and use gradients to update the embeddings (Described in Line 11–12, Algorithm 3).

4.2. Revised topical skip-gram model

The skip-gram model learns word embeddings in a similar way to neural language models but without a non-linear hidden layer. In this model, each word corresponds to a unique vector. The object of the Skip-gram is to maximize the probability of predicting context words for each pivot word in the corpus:

$$L_w(\mathscr{C}) = \frac{1}{M}\sum_{i=1}^{M}\sum_{-\varepsilon\le c\le\varepsilon, c\neq 0}\log p(w_{i+c}|w_i) \tag{14}$$

where ε is the context size of a pivot word. The probability of $p(w_c|w_i)$ is computed by using a softmax function:

$$p(w_c|w_i) = \frac{e^{\boldsymbol{x}_c\bullet\boldsymbol{x}_i}}{\sum_{c'\in\mathscr{C}} e^{\boldsymbol{x}_{c'}\bullet\boldsymbol{x}_i}} \tag{15}$$

$\boldsymbol{x}_c$ and $\boldsymbol{x}_i \in \mathbb{R}^{\tau}$ are vector representations of pivot word w_i and context word w_c. τ is the dimension of the embeddings. In the skip-gram model, each word is treated in turn as the pivot, and all pairs of words and context words are appended to the training dataset. Negative sampling and hierarchical softmax can then be used to learn the model (Mikolov, Sutskever, et al., 2013). In particular, the negative sampling approach defines the loss to be:

$$L_w^{neg}(\mathscr{C}) = \sum_{pos}\log\sigma(\boldsymbol{x}_c\bullet\boldsymbol{x}_i) + \sum_{\text{neg}}\log\sigma(-\boldsymbol{x}_c\bullet\boldsymbol{x}_i) \tag{16}$$

Our Revised Topical Skip-gram (**RTS**) model with a joint loss function aims to learn vector representations for words and topics separately and simultaneously. More specifically, we aim to maximize the following log-likelihood:

$$L_{wz}(\mathscr{C}) = \delta L_w(\mathscr{C}) + (1-\delta)L_z(\mathscr{C}) \tag{17}$$

where δ is the linear interpolation parameter that assigns the weight balance between the two constituent models. $L_z(\mathscr{C})$ is the loss that is conducted using a context topic vector and a pivot topic vector:

$$L_z(\mathscr{C}) = \frac{1}{K}\sum_{i=1}^{K}\sum_{-\varepsilon\le c\le\varepsilon, c\neq 0}\log p(z_{i+c}|z_i) \tag{18}$$

Like topical word embeddings (Liu, et al., 2015), we assume each topic appears in all positions of words assigned with this topic. Unlike their work, we use the pivot topic to predict context topics but not context words. They learn topic embeddings while keeping word embeddings unchanged and then concatenate word embeddings and topic embeddings for the final output. We argue that our assumption is more accurate. As assigning topics only at the input side instead of on both sides of the model will ignore some critical information, the vector of a topic will not represent the collective semantics of words and topics. In the above way, we can truly capture the mutual reinforcement features of both words and topics. This assumption is confirmed in our experiments. We also find that using concatenated embeddings is not as effective as using the word embeddings produced by joint learning.

Furthermore, we provide a different way here to compute the joint loss. We directly calculate the softmax and use the cross-entropy loss directly to maximize the probability. To make the computation possible, we propose a negative softmax loss to speed up the embeddings learning and at the same time maintain the generation performance. Instead of taking the probability of the context word compared to the entire possible context words in the vocabulary, our method uses a different way. **RTS** randomly select κ negative samples for each word-context pair, and then form $neg+1$ labels together with the original label and evaluate the probability only from these training pairs:

$$p'(w_c|w_i) = \frac{e^{\boldsymbol{x}_c\bullet\boldsymbol{x}_i}}{e^{\boldsymbol{x}_c\bullet\boldsymbol{x}_i} + \sum_{c'\in neg} e^{\boldsymbol{x}_{c'}\bullet\boldsymbol{x}_i}} \tag{19}$$

We depict an overview of our CLEs learning framework in Fig. 1. It is well-known that noises in data and learning inappropriate representations is a major drawback of embedding-based techniques. Our method attempts to avoid noises that both semantic models may introduce in the following way. In the generation of bilingual dictionaries, we use

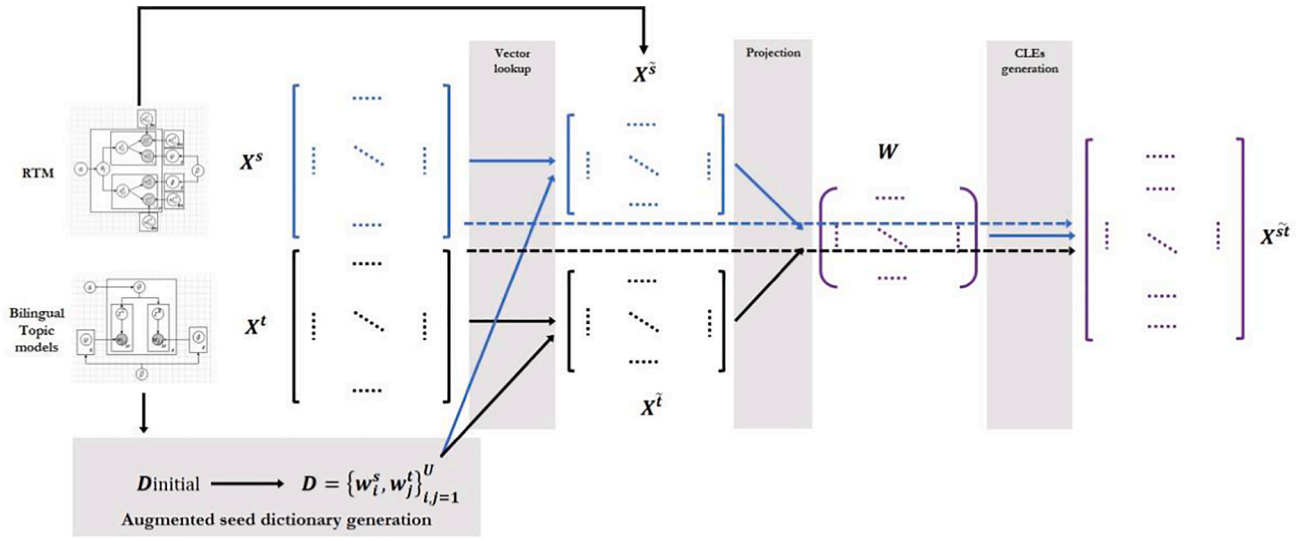


Fig. 1. An overview of our CLEs learning framework.

similarities between words and use information acquired from their contexts. The intuition is that the context helps to disambiguate the true meaning of the occurrence of a word. In other words, after observing the context of a word, fewer latent cross-lingual concepts will share most of the probability mass in the word representation. In the process of CLEs generation, we mutually enhance topic models and word embedding models. However, we only use word embedding models as the main model and output high quality monolingual embeddings instead of monolingual topics. Though there are still some noises in our methods, we demonstrate in the experiments below that our methods work better than previously proposed ones. We will further investigate this issue in our future work.

