Information Sciences 608 (2022) 809–824

Contents lists available at ScienceDirect

Information Sciences

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

Neural topic-enhanced cross-lingual word embeddings for CLIR

Dong Zhou [a,*], Wei Qu [b], Lin Li [c], Mingdong Tang [a], Aimin Yang [d]

[a] School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou, Guangdong 510006, China
[b] School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan, Hunan 411201, China
[c] School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan, Hubei 430070, China
[d] School of Computer Science and Intelligence Education, Lingnan Normal University, Zhanjiang, Guangdong 524048, China

ARTICLE INFO

Article history:
Received 5 November 2021
Received in revised form 20 June 2022
Accepted 22 June 2022
Available online 27 June 2022

Keywords:
Cross-Lingual Information Retrieval
Cross-lingual Word Embeddings
Neural Generative Models
Word Embedding Models

ABSTRACT

Cross-lingual information retrieval (CLIR) methods have quickly made the transition from translation-based approaches to semantic-based approaches. In this paper, we examine the limitations of current unsupervised neural CLIR methods, especially those leveraging aligned cross-lingual word embedding (CLWE) spaces. At the moment, CLWEs are normally constructed on the monolingual corpus of bilingual texts through an iterative induction process. Homonymy and polysemy have become major obstacles in this process. On the other hand, contextual text representation methods often fail to outperform static CLWE methods significantly for CLIR. We propose a method utilizing a novel neural generative model with Wasserstein autoencoders to learn neural topic-enhanced CLWEs for CLIR purposes. Our method requires minimal or no supervision at all. On the CLEF test collections, we perform a comparative evaluation of the state-of-the-art semantic CLWE methods along with our proposed method for neural CLIR tasks. We demonstrate that our method outperforms the existing CLWE methods and multilingual contextual text encoders. We also show that our proposed method obtains significant improvements over the CLWE methods based upon representative topical embeddings.

1. Introduction

Extracting meaningful semantic representations from lexical forms of words is very attractive in natural language processing (NLP) and information retrieval (IR) [34]. It is particularly beneficial for the research community to utilize these representations to solve the cross-lingual information retrieval (CLIR) problem, an extreme case that documents and queries are written in distinct languages. Nevertheless, only until recently encouraging results have been reported by using cross-lingual word embeddings (CLWEs) [4,27,28,48]. CLWEs promise to induce bilingual semantic representations or lexicons with only a very small dictionary or without any aligned bilingual signals at all [35,46], and would be a major step toward CLIR to, from, or even between resource-lean languages.

Vulić and Moens [44] made the pilot study employing neural-net-inspired CLWEs in the CLIR setting. They used CLWEs trained from comparable documents and showed that the approach outperforms bilingual topic models like bilingual latent Dirichlet allocation (BLDA) [43]. However, their method still needs a certain amount of bilingual training data aligned at the document level. Litschko et al. [26] found that the recent advances in CLWEs made it possible to execute CLIR with a minimal

* Corresponding author.
E-mail addresses: dongzhou@gdufs.edu.cn (D. Zhou), 286116867@qq.com (W. Qu), cathylilin@whut.edu.cn (L. Li), mdtang@gdufs.edu.cn (M. Tang), amyang18@163.com (A. Yang).

https://doi.org/10.1016/j.ins.2022.06.081
0020-0255/

amount of aligned bilingual data (with or without a very small seed dictionary). Their paper offers an effective way of using CLWEs to conduct unsupervised CLIR methods with competitive performance. In a follow-up study [27], they investigated the effectiveness of CLWEs in resource-lean settings and confirmed that CLWEs indeed offer a competitive and viable solution for CLIR. Furthermore, they evaluated pre-trained multilingual text encoders based on the Transformer architecture [28] for CLIR. They discovered that the off-the-shelf contextual multilingual embeddings fail to significantly outperform the models based on CLWEs.

All attempts mentioned above simply borrow the extant CLWE models [39] for CLIR. Some studies also consider another popular representative semantic model (i.e., topic models [3]) to perform CLIR tasks. Research has shown that these two models actually have certain connections and differences from each other [30]. Word embeddings can be considered a continuous process by learning center words in the context (or vice versa) to generate semantic representations. Topic models adopt a global approach to learn semantic representations by equally considering all words in a document. This attempt is generally considered as a discrete learning process. Currently, two models are usually utilized separately for most cross-lingual tasks, as is the case in CLIR.

On the contrary, in the monolingual setting, semantic representations can be generated more accurately by combing these two powerful models. Topic models such as probabilistic LSA [18] and LDA [3] can be greatly enhanced by word embeddings [25]. Alternatively, to solve the problems of homonymy and polysemy, word embeddings can be intensified by topic models [30]. Jointly learning the two models so as to generate better semantic representations is studied in [38]. Recently, the use of neural topic models resulted in good performance [20,32,45]. There are also some early attempts to integrate the neural models with word embeddings [10].

Though this field of work shows some promising results, there are two prominent shortages of the current studies. 1) The effectiveness of topic-enhanced word embeddings in bilingual settings is not fully explored; 2) There is a lack of study on utilizing more recent and advanced neural topic models [32,45] in cross-lingual embedding generation processes. It is well known that topic models like LDA commonly require expensive iterative inference methods which can be avoided by these neural models [29].

We try to fill the gap by proposing a method utilizing a novel neural generative model with Wasserstein autoencoders to learn Enhanced CROSS-Lingual word Embeddings (EXLE). We approach the problem using a three-part method to model neural topic relevance and word embeddings in mutual reinforcement. In the first part, to make the topics and embeddings more informative than the widely adopted semantic models, we develop a neural generative model with Wasserstein autoencoders by simultaneously processing topics and enhanced word embeddings. The second part develops a Rectified Skip-Gram Model (R-SGM) to train original embeddings, neural topics, and enhanced embeddings in a unified way. After obtaining embeddings in two individual languages, in the third part, we employ a projection-based method [19,39] to map these two separate embeddings to a shared bilingual space to produce CLWEs. We then follow previous work [26,27] to use CLWEs for CLIR.

We utilize CLIR tasks on standard CLEF[1] test collections as benchmarks for a systematic evaluation of our proposed method for a range of different language combinations. We perform a series of extrinsic evaluations. We first evaluate the CLIR performance based on our EXLE method as a whole. Results suggest that the EXLE outperforms other state-of-the-art semantic CLWE methods that use aligned monolingual static and contextual word embeddings (BERT [9], for example) for neural CLIR. Then we evaluate our CLWE training procedure to check if it produces better CLIR results than using classical monolingual topical word embeddings. In this experiment, the EXLE also demonstrates better performance than previously proposed topical word embeddings, delivering statistically significant improvements. Our method can even outperform the CLIR method which uses a neural machine translation system in some settings. We also run an intrinsic evaluation to investigate if the results align with those of the extrinsic evaluation.

The contributions of the current article are:

i. We present a fully unsupervised CLIR method based on neural topic-enhanced cross-lingual word embeddings. The method constructs a shared bilingual semantic space by projecting two enhanced topical embeddings in distinct languages. The process of considering neural topic models and word embeddings simultaneously for neural CLIR is first introduced by our method, to the best of our knowledge.
ii. We propose a three-part method to model neural topic relevance and word embeddings in a mutual reinforcement way. We use both a neural generative model and a rectified skip-gram model while training the enhanced monolingual word embeddings, which are significantly different from previous approaches.
iii. Extensive experiments on different language pairs demonstrate that our proposed method outperforms representative static and contextual CLWEs, as well as other topic-enhanced word embeddings.

In the remaining parts of this article, Section 2 summarizes the related work. Section 3 illustrates our EXLE method in detail for generating neural topic-enhanced CLWEs. We depict the experimental evaluation and results in Section 4 for unsupervised neural CLIR. We provide a conclusion and outline future work in Section 5.

[1] https://www.clef-initiative.eu.

2. Related work

2.1. Semantic-based CLIR

Queries and documents in cross-lingual information retrieval tasks are typically composed in two distinct languages. Therefore, it often requires incorporating some facility for language translation. Techniques employing direct translation of queries and documents using dictionaries, parallel or comparable corpora, and machine translation systems are thoroughly examined [49,50]. However, translation creates an extra burden and often leads to performance degradation. As a result, researchers are actively looking for CLIR systems to solve the query-document mismatch problem by transforming the query and the document representations into a "third" semantic space prior to comparison [17,24].

The effort dates back to over 30 years when Dumais et al. [12] used a technique known as latent semantic indexing (LSI) for CLIR. As an alternative, CLIR can be implemented using explicit semantic analysis [7]. Following the same line of thinking, monolingual and multilingual topic models are also used for CLIR, particularly for extracting semantics from document-aligned comparable corpora [43,44].