5. Experiments

In this section, we conduct two sets of extensive experiments to evaluate the proposed **C_ATM** model against the state-of-the-art alternatives. In the first set, we focus on evaluating the proposed **C_ATM** model in bilingual lexicon extraction, cross-lingual semantic word similarity and cross-lingual document classification tasks. Results suggest that our model outperforms state-of-the-art supervised and unsupervised CLE models using basic vector-based semantics only. In the second set of experiments, we focus on evaluating our model by replacing the basic monolingual embeddings with the representative monolingual topical word embeddings. In this evaluation, our proposed model again performs better, leading to a good solution in all cases tested. Furthermore, we compare our augmented seed dictionary with the commonly used dictionary in both supervised and unsupervised methods for generating CLEs. The results prove that using bilingual topic models for enriching the seed dictionary can further improve the performance. We also provide an detail analysis on our results and implications of our work in this section.

5.1. Experimental setup

5.1.1. Evaluation setup

We choose language pairs from English (EN) to {Italian (IT), Spanish (ES), German (DE), French (FR)}. These pairs consist of different degrees of similarities[1]. We use the training data for monolingual and cross-lingual word embeddings generation from document-aligned Wikipedia corpora[2] in all the above languages. As all projection-based methods for inducing cross-lingual embeddings perform similarly, we opt for Smith, et al. (2017)'s approach for supervised training and Conneau, et al. (2018)'s approach for unsupervised training of our **C_ATM** model (see section 3 above). These methods are chosen due to their competitive performance, broad coverage, and readily available implementation[3]. Readers are referred to original papers and online implementations for more information and technique details.

Table 2
Dataset description for primary tasks.

Task	Language pairs	Size
Bilingual Lexicon Extraction	EN, IT	1500
	EN, DE	1500
	EN, ES	1500
	EN, FR	1500
Cross-lingual semantic word similarity	EN, IT	970
	EN, DE	914
	EN, ES	914
	ES, IT	970
	DE, IT	888
	DE, ES	956

5.1.2. Evaluation tasks and datasets

We perform a series of experiments to test the effectiveness and robustness of our **C_ATM** model to learn cross-lingual embeddings. These include two intrinsic (primary) NLP tasks: bilingual lexicon extraction or bilingual dictionary induction & cross-lingual semantic word similarity, and one extrinsic (auxiliary) NLP task: cross-lingual document classification. Statistics of the primary tasks are listed in Table 2.

Bilingual Lexicon Extraction measures the accuracy of the induced word dictionary comparing to a gold standard. We use bilingual dictionaries created by (Conneau, et al., 2018) as the gold standard. Each test set in each language consists of 1500 gold translation pairs. CSLS is used to retrieve nearest neighbors as usual. We report *Precision@k* scores for $k = 1, 5, 10$ (denoted as *P@1, P@5, P@10*). This accounts for a fraction of pairs for which the correct translation of source words is in the k-th nearest neighbors.

Cross-lingual semantic word similarity measures how well the

[1] EN and DE are Germanic languages, IT, ES and FR are Romantic languages

[2] http://linguatools.org/tools/corpora/wikipedia-comparable-corpora/

[3] https://github.com/facebookresearch/MUSE

Table 3
A description of various baseline models together with their corresponded references.

Baseline model	Reference	Description
CLE + WE (supervised)	Smith et al. (2017)	Supervised CLEs learning with readily available bilingual dictionaries
CLE + WE (unsupervised)	Conneau et al. (2018)	Unsupervised CLEs learning with iteratively generated bilingual dictionaries
CLE + TWE-Liu et al. (supervised)	Smith et al. (2017) & Liu et al. (2015)	Monolingual topical word embeddings with supervised CLEs learning
CLE + TWE-Liu et al. (unsupervised)	Conneau et al. (2018) & Liu et al. (2015)	Monolingual topical word embeddings with unsupervised CLEs learning
CLE + TWE-Shi et al. (supervised)	Smith et al. (2017) & Shi et al. (2017)	Monolingual jointly learned word embeddings with supervised CLEs learning
CLE + TWE-Shi et al. (unsupervised)	Conneau et al. (2018) & Shi et al. (2017)	Monolingual jointly learned word embeddings with unsupervised CLEs learning

cosine similarity between two words of different languages correlates with a human-labeled score. We use the SemEval 2017 competition data (Camacho-Collados, Pilehvar, Collier, & Navigli, 2017) and Pearson correlation is adopted as the evaluation metric to report the results.

Cross-lingual document classification is a downstream task using a binary classifier to train and evaluate every topic and every language pair, using pre-defined train and test splits. We use the TED cross-lingual document classification corpus (Glavaš, et al., 2019), and the F1 score is adopted as the evaluation metric to report the results. The results of this auxiliary task are reported in the Section 5.2.5.

5.1.3. Baseline models

To compare our iterative refinement process with previous published monolingual word embeddings and monolingual topical word embeddings, we evaluate our **C_ATM** model and compare with the existing state-of-the-art supervised and unsupervised CLE models trained on word embeddings (WEs), topical word embeddings (TWEs) and our **C_ATM** (**RTM + RTS**) embeddings (we report the primary results with the bootstrapped dictionary as described in algorithm 1, section 3). These form three sets of models in comparison: ***CLE + WE, CLE + TWE*** and ***CLE + C_ATM***. For ***CLE + TWE*** method we provide variants of two topical word embedding models introduced in (Liu, et al., 2015; Shi, et al., 2017). We use ***CLE + C_ATM + u*** to denote our **C_ATM** model with the non-bootstrapped dictionary. A brief description of various baseline models together with their corresponded references is shown in Table 3.