Methods resorting to distributed real-valued semantic representations have been investigated and proven effective for ad-hoc information retrieval [44]. Vulić et al. found that utilizing word embeddings produces better results than topic models for information retrieval in both monolingual and cross-lingual settings [44]. However, using semantic-based approaches alone still cannot obtain satisfactory performance. At the same time, the fast advancement of CLWEs provides an ideal solution for CLIR, especially to, from, or even between resource-lean languages. Litschko et al. [26] showed that various types of CLWEs can be used to construct an effective unsupervised CLIR method. In particular, unsupervised generated CLWEs are pretty appealing as no aligned bilingual data is needed. They further evaluated CLWEs in resource-rich and resource-lean languages [27], confirming that using CLIR tasks is more effective in evaluating CLWEs than using word translation tasks. Bonab et al. [4] proposed narrowing the translation gap between two languages by using CLWEs for an end-to-end neural CLIR task. Yu and Allan [48] examined interaction-based neural matching models for CLIR through CLWEs. Both studies confirm that adopting CLWEs is quite effective.

Contextual word embeddings such as multilingual BERT [9] can also be utilized for building a CLIR system. Jiang et al. [21] employed multilingual BERT directly as a matching model for CLIR. However, the work is performed in a supervised setting. Litschko et al. [28] evaluated pre-trained multilingual text encoders based on the Transformer architecture for unsupervised CLIR and discovered that the off-the-shelf contextual multilingual embeddings fail to outperform the models based on CLWEs significantly.

2.2. Cross-lingual word embeddings

Multilingual contextual embeddings have shown certain theoretical and practical attractions. However, the computational cost is very high due to the massive number of parameters. In comparison with contextual models, static CLWEs [35,41] are generally efficient and straightforward, showing wider applicability in multilingual applications. Earlier CLWEs are often obtained by aligning words, sentences, or documents from a bilingual corpus [31]. Due to the fast development of semantic representations, CLWE models have gradually evolved from just aligning vocabularies to aligning word embedding spaces. The most representative methods are projection-based models [1,11,39]. These models require little or no supervision to create a mapping or projection between two pre-trained word embeddings.

However, most of the above methods focus on developing a mapping/projection strategy. They usually assume that the pre-trained monolingual word embeddings are of high quality in this process. In fact, there are problems with this assumption. If word embeddings are of low quality, no matter how complicated the projection process is, it cannot ensure to generate good CLWEs. At the same time, these methods do not consider the use of topic models to enhance CLWEs.

2.3. Combining word embeddings with topics

With the rapid development of word embeddings and topic models, there are many studies combining the two models to produce better monolingual semantic representations. Liu et al. [30] treated topics as pseudo words in the process of jointly learning. They further concatenated topical and word embeddings as the final output. The method proposed by Das et al. [8] started from the multivariate Gaussian distribution of word embeddings and generated discrete tokens to construct semantic representations. Similarly, word embeddings can be used to enhance topic models. For example, auxiliary word embeddings were used to prompt semantically related words and further enhanced a topic model for short texts [25]. Word embedding and topics can also be learned at the same time. As in [38], matrices were utilized instead of vectors to train both topics and word embeddings. Fu et al. [13] presented a hybrid model training word embeddings with the help of topic models, and then further learning topics with the enhanced embeddings.

In the joint training method [38] mentioned above, word embeddings are updated in each iteration during the training of LDA. Instead we model word embeddings, enhanced neural topics, and enhanced word embeddings in a more aggregate way. Furthermore, we directly utilize embeddings in our method instead of matrices as used in their work. Our method is also fundamentally different from those methods [13,30] that treat topics as pseudo words and train different embeddings sep-

arately. Our approach avoids the inherent noise problem associated with the naïve word2vec models. In addition, we utilize a generative model with Wasserstein autoencoders instead of probabilistic topic models in our work.

3. The EXLE method

We now present our method utilizing a novel neural generative model with Wasserstein autoencoders to learn Enhanced CROSS-Lingual word Embeddings (i.e., the EXLE method) in detail. The method assumes that the two representation models mutually reinforce each other. Fig. 1 demonstrates the overall framework of the EXLE method. It consists of three parts to generate enhanced monolingual embeddings and CLWEs: (I) We first use the state-of-the-art techniques to obtain initial word embeddings from large document collections. These pre-trained embeddings are incorporated into a novel neural generative model that we name it Wasserstein autoencoders-based Neural Generative Model (W-NGM) for simultaneously updating. (II) We train a Rectified Skip-Gram Model (R-SGM) to integrate word embeddings, neural-enhanced topics, and neural-enhanced word embeddings to create high-quality monolingual topical word embeddings. We develop a joint loss function for this very purpose. (III) Projection methods are subsequently applied to get the final CLWEs. We summarize primary notations used in this paper in Table 1.

3.1. Part I: A novel neural generative model W-NGM

In the LDA model, we assume that a multinomial distribution of topics constitutes a document [3]. This famous generative probabilistic model, though effective, has apparent problems. It usually requires collapsed Gibbs sampling or the mean-field method for inference. The close form solution and the approximation methods of the LDA model may lead to the inaccurate inference of parameters and low efficiency with a large volume of data.

Algorithm 1 Generative Process of W-NGM

for each word position $i \in \{1, \cdots, N\}$:
sample a topic $z_i \sim Multnomial(\theta)$
sample a pivot word $w_i \sim Multnomial(\beta_{z_i})$
transform θ_i to ψ_i with an MLP
for each context word position $j \in \{1, \cdots, \varepsilon\}$:
sample an embedding index $a_{i,j} \sim Categorical(\psi_i)$
sample a context word $w_{i,j} \sim Multnomial\left(\gamma_{a_{i,j}}\right)$

Neural generative models that employ inference mechanisms with neural networks may solve the expensive inference problem. We opt for a model in the Wasserstein autoencoders [32] framework. However, a limitation of their model is that word embeddings are not considered simultaneously with the topics. Instead, our W-NGM model tries to model neural topic relevance and word embeddings in mutual reinforcement. We now describe the model in detail.

In the W-NGM model, we assume that there exists a corpus $\mathscr{C} = (w_1, \cdots, w_V)$ containing V words. Word w_i can be chosen and used as a center word or pivot word as in the Skip-Gram model [31]. $\boldsymbol{v}_i$ is used to denote the vector representation of w_i. The size of pivot words $N < V$. A surrounding word or a context word is denoted as w_{i+c} where $-\varepsilon \leq c \leq \varepsilon$ and ε is the context size of w_i. Given w_i, we use a vector $\boldsymbol{u}_i \in \mathbb{R}^V$ to denote the representation of a group of context words. By assuming the interchangeability of context words within the window ε, we use the classical bag-of-words (BoWs) to represent these words.

Let $z_i \in \{1, \cdots, K\}$ of w_i represents a topical distribution. This variable is used to generate the context words together with the global topics (corpus wide). K is the pre-defined number of topics. z_i is drawn according to the multinomial distribution $\theta \in \mathbb{R}^{N \times K}$ ($\sum \theta_k = 1, \theta_k \geq 0, k = 1, \cdots, K$). In the reconstruction phase, θ is transformed into a distribution $\psi \in \mathbb{R}^{N \times \tau}$ used for sampling embedding indexes. τ is the dimension of the word embeddings. Next, in order to obtain the semantic distributions of words through topic distributions, we first select an embedding index and generate the context words according to its corresponding multinomial distribution sampling. The generative process of our W-NGM model is given in Algorithm 1.

Unlike LDA, we use neural networks to infer the required distributions. We will illustrate them in detail below. β is a $K \times V$ matrix denoting global topics. γ is a $\tau \times V$ matrix representing corpus-wide word embeddings. Given the observed variables $\{\boldsymbol{v}_i, \boldsymbol{u}_i\}_{i=1}^N$, the objective of our model is to infer $p(\theta|\boldsymbol{v}, \boldsymbol{u})$, the similar posterior to the LDA. As illustrated in Fig. 2, our W-NGM model consists of two main components that are both constructed by a neural network.

3.1.1. Encoder

Our model follows the Wasserstein autoencoders framework. The first component is an encoder containing a multi-layer perceptron (MLP). MLP takes $\boldsymbol{v}$ and $\boldsymbol{u}$ as the inputs, and outputs θ from a SoftMax driven K units. The encoder outputs

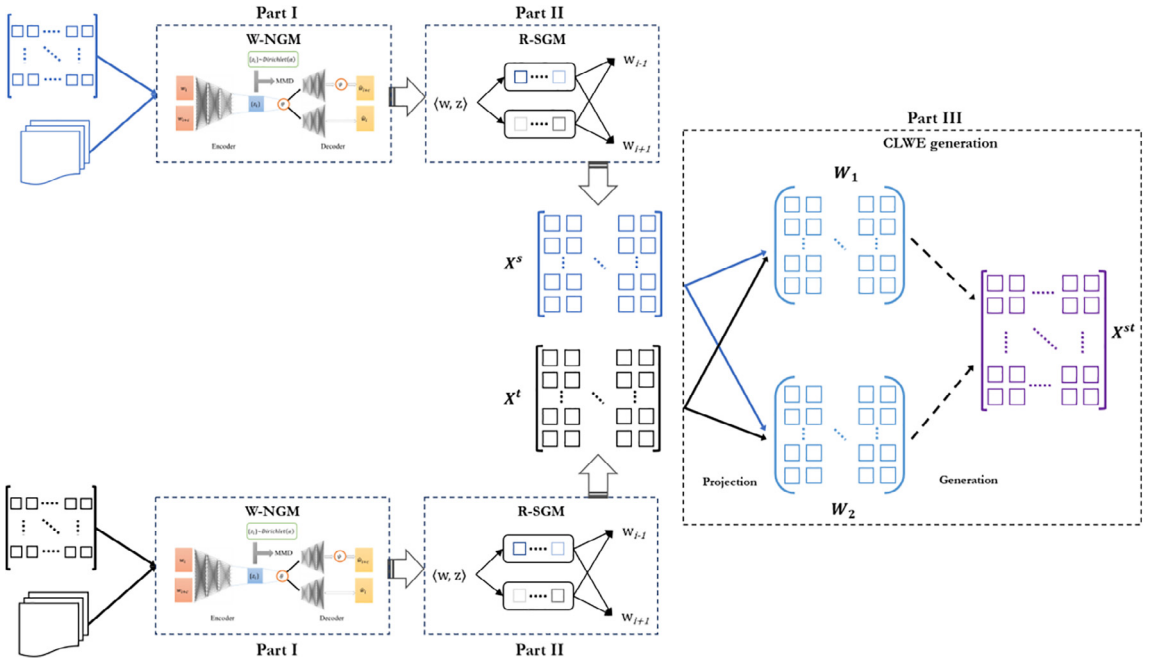

Fig. 1. The overall framework of our proposed EXLE method.

Table 1
Main notations used in the paper.

$\mathscr{C}$	a document corpus
$\mathscr{X}$	a pre-trained word embedding space
$\boldsymbol{W}$	a linear projection between two different semantic spaces
$\boldsymbol{s}$	a source language (usually used as superscript)
$\boldsymbol{t}$	a target language (usually used as superscript)
$\boldsymbol{v}/\boldsymbol{u}$	the vector representations of a word
$\boldsymbol{x}_i/\boldsymbol{x}_c$	the representations of a pivot word and a context word
V	total number of words
K	the number of topics
N	the number of pivot words
w	a word
c	a context word
ε	the context size of a word
z	a topical distribution index
θ	a multinomial distribution of topics
ψ	a topical distribution of words
τ	the dimension of the word embeddings
β	corpus-wide latent topics
γ	corpus-wide word embeddings
$b_{\boldsymbol{v}}/b_{\boldsymbol{u}}$	offset vectors
ξ	a balance factor
δ	the linear interpolation parameter of the R-SGM model
κ	the number of negative samples
λ	the linear interpolation parameter of the Wasserstein autoencoders
η	the linear interpolation parameter of the ICP model

$Q(\theta|\boldsymbol{v},\boldsymbol{u})$ to efficiently infer the posterior, i.e., $Q(\theta|\boldsymbol{v},\boldsymbol{u}) \approx p(\theta|\boldsymbol{v},\boldsymbol{u})$. We use a deterministic encoder $\theta = enc(\boldsymbol{v},\boldsymbol{u})$, which is conceptually and computationally simpler.

3.1.2. Decoder

The second component is the decoder corresponding to $p(\widehat{\boldsymbol{v}},\widehat{\boldsymbol{u}}|\theta)$. The decoder also consists of a neural network. It takes θ as the input, and outputs $\widehat{\boldsymbol{v}},\widehat{\boldsymbol{u}}$ from a SoftMax driven $2 \times V$ units. More specifically, we utilize the following two MLPs to generate the pivot words and the context words, respectively. It is obvious that the context words are produced independently.

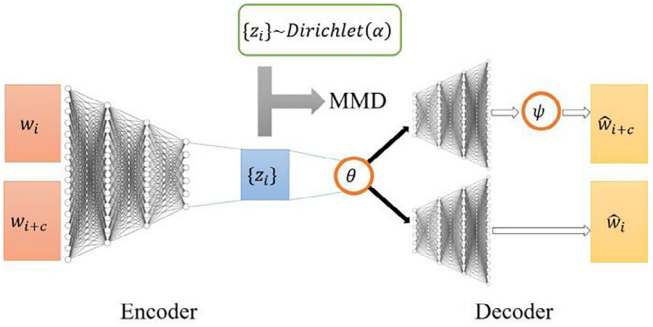

Fig. 2. The Wasserstein Autoencoders framework for the proposed W-NGM model.

$$p(\hat{\boldsymbol{v}}|\theta) \propto softmax(\boldsymbol{h})$$

$$\boldsymbol{h} = \beta\theta + b_{\boldsymbol{v}} \tag{1}$$

$$\psi \propto MLP(\theta)$$

$$p(\hat{\boldsymbol{u}}|\psi) \propto softmax(\boldsymbol{h}\prime)$$

$$\boldsymbol{h}\prime = \gamma\psi + b_{\boldsymbol{u}} \tag{2}$$

where $b_{\boldsymbol{v}}$ and $b_{\boldsymbol{u}}$ are offset vectors.

3.1.3. Wasserstein autoencoders

For the autoencoder, a negative cross-entropy loss between the input and output of the decoder is defined as the reconstruction loss:

$$cost((\boldsymbol{v},\boldsymbol{u}),(\hat{\boldsymbol{v}},\hat{\boldsymbol{u}})) = \left(-\sum_{i=1}^{V} \boldsymbol{v_i}^T \log \hat{\boldsymbol{v}}_{\boldsymbol{i}}\right) + \xi \cdot \left(-\sum_{i=1}^{N} \boldsymbol{u_i}^T \log \hat{\boldsymbol{u}}_{\boldsymbol{i}}\right) \tag{3}$$

where ξ serves as a balance factor. By the theory of the Wasserstein Autoencoders [42], minimizing the optimal transport distance between the output of the decoder and the target distribution for a scalar value λ is defined below, which is also the objective of our model:

$$\inf_{Q(\theta|\boldsymbol{v},\boldsymbol{u})} \mathbb{E}_{P_{\boldsymbol{v},\boldsymbol{u}}} \mathbb{E}_{Q(\theta|\boldsymbol{v},\boldsymbol{u})}[cost((\boldsymbol{v},\boldsymbol{u}),\mathrm{dec}(\theta))] + \lambda \cdot \mathscr{D}_{\Theta}(Q_{\Theta},P_{\Theta}) \tag{4}$$

where $cost(\,)$ is the cost function, $\mathrm{dec}(\,)$ is the output of the decoder (i.e. $(\hat{\boldsymbol{v}},\hat{\boldsymbol{u}})$). P_{Θ} and $Q_{\Theta} = \mathbb{E}_{P_{\boldsymbol{v},\boldsymbol{u}}} Q(\theta|\boldsymbol{v},\boldsymbol{u})$ are prior distribution and aggregated posterior, respectively. $\mathscr{D}_{\Theta}(Q_{\Theta},P_{\Theta})$ is an arbitrary divergence between Q_{Θ} and P_{Θ}. An important feature of the Wasserstein Autoencoders is that the aggregated posterior is used as the regularization term. We use the maximum mean discrepancy (MMD) [16] based divergence in our implementation to set $MMD_f(Q_{\Theta},P_{\Theta})$, where $f : \Theta \times \Theta \to \mathbb{R}$ denotes a kernel function. $MMD_f(Q_{\Theta},P_{\Theta})$ can be calculated as:

$$MMD_f(Q_{\Theta},P_{\Theta}) = \int_{\Theta} f(\theta), dP_{\Theta}(\theta) - \int_{\Theta} f(\theta), dQ_{\Theta}(\theta)_{\mathscr{H}_f} \tag{5}$$

where $\mathscr{H}_f$ is the reproducing kernel Hilbert space of the real-valued functions mapping Θ to $\mathbb{R}$. $f(\theta,\cdot)$ can be considered as transferring θ to a higher dimensional space. The information diffusion kernel [23] is utilized in our implementation for the Dirichlet distribution.