Previous studies use pre-trained vectors, usually fastText[4], as monolingual embeddings to induce CLEs. As our model needs to train word embedding models and topic models together, for a fair comparison, we need to follow the exact training procedures and use exactly the same corpora and pre-process procedures that are unavailable to us. Instead, we have our own implementation of various models, including baselines and our main focus is on producing better monolingual embeddings. So our ***CLE + WE*** baselines are not directly compared to published results (Conneau, et al., 2018; Søgaard, et al., 2018). However, we show that in our experiments, the TWE-based and C_ATM-based models can achieve comparable and even better performance than training the CLEs by using fine-tuned monolingual word vectors.

5.1.4. Parameter settings

There are many parameters involved in various models. Some are chosen based on previous experience and some trails, others are determined by extensive sensitivity experiments. Separate data was utilized for training the necessary parameters. Every effort was made to ensure no overlap between the training data and the test data used in our experiments.

In the **RTM** model, we find that the sensitivity of the hyper-parameters α and β is limited. For simplicity, we set $\alpha = 50/K$ and $\beta = 0.01$ as widely used before. The number of iterations for topic models is set to 1000. The total number of topics K used in our **RTM** model demonstrates the degree of hidden semantics discovered by topic models. There are many ways to determine this parameter (Blei, 2012). However, the primary goal of our current paper is to generate high-quality monolingual topic-enhanced word embeddings rather than high-quality topics. So that we use a fixed number of latent topics, although an automatic version is certainly possible. The selection of parameter K (topic numbers) in both tasks for our **C_ATM** model with language pair EN-IT is illustrated in Fig. 2[5]. The best value of this parameter is 150 with modest topic numbers preferred. Similar trends were observed in other language pairs and evaluation tasks, so we omit them here. As a result, the total number of topics K trained by our **RTM** model is set to 150 for the main results reported.

We now examine the effect of the linear interpolation parameter δ used in our **RTS** model. It assigns the weight balance between the two constituent models $L_w(\mathscr{C})$ and $L_z(\mathscr{C})$. A number of runs were executed with a spread of settings from 0.1 to 0.9 for the parameter. As shown in Fig. 3, interestingly, the best values[6] were obtained when parameter δ reaches 0.5. This further shows that the equal importance of words and topics when learning word embeddings. As a result, the linear interpolation parameter used in our **RTS** model is set to 0.5 in the rest of the paper.

The dimension size τ for all word embeddings is set to 300 as a standard practice. We select 16 negative samples, and the windows size is set to 5. Parameters in our baseline models are set according to their tuning procedures in the original papers or those obtaining the best performance. We refer the reader to their original papers for more implementation details. Our model, as well as previous baseline models were implemented or re-implemented in Python with Tensorflow.

5.1.5. Monolingual results

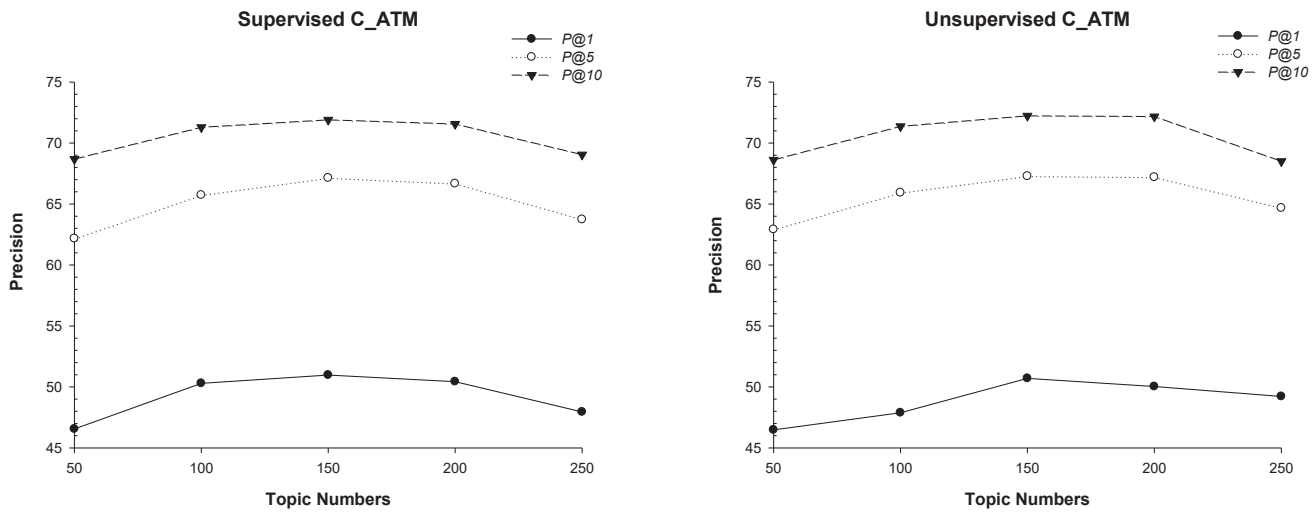
Before we show the effectiveness of our **C_ATM** model in a cross-lingual setting, we firstly demonstrate that our **ATM** model works well monolingually. To evaluate models of similarity for polysemous words, we use Huang, Socher, Manning, and Ng (2012)'s Stanford's Contextual Word Similarity (SCWS) dataset for this purpose. The dataset contains 2003 word pairs together with their context sentences. The ground truth scores were labeled by humans. Also, we compare our results with the two variants of representative topical word embedding models introduced in (Liu, et al., 2015; Shi, et al., 2017). The word similarity is evaluated using cosine similarity with two variants as in previous work *AvgSimC* and *MaxSimC*. Interested readers should refer to (Liu, et al., 2015; Shi, et al., 2017) for details on these two methods. Spearman correlation coefficient is adopted as the evaluation metric.

The results are shown in Table 4. Our **ATM** model outperforms the two state-of-the-art TWE models. However, the vectors in our model are much shorter as we do not concatenate word vectors with topic vectors. We also use a joint loss to train the topic models and word embedding models in a more aggregate way. All these results serve as a solid foundation for evaluating our **C_ATM** model and its variants in the cross-lingual setting, which we present below.

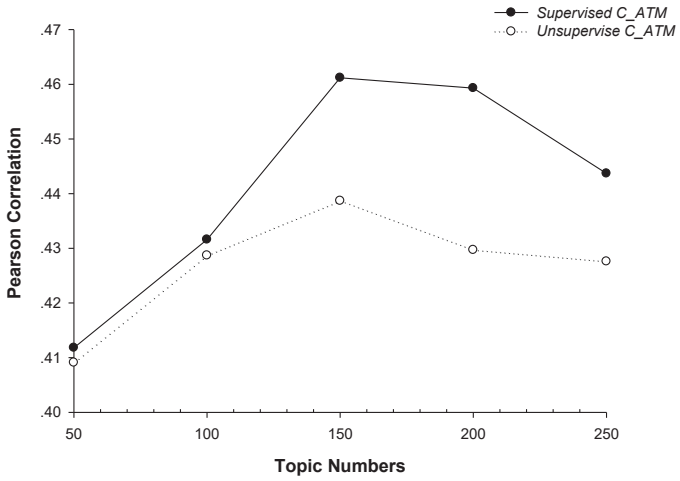
[4] https://fasttext.cc/docs/en/pretrained-vectors.html

[5] Note that in Figs. 2 and 3, separate data was utilized to train the necessary parameters. The size of training data is much smaller than the size of testing data used in the formal experiments.

[6] We only show the behavior of our supervised model in bilingual lexicon extraction task with language pair IT-EN when measured in P@1. Similar results were observed in other tasks and language pairs with different evaluation metrics.



(a). Bilingual lexicon extraction task with language pair EN-IT



(b). Cross-lingual semantic word similarity task with language pair EN-IT

Fig. 2. The impact of varying parameter *K*.

5.2. Results and discussion

5.2.1. Bilingual lexicon extraction task

For the bilingual lexicon extraction task, we report the results using the Conneau et al.'s dictionaries with 1.5 k source test queries in Table 5. In addition, we test results for significance with the two-tailed *t-test* with Bonferroni correction (Dror, Baumer, Shlomov, & Reichart, 2018). As it can be seen, our model clearly outperforms various baseline models, including embedding-based methods and topical embedding-based methods. In particular, our method significantly outperforms all the embedding-based methods ($\rho < 0.01$) and the majority of the topical embedding-based methods except for two runs between language pairs EN-DE.

The methods that only trained on basic word embeddings get the worst results of all by a large margin. Comparing to our **C_ATM** model that captures the inter-correlations of topic models and word embedding models, the performance of CLE models that use TWEs is less impressive. However, all models that use topical features consistently exceed models that are only using embedding features for every evaluation metric with a supervised and unsupervised way to train CLEs. The largest improvement of our **C_ATM** model is *13%* when compared to WE-based model and *11%* compared to other TWE-based models measured by P@1 in supervised CLE models. *17%* compared to the WE-based model and 16% compared to other TWE-based models measured by P@1 in unsupervised CLE models. Similar trends are observed in other evaluation metrics. This confirms the validity of considering topical features in building good word embeddings.

Pleasingly although the performance of ***CLE*** + ***WE*** in our implementation is worse than using the fine-tuned word vectors as in previous papers, our TWE-based and C_ATM-based models achieve competitive performance by using word vectors trained by skip-gram only. In particular, the results obtained by our ***CLE*** + ***C_ATM*** models are promising. Runs employing iteratively refined word embeddings produced the highest score across all evaluation metrics in every single run. Comparing our supervised and unsupervised results with (Conneau, et al., 2018), it can be seen that our proposed model is quite effective.

Our method is more effective than TWE-based models that are previously proposed. The main reason is that the vectors in (Liu, et al.,

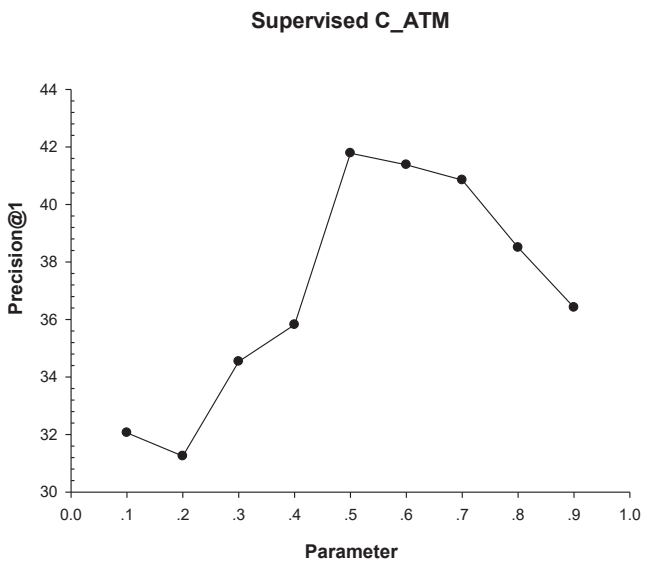


Fig. 3. The impact of varying parameter δ in Bilingual lexicon extraction task with language pair IT-EN.

Table 4
Monolingual results by using spearman correlation $\rho \times 100$ for the SCWS dataset. For illustration purposes the results in the upper part of the baseline methods are copied from (Shi, et al., 2017).

Method	Similarity metrics	$\rho \times 100$
C&W	Cosine Similarity	57.0
Skip-gram	Cosine Similarity	65.7
TFIDF	Cosine Similarity	26.3
Pruned TFIDF	Cosine Similarity	62.5
LDA-S	Cosine Similarity	56.9
LDA-C	Cosine Similarity	50.4
TWE-Liu et al.	*AvgSimC*	68.1
TWE-Liu et al.	*MaxSimC*	67.3
TWE-Shi et al.	*AvgSimC*	68.0
TWE-Shi et al.	*MaxSimC*	67.7
ATM	*AvgSimC*	**69.0**
ATM	*MaxSimC*	**68.2**

2015) have twice as many dimensions as those in our model, as they concatenate topic vectors and word vectors together. We also present a performance of using Shi, et al. (2017)'s TWE model to build CLEs (See ***CLE*** + ***TWE-Shi et al.*** in Table 1). Their monolingual TWE model is a by-product when training topic models and word embedding models together by using a BTM similar process. Again our model achieves better results than theirs in both supervised and unsupervised generation of CLEs. Furthermore, we notice that their model achieves better performance than Liu et al's model except in very few cases (such as in EN-ES). This demonstrates the effectiveness of jointly training topic models and word embedding models. Overall, the model proposed in our paper can effectively improve the bilingual lexicon extraction accuracy comparing to CLE models trained on basic word embeddings and representative monolingual topical word embeddings.