3.2. Part II: A rectified Skip-Gram model R-SGM

The skip-gram model does not contain non-linear hidden layers opposite to the neural language model. The objective of predicting the context words w_c given a pivot word w_i is defined as:

$$L_w(\mathscr{C}) = \frac{1}{M}\sum_{i=1}^{M} \sum_{-\varepsilon \le c \le \varepsilon, c \ne 0} \log p(w_{i+c}|w_i) \tag{6}$$

The probability of $p(w_c|w_i)$ can be calculated by SoftMax:

$$p(w_c|w_i) = \frac{e^{\boldsymbol{x}_c \cdot \boldsymbol{x}_i}}{\sum_{c' \in \mathscr{C}} e^{\boldsymbol{x}_{c'} \cdot \boldsymbol{x}_i}} \tag{7}$$

$\boldsymbol{x}_i$ and $\boldsymbol{x}_c \in \mathbb{R}^\tau$ denote the vectors of w_i and w_c, respectively. The skip-gram model conducts the training process iteratively until high-quality word embeddings are obtained: 1) all words in a corpus are taken in turn as the pivot word; 2) the corresponding context words of the pivot word are generated; 3) pairs of pivot and context words are composed as the training set; 4) two methods (hierarchical SoftMax or negative sampling) are adopted for training the model until completion [31]. We are more interested in the negative sampling method and use it to develop our own model. The loss of the method is defined as:

$$L_w^{neg}(\mathscr{C}) = \sum_{pos} \log\sigma(\boldsymbol{x}_c \cdot \boldsymbol{x}_i) + \sum_{neg} \log\sigma(-\boldsymbol{x}_c \cdot \boldsymbol{x}_i) \tag{8}$$

We aim to train word embeddings and topics separately and simultaneously in our R-SGM model. In addition, we own updated word embeddings generated from the W-NGM Model. We integrate word embeddings ($L_w(\mathscr{C})$), W-NGM-enhanced topics ($L_z(\mathscr{C})$), and W-NGM-enhanced word embeddings ($L_{\widetilde{w}}(\mathscr{C})$) all together to create high-quality topical word embeddings in each language. We define the loss of the R-SGM as the follows:

$$L_{wz\widetilde{w}}(\mathscr{C}) = (1-\delta)L_w(\mathscr{C}) + \frac{\delta}{2}L_z(\mathscr{C}) + \frac{\delta}{2}L_{\widetilde{w}}(\mathscr{C}) \tag{9}$$

where δ is a free parameter to balance the weights between different components. $L_z(\mathscr{C})$ is defined by using the topic vectors instead of the word vectors:

$$L_z(\mathscr{C}) = \frac{1}{K}\sum_{i=1}^{K} \sum_{-\varepsilon \le c \le \varepsilon, c \ne 0} \log p(z_{i+c}|z_i) \tag{10}$$

$L_{\widetilde{w}}(\mathscr{C})$ has the same form as the $L_w(\mathscr{C})$. The difference is that the lookup embeddings are not initialized randomly but use the enhanced embeddings from part I.

Unlike the topical word embeddings proposed by Liu et al. [30], in our model, pivot topics are used to predict the context topics rather than the context terms. At the same time, we also assume that each topic will appear in all locations of the related words. Obviously, our R-SGM model will make further use of topics to improve the performance of word embeddings while learning topical word embeddings. In their work, word embeddings are always unaltered. We believe that the mutual reinforcement process is in-line with the requirement of combination. Strengthening one model while keeping the other unchanged will ignore some important information, so that the topic-based vectors cannot represent the assembled semantics of words and topics. Our experimental results confirm this assumption. Finally, their model concatenates word and topic embeddings to generate a long embedding for output. Our R-SGM model does not adopt this approach. The results in experiments also prove that the word embeddings produced by joint learning is more efficient.

In view of the high computational complexity of the original skip-gram model, we propose a method to randomly pick κ negative samples for every pivot-context pair [15] to computer the joint loss. These negative samples are combined with the original labels to form $\kappa^{neg} + 1$ labels, and the following probabilities are calculated for these training pairs:

$$p\prime(w_c|w_i) = \frac{e^{\boldsymbol{x}_c \cdot \boldsymbol{x}_i}}{e^{\boldsymbol{x}_c \cdot \boldsymbol{x}_i} + \sum_{c' \in \kappa^{neg}} e^{\boldsymbol{x}_{c'} \cdot \boldsymbol{x}_i}} \tag{11}$$

This method can be regarded as a negative SoftMax loss method. By directly calculating SoftMax and maximizing the probability of the cross-entropy loss, the training speed of our R-SGM model can be significantly accelerated on the basis of guaranteeing the performance of generation.

3.3. Part III: Cross-Lingual word embeddings generation

We now describe the projection methods used to obtain the final bilingual topical word embeddings based on the monolingual embeddings acquired in the last part.

We define the pre-trained monolingual word embedding spaces obtained from the last part to be $\mathscr{X}^{\boldsymbol{s}}$ and $\mathscr{X}^{\boldsymbol{t}}$ of $\tau \times U$ matrices. U is defined to be the vocabulary size of either language in processing. They form dictionaries containing word pairs $\left\{w_i^s, w_j^t\right\}_{i,j=1}^{U}$. Our method aims to minimize the Euclidean distance between the source and the target space for alignment, so that the resulting projection will make the corresponding translation words between the two languages close together in the shared space. In particular, we define the real number quadratic matrix space $\boldsymbol{W}$ (i.e., the mapping) as the follows:

$$\boldsymbol{W}^* = argmin_{\boldsymbol{w}} \|\boldsymbol{W}\mathscr{X}^{\boldsymbol{s}} - \mathscr{X}^{\boldsymbol{t}}\|_F \tag{12}$$

By restricting $\boldsymbol{W}$ to an orthogonal matrix, the alignment performance can be greatly improved. At the same time, the above optimization problem can be transformed into the Procrustes problem [36], whose closed form solution can be acquired by singular value decomposition (SVD) [47] of $\mathscr{X}^{\boldsymbol{t}}\mathscr{X}^{\boldsymbol{s}\boldsymbol{T}}$:

$\boldsymbol{W}^{*} = argmin_{\boldsymbol{w}}\|\boldsymbol{W}\mathscr{X}^{\boldsymbol{s}} - \mathscr{X}^{\boldsymbol{t}}\|_{F} = \boldsymbol{U}\boldsymbol{V}^{\boldsymbol{T}}$, subject to.$\boldsymbol{W}\boldsymbol{W}^{\boldsymbol{T}} = \boldsymbol{I}$

$$\boldsymbol{U}\Sigma\boldsymbol{V}^{\boldsymbol{T}} = \boldsymbol{SVD}\left(\mathscr{X}^{\boldsymbol{t}}\mathscr{X}^{\boldsymbol{s}\boldsymbol{T}}\right) \tag{13}$$

The alignment method mentioned above can be considered *unsupervised* if a readily available bilingual dictionary is not used, or *supervised* otherwise [26]. In some work, two projection matrices (source to target language or vice versa) are utilized. Here we employ a simplified strategy by using only one projection matrix $\boldsymbol{W}$. The strategy can also achieve satisfied results. In terms of the concrete *supervised* method used in our CLWEs generation, we adopt the model proposed by Smith et al. [39], and use the off-the-shelf dictionaries for training. Although it is possible to use the dictionaries of different sizes, it is mainly designed for the resource-lean settings [27] and are clearly beyond the scope of this article. We save it in the follow-up work.

If bilingual dictionaries are not ready, they can be automatically constructed by self-learning in an unsupervised method. We opt for Hoshen and Wolf's method [19] termed the iterative closet point (ICP) model, because it produces better unsupervised CLWEs than others [14]. Unlike the supervised model, the ICP model seeks two projection matrices $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ instead of one to minimize the sum of distances between two embeddings. Firstly, PCA is used to project the most commonly used words in source and target languages into a low-dimensional space to induce a seed dictionary. Then the model fixes the matrices $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ to search the optimal alignment. Finally, the alignment will be used to update the projection matrices. More formally, the model tries to minimize:

$$\sum_{\boldsymbol{i}}\|x_i^{\boldsymbol{s}}\boldsymbol{W}_1 - x_{ind(i)}^{\boldsymbol{t}}\| + \sum_j\|x_j^{\boldsymbol{t}}\boldsymbol{W}_2 - x_{ind(j)}^{\boldsymbol{s}}\| + \eta\sum_i\|x_i^{\boldsymbol{s}} - x_i^{\boldsymbol{s}}\boldsymbol{W}_1\boldsymbol{W}_2\| + \eta\sum_j\|x_j^{\boldsymbol{t}} - x_j^{\boldsymbol{t}}\boldsymbol{W}_2\boldsymbol{W}_1\| \tag{14}$$

where $ind()$ refers to the word index. $x_i^{\boldsymbol{s}}$ and $x_j^{\boldsymbol{t}}$ denote the word vectors from the source and target languages, respectively. η is usually set to 0.1. During the training process of this *unsupervised* method, we utilize an iterative bootstrapping strategy as in the supervised counterparts to further improve the projection.