5.2.2. Cross-lingual semantic word similarity task

In the cross-lingual semantic word similarity task, results are reported in Fig. 4 for six language pairs: EN-IT, EN-DE, EN-ES, ES-IT, DE-IT, and DE-ES. One can see that our model again provides a strong and robust gain in performance across all language pairs when comparing to baseline models. We get up to *14%* when compared to the WE-based model and 10% compared to other TWE-based models in supervised CLE models, 22% when compared to the WE-based model and 18% compared to other TWE-based models in unsupervised CLE models. Again we notice a performance drop in unsupervised models, which is in line with the results of bilingual lexicon extraction.

All the results described above confirm the robustness of the proposed model. Indeed, integrating topic models and word embedding models can significantly enhance the embedding performance, which in turn improves the quality of CLE. Our **C_ATM** model converges to a good solution in all runs without exception. The accuracy achieved by our model is always better, surpassing previous embedding-only methods and topical embedding methods.

5.2.3. The effects of different language pairs

The language pairs used in our paper consist of different degree of similarities. EN and DE are Germanic languages, IT, ES, and FR are Romantic languages. They demonstrate different behavior in the experiments. In the bilingual lexicon extraction task, all models achieve better performance for less similar language pairs in EN-IT, EN-ES, and EN-FR than similar language pairs in EN-DE. In particular, results in EN-ES are quite impressive. In the cross-lingual semantic word similarity task, similarities in language pairs from different language families are worse than those from the same language families. For example, results for EN-DE are much better than those for DE-IT. As pointed by (Søgaard, et al., 2018), whether the performance is caused by language differences or by the nature of test collections is unclear. Further investigation is very much needed here. Finally, we notice that the improvements are limited in language pairs EN and DE (from the same family). The morphological properties of two languages may cause this. Clearly, future work can consider incorporating more syntactic and structural information into the process.

5.2.4. Supervised vs unsupervised CLE models

We notice that the behavior of supervised models and unsupervised models is not on par with previous findings (Conneau, et al., 2018). Furthermore, we often observe supervised models achieve better results than their unsupervised counterparts, quite the opposite of previous findings. Finally, we notice this is also confirmed by a number of researchers (Hoshen & Wolf, 2018; Litschko, Glavas, Vulić, & Dietz, 2019). Further investigation is very much needed here.

5.2.5. Bootstrapping strategies

As we have already demonstrated in Sections 5.2.1–5.2.3, our **C_ATM** model outperforms various state-of-the-art alternatives for CLEs generation with the bootstrapped dictionary (for both supervised and unsupervised model) as described in Section 3. However, we are still interested in the research question: if the bootstrapped dictionary improves the performance? Therefore, we depict the results of the ***CLE*** + ***C_ATM*** model and the ***CLE*** + ***C_ATM*** + ***u*** model in the *bilingual lexicon extraction* task in a supervised setting with the metric P@10 in Fig. 5-(a) and in an unsupervised setting with the metric P@10 in Fig. 5-(b) (as similar results were observed in other tasks, settings with different evaluation metrics). Both results show that our bootstrapping approach boots the performance when using bilingual topic models. As expected, the improvements in unsupervised methods are more significant than those in supervised methods. As most unsupervised method just uses a simple seed dictionary for final embedding generation, the extended bilingual dictionary can guarantee to increase the performance. In the supervised methods, when the dictionary is further improved, the performance can also be further enhanced. This is in line with prior findings (Vulić & Moens, 2014) that topic models can still capture certain semantic similarities between words.

5.2.6. Extrinsic evaluation

It has been shown previously that intrinsic evaluation sometimes does not correspond to extrinsic evaluation (i.e., downstream tasks) (Glavaš, et al., 2019). It is interesting to see the performance of our proposed model in such a setting. We run another evaluation task – cross-lingual document classification to examine if our model can

Table 5
Bilingual lexicon extraction results, best results are marked in bold, statistically significant differences between our method and ***CLE + TWE-Shi et al.*** & ***CLE + WE*** are indicated by † and *, respectively.

	EN-IT			IT-EN			EN-ES			ES-EN		
	P@1	P@5	P@10	P@1	P@5	P@10	P@1	P@5	P@10	P@1	P@5	P@10
Supervised Methods												
CLE + WE	70.89	79.53	83.52	69.08	77.22	82.14	75.39	86.26	90.62	77.59	87.49	90.01
CLE + TWE-Liu et al.	77.70	84.81	89.03	70.63	80.07	86.55	79.98	88.73	91.59	83.33	89.59	92.60
CLE + TWE-Shi et al.	78.80	85.43	90.20	72.46	82.21	87.22	80.08	87.38	92.07	82.44	91.18	93.07
CLE + C_ATM	**79.50***†	**87.15***†	**90.73***†	**78.36***†	**85.36***†	**88.74***†	**82.94***†	**90.94***†	**94.68***†	**84.10***†	**91.68***†	**93.55***†
Unsupervised Methods												
CLE + WE	61.85	72.51	78.02	60.31	71.37	77.24	76.50	86.88	90.70	79.34	86.99	90.50
CLE + TWE-Liu et al.	62.40	74.24	81.16	61.88	69.52	78.65	80.73	90.00	91.34	82.84	91.09	92.20
CLE + TWE-Shi et al.	70.05	81.04	84.66	69.13	76.50	79.68	80.93	90.60	92.32	83.52	90.34	92.75
CLE + C_ATM	**72.55***†	**82.38***†	**86.66***†	**70.29***†	**77.06***†	**80.64***†	**83.17***†	**91.53***†	**94.93***†	**84.01***†	**91.25***†	**93.09***†
	EN-DE			**DE-EN**			**EN-FR**			**FR-EN**		
	P@1	P@5	P@10	P@1	P@5	P@10	P@1	P@5	P@10	P@1	P@5	P@10
Supervised Methods												
CLE + WE	69.49	82.45	85.88	63.86	76.93	79.43	76.34	87.04	91.34	73.84	85.40	89.75
CLE + TWE-Liu et al.	74.03	83.57	87.03	65.09	79.47	82.83	82.49	92.02	95.34	80.05	90.36	94.29
CLE + TWE-Shi et al.	74.15	83.66	87.91	71.57	**83.07**	84.25	82.71	93.34	95.85	81.19	90.09	93.42
CLE + C_ATM	**74.72***†	**85.54***†	**88.44***†	**72.27***†	**83.07***	**86.63***†	**83.59***†	**94.00***†	**96.19***†	**83.04***†	**91.35***†	**94.64***†
Unsupervised Methods												
CLE + WE	66.61	80.10	85.07	68.82	82.19	85.28	76.23	88.07	91.41	77.28	86.66	89.05
CLE + TWE-Liu et al.	70.67	84.13	87.65	68.93	82.37	85.28	78.52	92.30	95.08	81.18	90.00	93.77
CLE + TWE-Shi et al.	71.59	84.70	87.91	69.22	82.46	85.95	82.05	94.08	96.10	81.18	90.85	93.94
CLE + C_ATM	**72.28***†	**84.88***	**88.18***†	**71.06***†	**82.98***†	**86.41***†	**83.15***†	**94.19***†	**96.44***†	**82.64***†	**91.16***†	**94.20***†

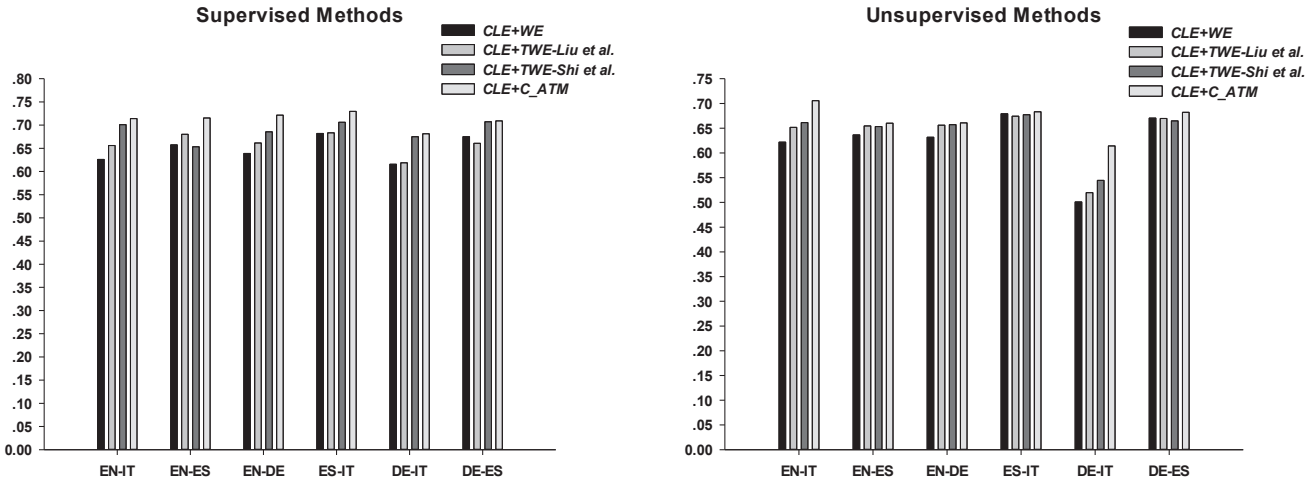


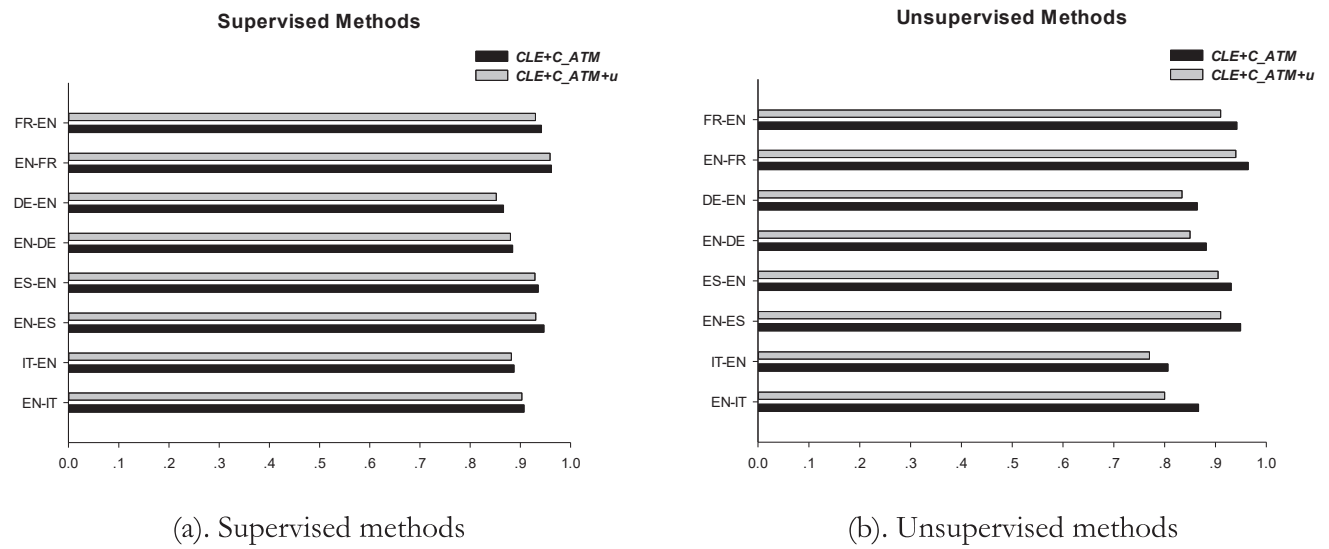
Fig. 4. Cross-lingual semantic word similarity results.

improve over simpler models. Follow (Glavaš, et al., 2019), we use a lightweight CNN-based binary classifier and pre-defined train and test data. Note that the results of our own implementation are slightly different from the figures reported in Glavaš et al.'s work. We report results with unsupervised methods on language pairs EN-IT, EN-DE, EN-FR in Fig. 6.

In contrast to the results of the intrinsic evaluation, CLE models that use TWEs underperforms the basic word embedding-based model. It can be seen that our ***C_ATM*** model again achieves improvements over all baseline models, including CLE models trained on basic word embeddings and representative monolingual topical word embeddings. Moreover, the improvements are always statistically significant. There exist other options for generating CLEs. However, the primary goal of the current paper is to study the cross-lingual embedding problem from a different angle rather than focus on the specific projection methods. We leave the further experiments and analyses as our future work.