4. Experiments

We organize a series of experiments to evaluate the performance of the proposed EXLE method in this article. As proved by previous studies [14,27], downstream tasks are more suitable for evaluating cross-lingual word embeddings than NLP-focused tasks. We first evaluate our EXLE method for CLIR as a whole, comparing it to the state-of-the-art semantic CLWE methods that use aligned monolingual static and contextual word embeddings for neural CLIR, to check the effectiveness of our method. Then we evaluate our CLWE training procedure to check if it produces better CLIR results than using the representative monolingual topical word embeddings. Furthermore, we compare our method with the CLIR method which uses a neural machine translation system. Lastly, we run an intrinsic evaluation through a word similarity task to investigate if the results align with those of the extrinsic evaluation for further investigation.

4.1. Evaluation setup

4.1.1. Datasets and tasks

All experiments selected language pairs with different degrees of similarities. These include English (EN) with German (DE), Spanish (ES), and Finnish (FI), respectively. The most similar language pairs are EN and DE, both are Germanic languages. FI is non-Indo-European (Uralic language) and least similar to English. ES is a romance language at the middle level. In terms of monolingual and cross-lingual word vectors generation, the method of Smith et al. [39] was chosen for supervised training, while the method of Hoshen and Wolf [19] was used for unsupervised training of CLWEs. Both approaches are competitive, have wide coverage, and are easily available[2]. Training data for all the above languages were obtained from the document-aligned Wikipedia corpora[3]. More information and technical details can be acquired from the original papers.

All our CLIR experiments were performed on the CLEF 2001–2003 test collections from the ad-hoc retrieval task. The target document collections in DE, ES, and FI were selected and paired with their correspondent EN queries. We provide statistics for the complete experimental setup in Table 2. As a standard practice, queries were concatenated from the *title* and the *description* fields of each topic provided by CLEF. All texts in documents as well queries were converted to lowercase. Punctuations, stop-words, and noise words were discarded.

4.1.2. Unsupervised neural CLIR methods

The unsupervised neural CLIR retrieval methods used in our experiments follow previous work on using neural vectors for monolingual and cross-lingual information retrieval [26,27,44]. We chose two representative methods provided in previous

[2] https://github.com/codogogo/xling-eval.
[3] https://linguatools.org/tools/corpora/wikipedia-comparable-corpora/.

Table 2
Statistics of the used CLEF test collections.

Language	Year	Name	No. of documents	Size (MB)
German	2001–2002	Frankfurter Rundschau 94 Der Spiegel 94–95 SDA German 94	225,371	527
	2003	Frankfurter Rundschau 94 Der Spiegel 94–95 SDA German 94–95	294,809	672
Spanish	2001–2002	EFE 94	215,738	509
	2003	EFE 94–95	454,045	1105
Finnish	2002–2003	Aamulehti 94–95	55,344	137

papers for retrieving documents. The first method (denoted as IDF) uses IDF-weighted sums of embeddings for representing queries and documents and employs the cosine similarity as the ranking function. The second method (denoted as TbT) utilizes word-by-word query translations with the help of the shared CLWEs. It reduces the cross-lingual retrieval task to a monolingual retrieval task by using a standard query likelihood language model for retrieving documents. More information and technical details can be acquired from the original papers.

4.1.3. Baseline methods and evaluation metrics

We compare our EXLE method to a number of state-of-the-art semantic CLWE methods that use aligned monolingual static and contextual word embeddings for the unsupervised neural CLIR.

The first set of baseline methods uses the monolingual word embeddings together with: the supervised projection model for learning cross-lingual word embeddings as in [39] (WE-SUP), and the unsupervised projection model for learning CLWEs as in [19] (WE-UNSUP). In both baseline methods, the monolingual word embeddings were trained by the original skip-gram model as in the word2vec tool instead of more complex methods like fastText [2] or Glove [33].

It has been shown that monolingual topical word embeddings previously proposed can improve the tasks like word similarity and text classification [30,38]. However, their effectiveness in information retrieval is not yet explored. In this set of baseline methods, we replace the word embeddings produced by the skip-gram model with the topical word embeddings in [30] together with the supervised (TWE-SUP) and the unsupervised (TWE-UNSUP) learning method for constructing the cross-lingual spaces and use for CLIR.

In the next set of baseline methods, we use the supervised (TWE2-SUP) and the unsupervised (TWE2-UNSUP) method together with another representative monolingual topical word embeddings [38] for CLWEs learning and CLIR. Note that in Shi et al.'s work [38], they also jointly trained the topic models and word embeddings. However, they adopted a different approach and produced very different embeddings to us.

For contextual alignment, we include the rotation-based BERT (RBERT-SUP) as described in [6]. It is proved more effective than aligning the "average sense" of a word [37], and it is essentially a supervised method. Note that all alignment methods evaluated in this paper are rotation-based. Fine-tune approaches are excluded from the experiments.

In our method, we use W-NGM + R-SGM together with the supervised (EXLE-SUP) and unsupervised (EXLE-UNSUP) projection models to learn CLWEs. Then we use it for CLIR as in the baseline methods.

To demonstrate the effectiveness of our method in a more general setting, we further compare our method to two representative baselines. First is a sanity check baseline: a monolingual language model-based method (LM). Second is a much stronger baseline: translates queries using a state-of-the-art neural machine translation toolkit [22], then performs a monolingual retrieval (NMT). This approach demonstrates better performance than using the off-the-shelf translation tools like Google Translator [27,28]. It requires large sentence-aligned corpora, while our method just needs document-aligned corpora.

In terms of evaluation metrics, we use the mean average precision (MAP) and the normalized discount cumulative gain (NDCG@10) in the main results. All results are calculated based on the average performance of the test topics. At a 95% confidence level, we utilize a two-tailed paired *t-test* to determine that the differences are statistically significant (i.e., $p < 0.05$).

4.1.4. Parameter settings

We need to set some parameters for the different models mentioned in this article. Most of the parameters were determined by extensive sensitivity experiments, and a small number were selected based on previous experience. We employed separate data to train the parameters and ensure that there was no overlap between this portion and the portion we used for the main experiments.

For the parameters in the W-NGM model, the number of latent topics K is experimentally set to 150 in all language pairs, indicating that a moderate number of topics is preferred. While it is theoretically possible to automatically tune the number of topics, we chose a fixed number for simplicity. In the autoencoder, the balance factor ξ is set to 0.5. λ is empirically set to 10. Similarly, the Dirichlet parameter in MMD is set to 0.1. To overcome the initial local minimum, the ADAM optimizer is set to a high momentum value of 0.99 with 0.002 for the learning rate. As a common practice, the dimension size τ is set to 300 in all word embeddings.

For the parameters in the R-SGM model, words and topics are considered to be equally important when training word embeddings. Therefore, the free parameter δ is set to 0.5. The number of negative samples and context window size are set to 16 and 10, respectively, the same setting as in the W-NGM model. All models, including the baseline methods, were re-implemented in Python using TensorFlow. We used only a single GPU for training. The mini-batch size and the dropout value are set to 128 and 0.2, respectively. It is worth pointing out that we set the parameters in the baseline methods based on the tuning procedures in their published papers to ensure the best performance. More technical details can be acquired from the original papers.

4.2. Results and Discussion

4.2.1. Main results

We present the performance for all methods in comparison on the CLEF 2002–2003 test collections in Tables 3 & 4. As no Finish data in the CLEF 2001 test collection, we report them in the next subsection. Note that the arbitrary value of the metrics (such as MAP) may seem low. However, the performance obtained in our paper is consistent with those reported in the original CLEF campaign[4] and various CLIR papers [26,27]. Note that we are more interested in comparing the EXLE with the state-of-the-art baseline methods implemented in the current article. The improvement of overall accuracy is beyond the scope of consideration here and will be further studied in the future.

Comparison to CLWEs We firstly focus on the results in Tables 3 & 4. The scores in the upper half of the table (above the dashed line in each CLIR method) correspond to the supervised models, whereas the scores in the lower half respond to unsupervised models. The statistically significant differences are marked as ○, ◇ and ^ with respect to the WE, TWE2, and RBERT baselines (the TWE model often shows worse performance than the WE model, so we do not include it for statistical significance computation). In addition, in some cases, our method works better than the NMT baseline. We mark the statistical difference with the NMT baseline as ▲. Improvements of our method (A) over another baseline method (B) are calculated by taking the score for method A minus the score for method B, and dividing it by the score for method B to get the percentage.