5.2.7. *Implications of the work*

As mentioned at the beginning of this paper, CLEs are appealing due to their excellent capability to enable multilingual semantics and increase the possibility of knowledge transfer. Previous work mainly focuses on the mapping/projection process, given that high-quality monolingual embeddings are readily available. However, this assumption is actually in doubt. So that in this paper, we tackle the problem from a different angle. Our paper can be viewed as a pioneer work for creating better CLEs based on high-quality monolingual embeddings. It is important to note that topic models are not the only way to enhance word embeddings. The subsequent work may integrate other representative hidden semantic models such as manifold-learning-based models (Zhou, Bousquet, Lal, Weston, & Scholkopf, 2004). Our method is built upon mutual reinforcement of the two models. Clearly, more advanced framework can be proposed and better integrate two or more semantic models. Furthermore, the CLEs generated by our proposed method can be used as a backbone in many real-world applications, particularly machine translation, cross-lingual or multilingual information retrieval



(a). Supervised methods (b). Unsupervised methods

Fig. 5. Effectiveness of the bootstrapping strategy.

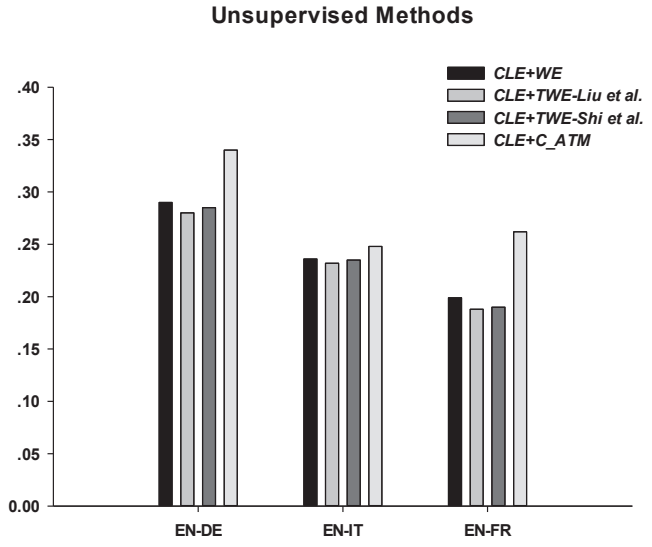


Fig. 6. Cross-lingual document classification results.

and processing, and so on. Finally, our method is suited to handle multilingual applications between resource-rich and resource-lean language pairs or just between resource-lean language pairs. For languages with a very limited amount of parallel data, our bootstrapping method for generating a bilingual dictionary will be quite valuable.

6. Conclusion

Despite the exciting and novel development of cross-lingual word embedding models, topical semantic models in such a setting is not yet exploited. In this work, we show for the first time that better cross-lingual embeddings can be enhanced by monolingual and bilingual topic models. We try to tackle two fundamental research problems in this paper. One is to generate good CLEs by inducing high-quality monolingual word embedding spaces through iteratively refining topic models and word embedding models. Another is that we try to create a better bilingual seed dictionary used for both supervised and unsupervised methods to generate CLEs. We show that our proposed model outperforms existing supervised and unsupervised CLE models trained on basic word embeddings. Our model also exceeds CLE models trained on representative monolingual topical word embeddings. We further prove that using bilingual topic models for enriching the seed dictionary can further improve performance. We hope that this paper will guide the further development of better CLE models from a different angle. In the future, we will evaluate our method on more language pairs with different projection methods and tasks to investigate the differences. We also plan to investigate a more unified way to incorporate our monolingual topical word embeddings into the generation process of cross-lingual embeddings.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We would like to thank anonymous reviewers for their helpful comments and suggestions. This work was supported by the National Natural Science Foundation of China under Project No. 61876062 and General Key Laboratory for Complex System Simulation, China under Project No. XM2020XT1004.

References

Adams, O., Makarucha, A., Neubig, G., Bird, S., & Cohn, T. (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 (EACL)* (pp. 937-947). Valencia, Spain.

Artetxe, M., Labaka, G., & Agirre, E. (2017). Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th annual meeting of the association for computational linguistics (ACL)* (pp. 451-462). Vancouver, Canada: Association for Computational Linguistics.

Artetxe, M., Labaka, G., & Agirre, E. (2018). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th annual meeting of the association for computational linguistics (Volume 1: Long Papers)* (pp. 789-798). Melbourne, Australia: Association for Computational Linguistics.

Bai, X., Cao, H., Chen, K., & Zhao, T. (2019). A Bilingual adversarial autoencoder for unsupervised bilingual lexicon induction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27*(10), 1639–1648.

Batmanghelich, K., Saeedi, A., Narasimhan, K., & Gershman, S. (2016). Nonparametric spherical topic modeling with word embeddings. In *Proceedings of the 54th annual meeting of the association for computational linguistics (Volume 2: Short Papers)* (pp. 537-542). Berlin, Germany: Association for Computational Linguistics.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77–84.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Camacho-Collados, J., Pilehvar, M. T., Collier, N., & Navigli, R. (2017). SemEval-2017 Task 2: Multilingual and cross-lingual semantic word similarity. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)* (pp. 15-26). Vancouver, Canada: Association for Computational Linguistics.

Cheng, X., Yan, X., Lan, Y., & Guo, J. (2014). BTM: Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering, 26*(12), 2928–2941.

Choudhury, M., & Deshpande, A. (2021). How linguistically fair are multilingual pre-trained language models? In *Proceedings of the 35th AAAI Conference on Artificial Intelligence* (pp. 12710-12718). Online.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., & Jegou, H. (2018). Word translation without parallel data. In *Proceedings of the 2018 international conference on learning representations* (pp. 1-14). Vancouver, BC, Canada.

Dabre, R., Chu, C., & Kunchukuttan, A. (2020). A survey of multilingual neural machine translation. *ACM Computing Surveys, 53, Article 99*(5), 1–38.

Dinu, G., & Baroni, M. (2015). Improving zero-shot learning by mitigating the hubness problem. In *International Conference on Learning Representations (workshops)* (pp. 1-10). San Diego, CA, USA.

Dolamic, L., & Savoy, J. (2010). Retrieval effectiveness of machine translated queries. *Journal of the Association for Information Science and Technology, 61*(11), 2266–2273.

Dror, R., Baumer, G., Shlomov, S., & Reichart, R. (2018). The Hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th annual meeting of the association for computational linguistics (Volume 1: Long Papers)* (pp. 1383-1392). Melbourne, Australia: Association for Computational Linguistics.

Faruqui, M., & Dyer, C. (2014). Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th conference of the European chapter of the association for computational linguistics* (pp. 462-471). Gothenburg, Sweden: Association for Computational Linguistics.

Fu, X., Wang, T., Li, J., Yu, C., & Liu, W. (2016). Improving distributed word representation and topic model by word-topic mixture model. In *Proceedings of the 8th Asian Conference on Machine Learning* (pp. 190-205). Hamilton, New Zealand.

Glavaš, G., Litschko, R., Ruder, S., & Vulić, I. (2019). How to (Properly) Evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 710-721). Florence, Italy: Association for Computational Linguistics.

Guo, J., Che, W., Yarowsky, D., Wang, H., & Liu, T. (2016). A distributed representation-based framework for cross-lingual transfer parsing. *Journal of Artificial Intelligence Research, 55*, 995–1023.

Heyman, G., Vulić, I., & Moens, M.-F. (2017). Bilingual lexicon induction by learning to combine word-level and character-level representations. In *Proceedings of the 15th conference of the european chapter of the association for computational linguistics: Volume 1, Long Papers* (pp. 1085-1095). Valencia, Spain: Association for Computational Linguistics.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 50-57). New York, NY, USA: ACM.

Hoshen, Y., & Wolf, L. (2018). Non-adversarial unsupervised word translation. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 469-478). Brussels, Belgium: Association for Computational Linguistics.

Huang, E., Socher, R., Manning, C., & Ng, A. (2012). Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th annual meeting of the association for computational linguistics* (pp. 873-882). Jeju Island, Korea: Association for Computational Linguistics.

Joulin, A., Bojanowski, P., Mikolov, T., Jégou, H., & Grave, E. (2018). Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 2979-2984). Brussels, Belgium: Association for Computational Linguistics.

Karimzadehgan, M., & Zhai, C. (2010). Estimation of Statistical Translation Models based on Mutual Information for Ad Hoc Information Retrieval. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '10* (pp. 323–330). New York, New York, USA: ACM.

Li, C., Wang, H., Zhang, Z., Sun, A., & Ma, Z. (2016). Topic Modeling for short texts with auxiliary word embeddings. In *Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval* (pp. 165-174). Pisa, Italy: ACM.

Litschko, R., Glavas, G., Ponzetto, S. P., & Vulić, I. (2018). Unsupervised cross-lingual information retrieval using monolingual data only. In *Proceedings of the 41st international ACM SIGIR conference on research & development in information retrieval* (pp. 1253-1256). Ann Arbor, MI, USA: ACM.

Litschko, R., Glavas, G., Vulić, I., & Dietz, L. (2019). Evaluating resource-lean cross-lingual embedding models in unsupervised retrieval. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval* (pp. 1109-1112). Paris, France: ACM.

Liu, Y., Liu, Z., Chua, T.-S., & Sun, M. (2015). Topical word embeddings. In *Proceedings of the twenty-ninth AAAI conference on artificial intelligence (AAAI)* (pp. 2418-2424). Austin, Texas, USA.

Mikolov, T., Le, Q. V., & Sutskever, I. (2013). Exploiting Similarities among Languages for Machine Translation. *CoRR, abs/1309.4168*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International conference on neural information processing systems - volume 2* (pp. 3111-3119). Lake Tahoe, Nevada: Curran Associates Inc.

Mogadala, A., & Rettinger, A. (2016). Bilingual word embeddings from parallel and non-parallel corpora for cross-language text classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies (NAACL-HLT)* (pp. 692-702). San Diego, California, USA.

Nguyen, D. Q., Billingsley, R., Du, L., & Johnson, M. (2015). Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics, 3*, 299–313.

Ni, X., Sun, J.-T., Hu, J., & Chen, Z. (2009). Mining multilingual topics from wikipedia. In *Proceedings of the 18th international conference on World wide web - WWW '09* (pp. 1155-1156). New York, New York, USA: ACM.

Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., & Welling, M. (2008). Fast Collapsed Gibbs Sampling for Latent Dirichlet Allocation. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '08* (pp. 569–577). New York, New York, USA: ACM.

Ruder, S., Vulić, I., & Søgaard, A. (2019). A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research, 65*, 569–631.

Shi, B., Lam, W., Jameel, S., Schockaert, S., & Lai, K. P. (2017). Jointly learning word embeddings and latent topics. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval* (pp. 375-384). Shinjuku, Tokyo, Japan: ACM.

Smith, S. L., Turban, D. H. P., Hamblin, S., & Hammerla, N. Y. (2017). Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *CoRR, abs/1702.03859*.

Søgaard, A., Ruder, S., & Vulić, I. (2018). On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 778-788). Melbourne, Australia: Association for Computational Linguistics.

Vulić, I., De Smet, W., & Moens, M.-F. (2013). Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora. *Information Retrieval, 16*(3), 331–368.

Vulić, I., De Smet, W., Tang, J., & Moens, M.-F. (2015). Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications. *Information Processing & Management, 51*(1), 111–147.

Vulić, I., & Moens, M.-F. (2014). Probabilistic models of cross-lingual semantic similarity in context based on latent cross-lingual concepts induced from comparable data. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 349-362). Doha, Qatar: Association for Computational Linguistics.

Vulić, I., & Moens, M.-F. (2015). Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval* (pp. 363-372). Santiago, Chile: ACM.

Vulić, I., Ruder, S., & Søgaard, A. (2020). Are all good word vector spaces isomorphic? *CoRR, abs/2004.04070*.

Wang, Z., Xie, J., Xu, R., Yang, Y., Neubig, G., & Carbonell, J. G. (2020). Cross-lingual alignment vs joint training: A comparative study and a simple unified framework. In *Proceedings of the 8th international conference on learning representations* (pp. 1-15). Addis Ababa, ETHIOPIA.

Xing, C., Wang, D., Liu, C., & Lin, Y. (2015). Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 1006-1011). Denver, Colorado: Association for Computational Linguistics.

Zhou, D., Bousquet, O., Lal, T. N., Weston, J., & Scholkopf, B. (2004). Learning with local and global consistency. *Advances in Neural Information Processing Systems, 16*, 321–328.

Zhou, D., Wu, X., Zhao, W., Lawless, S., & Liu, J. (2017). Query expansion with enriched user profiles for personalized search utilizing folksonomy data. *IEEE Transactions on Knowledge and Data Engineering, 29*(7), 1536–1548.

Zhou, D., Zhao, W., Wu, X., Lawless, S., & Liu, J. (2018). An iterative method for personalized results adaptation in cross-language search. *Information Sciences, 430-431*, 200–215.

Zou, W. Y., Socher, R., Cer, D., & Manning, C. D. (2013). Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 conference on empirical methods in natural language processing (EMNLP)* (pp. 1393-1398). Seattle, Washington, USA.