Our EXLE method works well in all CLIR evaluations, producing better than other baselines that use cross-lingual word embeddings measured by all evaluation metrics in all test collections with significant improvements observed in most cases. This finding holds for both supervised and unsupervised projection-based methods. When measured with the MAP, the greatest improvement reaches 57.50% (EN-FI, TbT, CLEF-2002) and 48.13% (EN-FI, IDF, CLEF-2003) compared to the supervised methods using the monolingual word embeddings trained by the original skip-gram model. 73.95% (EN-FI, TbT, CLEF-2002) and 33.85% (EN-FI, TbT, CLEF-2003) comparing to the unsupervised alternatives. This shows that our proposed EXLE method can capture bilingual semantic information of texts more accurately than the most advanced cross-language word embedding generation methods for CLIR. The reason lies in the different mechanisms of leveraging topics into word embeddings. We learn neural topic-enhanced cross-lingual word embeddings based on simultaneously model topic relevance and word embeddings. Furthermore, our method not only avoids the inherent noises introduced by the naïve word2vec methods, but also includes a unified mutually reinforcing framework for joint training of topic models and word embeddings.

Comparison to topical CLWEs We now focus on comparing our proposed model with the models that use classical monolingual topical word embeddings. Firstly, we observe mixture results obtained using Liu et al.'s TWE model and Shi et al.'s joint model. For the TWE model, we can only observe the CLIR methods work better than the WE baseline in very few cases, for example, in EN-DE (2002) when using TbT for CLIR. In other cases, a substantial decrease in performance is frequently detected. This result indicates that concatenating word embeddings and topical embeddings may not be a wise option for generating good topical word embeddings. Furthermore, it also shows that the joint loss function used by the TWE model cannot catch the interconnections between topics and word embeddings very well. On the contrary, we can observe that methods using Shi et al's joint training model (TWE2) in many language pairs outperform simpler monolingual word embeddings with no consideration of topical information. This confirms the validity of jointly training topics and word embeddings.

Pleasingly, again our method outperforms the TWE and TWE2 in every case, with statistically significant results. For example, our model gains an improvement of up to 122.09% (EN-FI, TbT, UNSUP, CLEF-2002) and 79.82% (EN-FI, TbT, UNSUP, CLEF-2003) over the TWE baseline methods; 37.43% (EN-FI, TbT, UNSUP, CLEF-2002) and 29.47% (EN-DE, TbT, UNSUP, CLEF-2003) over the TWE2 baseline methods when measured with the MAP. Our EXLE method beats simpler monolingual TWE models. It tackles two main challenges in merging topic models and word embeddings: How to train good topic models by simultaneously updating word embeddings and how to train good word embeddings by correctly leveraging topical information.

Comparison to NMT The methods based on CLWEs perform better than the LM baseline except for very few cases between less similar languages. This is reasonable as the vocabulary between the specific language pairs is limited. In most runs, the CLWE-based methods underperform the NMT baseline. However, it is very encouraging to find out that our EXLE method beats it in several cases (EN-DE and EN-FI in CLEF 2002, EN-DE, EN-ES, and EN-FI in CLEF 2003, measured in the

[4] https://www.clef-initiative.eu/.

Table 3
CLIR performance of various methods in CLEF 2002 test collection. Bold numbers indicate the best scores in the group excluding NMT. NMT Results inferior to our methods are underlined. Statistically significant differences between our methods and WE, TWE2, RBERT and NMT are indicated by ○, ◇, ^ and ▲ respectively. Improvements of our methods over the baseline methods are shown in percentage for MAP.

CLIR Methods	CLWE Methods	EN-DE			EN-ES			EN-FI		
		MAP	Improv.	NDCG	MAP	Improv.	NDCG	MAP	Improv.	NDCG
LM		0.1678	–	0.2340	0.1290	–	0.2027	0.1064	–	0.1676
NMT		0.2962	–	0.4131	0.4128	–	0.6486	0.2331	–	0.3672
IDF	WE-SUP	0.1994	*24.19%*	0.3774	0.1678	*11.52%*	0.3000	0.0962	*26.89%*	0.1407
	TWE-SUP	0.1702	*45.57%*	0.3412	0.1368	*36.76%*	0.2526	0.0967	*26.26%*	0.1976
	TWE2-SUP	0.2302	*7.62%*	0.3980	0.1723	*8.56%*	0.2971	0.1025	*19.16%*	0.2184
	RBERT-SUP	0.2348	*5.51%*	0.4060	0.1758	*6.44%*	0.3031	0.1086	*12.42%*	0.2315
	EXLE-SUP	**0.2477**○◇^	–	**0.4144**○◇^	**0.1871**○◇^	–	**0.3243**○◇^	**0.1221**○◇^	–	**0.2620**○◇^
	WE-UNSUP	0.2162	*7.92%*	0.3851	0.1533	*20.49%*	0.2859	0.1063	*6.49%*	0.2244
	TWE-UNSUP	0.1478	*57.84%*	0.3128	0.1241	*48.90%*	0.2238	0.1063	*6.49%*	0.1947
	TWE2-UNSUP	0.2201	*6.03%*	0.3930	0.1577	*17.15%*	0.2877	0.1090	*3.91%*	0.**2624**
	EXLE-UNSUP	**0.2334**○◇	–	**0.4121**○◇	**0.1847**○◇	–	**0.3263**○◇	**0.1132**○◇	–	**0.**2572○
TbT	WE-SUP	0.2333	*31.27%*	0.3655	0.3046	*23.02%*	0.4761	0.1695	*57.50%*	0.2279
	TWE-SUP	0.2565	*19.40%*	0.3734	0.2929	*27.96%*	0.4764	0.1873	*42.49%*	0.2916
	TWE2-SUP	0.2556	*19.85%*	0.3844	0.3298	*13.62%*	0.5258	0.1980	*34.83%*	0.3226
	RBERT-SUP	0.2658	*15.24%*	0.3997	0.3430	*9.25%*	0.5468	0.2138	*24.84%*	0.3485
	EXLE-SUP	**0.3063**○◇^▲	–	**0.4247**○◇^▲	**0.3748**○◇^	–	**0.5701**○◇^	**0.2669**○◇^▲	–	**0.3704**○◇^▲
	WE-UNSUP	0.1932	*57.08%*	0.3258	0.2772	*37.49%*	0.4533	0.1087	*73.95%*	0.1258
	TWE-UNSUP	0.1953	*55.40%*	0.3286	0.2562	*48.76%*	0.3989	0.0851	*122.09%*	0.1465
	TWE2-UNSUP	0.2329	*30.30%*	0.3608	0.3254	*17.14%*	0.5132	0.1376	*37.43%*	0.2516
	EXLE-UNSUP	**0.3035**○◇▲	–	**0.4076**○◇	**0.3811**○◇	–	**0.6033**○◇	**0.1890**○◇	–	**0.2885**○◇

Table 4
CLIR performance of various methods in CLEF 2003 test collection. Bold numbers indicate the best scores in the group excluding NMT. NMT Results inferior to our methods are underlined. Statistically significant differences between our methods and WE, TWE2, RBERT and NMT are indicated by ○, ◇, ^ and ▲ respectively. Improvements of our methods over the baseline methods are shown in percentage for MAP.

CLIR Methods	CLWE Methods	EN-DE			EN-ES			EN-FI		
		MAP	Improv.	NDCG	MAP	Improv.	NDCG	MAP	Improv.	NDCG
LM		0.1428	–	0.2163	0.1026	–	0.1478	0.1423	–	0.2258
NMT		0.2639	–	0.3997	0.3343	–	0.4816	0.2765	–	0.4387
IDF	WE-SUP	0.2427	*15.60%*	0.3629	0.1958	*14.64%*	0.3477	0.2181	*48.13%*	0.3089
	TWE-SUP	0.1863	*50.60%*	0.2709	0.1665	*34.85%*	0.2574	0.2734	*18.14%*	0.3986
	TWE2-SUP	0.2469	*13.64%*	0.3827	0.2023	*10.97%*	0.3306	0.2950	*9.50%*	0.3916
	RBERT-SUP	0.2494	*12.51%*	0.3865	0.2104	*6.70%*	0.3438	0.3009	*7.35%*	0.3994
	EXLE-SUP	**0.2806**○◇^▲	–	**0.4124**○◇^▲	**0.2245**○◇^	–	**0.3881**○◇^	**0.3230**○◇^▲	–	**0.4663**○◇^▲
	WE-UNSUP	0.2595	*11.69%*	0.3915	0.1944	*13.20%*	0.3426	0.2146	*17.11%*	0.3168
	TWE-UNSUP	0.2271	*27.63%*	0.3169	0.1348	*63.23%*	0.2372	0.1813	*38.63%*	0.2552
	TWE2-UNSUP	0.2784	*4.11%*	0.4046	0.1958	*12.37%*	0.3234	0.2220	*13.23%*	0.2913
	EXLE-UNSUP	**0.2898**○◇▲	–	**0.4316**○◇▲	**0.2201**○◇	–	**0.3620**○◇	**0.2513**○◇	–	**0.3503**○◇
TbT	WE-SUP	0.2239	*37.11%*	0.3038	0.2775	*21.99%*	0.3924	0.1982	*33.99%*	0.2426
	TWE-SUP	0.2184	*40.54%*	0.2918	0.2585	*30.98%*	0.3865	0.2052	*29.46%*	0.2858
	TWE2-SUP	0.2472	*24.18%*	0.3514	0.2870	*17.97%*	0.4214	0.2127	*24.86%*	0.3104
	RBERT-SUP	0.2496	*22.95%*	0.3549	0.2984	*13.43%*	0.4383	0.2234	*18.92%*	0.3259
	EXLE-SUP	**0.3069**○◇^▲	–	**0.4272**○◇^▲	**0.3385**○◇^▲	–	**0.4870**○◇^▲	**0.2656**○◇^	–	**0.3517**○◇^
	WE-UNSUP	0.2505	*17.25%*	0.3560	0.2754	*23.03%*	0.3629	0.1676	*33.85%*	0.2108
	TWE-UNSUP	0.1791	*63.98%*	0.2712	0.2690	*25.98%*	0.3860	0.1247	*79.82%*	0.1453
	TWE2-UNSUP	0.2606	*12.71%*	0.3478	0.2864	*18.32%*	0.4028	0.1733	*29.47%*	0.2108
	EXLE-UNSUP	**0.2937**○◇▲	–	**0.4200**○◇▲	**0.3388**○◇▲	–	**0.4485**○◇	**0.2243**○◇	–	**0.2619**○◇

MAP). It has been shown that CLWE-methods can outperform MT in sentence retrieval in some cases [27]. We show that in this work, with the help of neural topics, they perform equally well in document retrieval.

Comparison to Contextual Embeddings It shows that in general, the R_BERT works better than the topic-enriched CLWEs but worse than the neural topic-enhanced CLWEs produced by our EXLE method in different CLIR retrieval methods

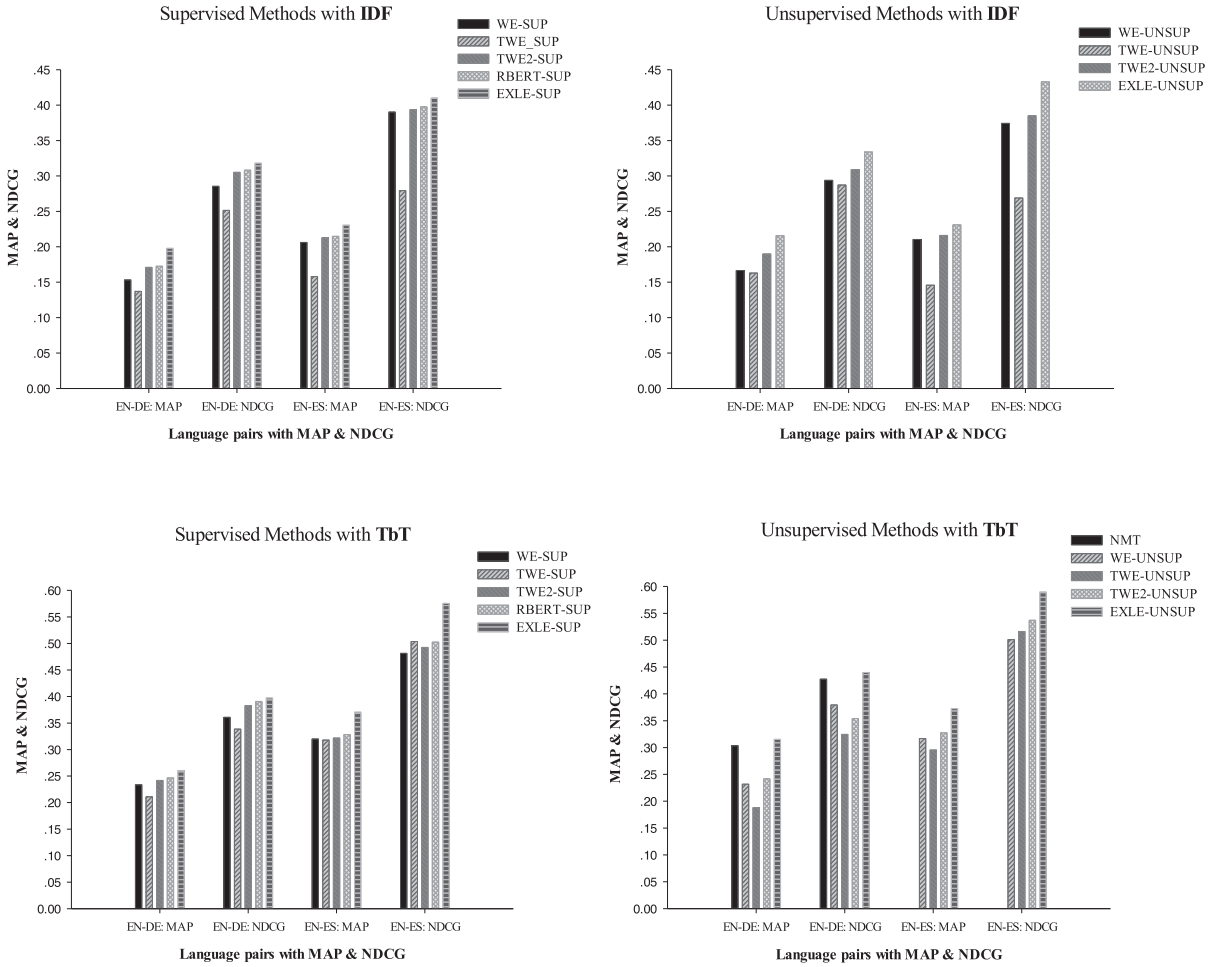

Fig. 3. Cssxdc LIR performance of various methods on the CLEF 2001 test collection. LM results are omitted due to their consistently lowest performance. NMT results are shown only when our proposed methods surpass them (EN-DE, TbT, UNSUP).

with an improvement of up to 14.58% (EN-DE, IDF, SUP, CLEF-2001), 24.84% (EN-FI, TbT, UNSUP, CLEF-2002), 22.95% (EN-DE, TbT, UNSUP, CLEF-2003). This result supports the intuition that aligning contextual models is more complicated than aligning static word vectors. Sub-par alignments are produced by rotation-based methods, as reported in [6]. Fine-tune may be a better solution, and should be investigated further.

Finally, the above results not only confirm the effectiveness of our EXLE method for unsupervised neural CLIR, but also prove its robustness through different experimental permutations. They also verify the reasonability of using a three-part method to model topic relevance and word embeddings in a mutual reinforcement way, which in turn produces high-quality CLWEs. In addition, using the CLWEs produced by our proposed method for the unsupervised CLIR consistently achieves better accuracy than the CLWEs generated by the classical embedding-only methods and topical embedding methods. In all runs, the best results can be obtained, without exception.

4.2.2. Results on the CLEF 2001 Test Collection

We present the performance of all methods in comparison on the CLEF 2001 test collection in Fig. 3 for a better illustration and comparison. Like the results reported in other CLEF test collections, our EXLE method works well in all CLIR evaluations, producing better than other baselines using cross-lingual word embeddings measured by all evaluation metrics in all test collections with significant results improvements are observed in most of the cases. This finding holds for both supervised and unsupervised projection-based methods. When measured with the MAP, the greatest improvement reaches 29.02% (EN-DE, IDF, CLEF-2001, $MAP = .1978$ for EXLE-SUP) compared to the supervised methods using the monolingual word embeddings trained by the original skip-gram model ($MAP = .1533$ for WE-SUP); 29.74% (EN-DE, IDF, CLEF-2001, $MAP = .2159$ for EXLE-SUP) compared to the unsupervised alternatives ($MAP = .1664$ for WE-UNSUP). This shows that our proposed EXLE method can capture the bilingual semantic information of texts more accurately than the most advanced cross-language word embedding generation methods for CLIR.

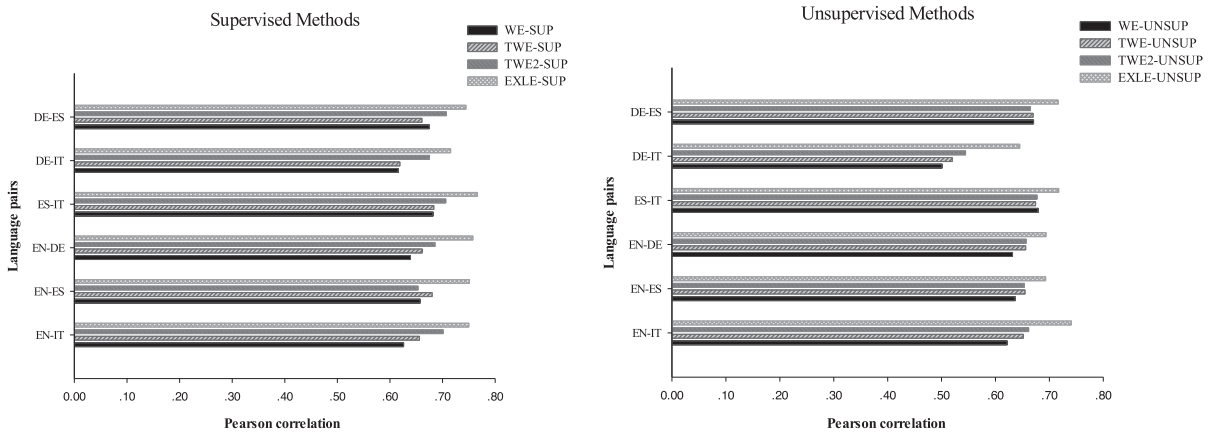

Fig. 4. Results of the intrinsic evaluation.

Our method outperforms the TWE and TWE2 in every case, with statistically significant results. Our method improves up to 58.52% (EN-ES, IDF, UNSUP, CLEF-2001, *MAP* = .2310 for EXLE-SUP) over the TWE baseline methods, and 21.38% (EN-DE, TbT, UNSUP, CLEF-2001, *MAP* = .1457 for TWE-UNSUP) over the TWE2 baseline methods when measured with the MAP. In this set of results, the R_BERT works better than the topic-enhanced CLWEs but worse than the neural topic-enhanced CLWEs produced by our EXLE method with different CLIR retrieval methods with an improvement of up to 14.58% (EN-DE, IDF, SUP, CLEF-2001, *MAP* = .1978 for EXLE-SUP & *MAP* = .1726 for RBERT-SUP).

All the results reported on the CLEF 2001 test collection align with our results on the CLEF 2002–2003 test collections. Thus, they confirm the robustness of our proposed EXLE method for CLIR.

4.2.3. Supervised vs. Unsupervised methods

Next, we consider the performance of different supervised and unsupervised models that learning CLWEs. From Tables 3 & 4 and Fig. 3 we can learn that the results are mixed. Generally speaking, in most cases, the supervised methods perform better than the unsupervised methods in the CLEF 2002–2003 test collections. However, for results on the CLEF 2001 test collection, the unsupervised methods display better performance than the supervised methods. Especially for our proposed method, the unsupervised version is always better in the CLEF 2001 test collection. We do not experiment with different sizes of dictionaries as in [14] because it is not the main focus of the current paper. It is shown that a CLIR system may be built without any cross-lingual information needed, but a reliable CLIR system is better constructed with additional cross-lingual information.

4.2.4. Different language pairs

We now examine the evaluation performance by using different language pairs as queries and documents. Most methods achieve better performance for the most similar language pair EN-DE. For all three test collections, our method performs better than the NMT baseline. It can be seen that the results obtained in EN-FI (a least similar language pair) are pretty impressive. Our method beats the NMT baseline in two of three test collections. At the same time, other methods can also obtain reasonable results. However, for the middle similar language pair EN-ES the results are less impressive. Only in one test collection (CLEF 2003) the performance seems better. As pointed out by [40], English, German, and Spanish are dependent-marking, but Finnish is mixed-marking. Further investigation is very much needed to determine whether it originates from the differences between languages or from the nature of test collections.

4.2.5. CLIR methods

We compare the two CLIR retrieval methods (IDF and TbT) using cross-lingual word embeddings. Again, the results are mixed. In EN-DE and EN-ES, the TbT method outperforms the IDF method in a large margin, whereas in EN-FI, we observe that the IDF method outperforms the TbT method. We notice that these results are slightly different from previous findings [27]. This may be due to the different language pairs used and pre-processing procedures we employ.

4.2.6. Intrinsic evaluation

Previous studies have shown that the results of extrinsic evaluation sometimes do not match those of intrinsic evaluation, such as in similarity tasks [14]. Therefore, it is necessary to verify further whether our proposed method can also achieve good results in the intrinsic evaluation. In this set of evaluations, cross-linguistic semantic word similarity is a common task, which reports the degree of correlation between the cosine similarity of words in two different languages and a score labeled by a human. We adopt this task and compare it with the baseline methods. In the experiment, we use the data provided by SemEval 2017 competition [5]. We employ Pearson correlation as the evaluation metric.

Fig. 4 shows that our EXLE method achieves robust improvements over the baseline methods in all six language pairs. In the supervised CLWEs, compared with the WE-based methods (EN-IT, *Pearson* = .7498 for EXLE-SUP & *Pearson* = .6259 for WE-SUP) and other TWE-based methods (EN-ES, *Pearson* = .7512 for EXLE-SUP & *Pearson* = .6534 for TWE2-SUP), we get 19.8% and 14.96% improvements respectively. In the unsupervised CLWEs, we obtained 28.75% (DE-IT, Pearson = 0.6450 for EXLE-SUP & Pearson = 0.5010 for WE-SUP) and 18.44% (DE-IT, Pearson = 0.6450 for EXLE-SUP & Pearson = 0.5446 for TWE2-SUP) improvements respectively by the same comparison.

The performance of the unsupervised methods has decreased compared with the supervised methods, which is consistent with the results of the extrinsic evaluation. It means that it is impossible to determine which mapping method is the best for generating CLWEs. From another point of view, this article mainly studies the mutual reinforcement of topic models and word embeddings, rather than trying to find better supervised or unsupervised projection methods. We will leave it in the future work.

5. Conclusion

We introduce a novel method EXLE to generate powerful neural topic-enhanced cross-lingual word embeddings for a fully unsupervised CLIR in this article. This method is built upon mapping two separate monolingual topical embeddings to a shared bilingual space. We propose a three-part method to model neural topic relevance and word embeddings in a mutual reinforcement way. We use both a neural generative model and a rectified skip-gram model while training the enhanced monolingual word embeddings, which are significantly different from the previous approaches. The process of considering neural topic models and word embeddings simultaneously for the neural CLIR is firstly introduced by this article, to the best of our knowledge. Results suggest that the EXLE outperforms the state-of-the-art methods that use aligned monolingual static and contextual word embeddings for generating CLWEs. Our proposed method also performs better than the previously proposed topical word embeddings, delivering statistically significant improvements. In some settings, our method can even outperform the CLIR method which uses a neural machine translation system. As to future work, we aim to experiment on more language combinations, especially between resource-rich and resource-lean languages with more projection approaches to improve the performance of generated CLWEs. Furthermore, we believe that topic-enhanced word embeddings can be naturally incorporated into the overall training framework to a greater extent. We also want to investigate more on correctly aligning multilingual contextual embeddings for the neural CLIR.

CRediT authorship contribution statement

Dong Zhou: Conceptualization, Methodology, Software, Writing – original draft. **Wei Qu:** Data curation, Validation. **Lin Li:** Visualization, Investigation. **Mingdong Tang:** Writing – review & editing. **Aimin Yang:** Writing – review & editing.

Data availability

Data will be made available on request.

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

This work was supported in part by the National Natural Science Foundation of China under Project No. 61876062 & No. 61976061, the Scientific Research Fund of Hunan Provincial Education Department under Project No. 21A0319 and the Hunan Provincial Natural Science Foundation of China under Project No. 2022JJ30020.

References

[1] M. Artetxe, S. Ruder, and D. Yogatama, On the Cross-lingual Transferability of Monolingual Representations, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2018*, Online,4623-4637, 2020.
[2] B. Athiwaratkun, A. Wilson, and A. Anandkumar, Probabilistic FastText for Multi-Sense Word Embeddings, in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018*, Melbourne, Australia, 1-11, 2018.
[3] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, Journal of Machine Learning Research 3 (2003) 993–1022.
[4] H. Bonab, S.M. Sarwar, J. Allan, Training Effective Neural CLIR by Bridging the Translation Gap, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2020, Virtual Event, China, 2020, pp. 9–18.
[5] J. Camacho-Collados, M.T. Pilehvar, N. Collier, R. Navigli, SemEval-2017 Task 2: Multilingual and Cross-lingual Semantic Word Similarity, in: Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval-2017, Vancouver, Canada, 2017, pp. 15–26.
[6] S. Cao, N. Kitaev, and D. Klein, Multilingual Alignment of Contextual Word Representations, in *Proceedings of the 8th International Conference on Learning Representations, ICLR 2020*, Addis Ababa, Ethiopia, 1-13, 2020.

[7] P. Cimiano, A. Schultz, S. Sizov, P. Sorg, and S. Staab, Explicit Versus Latent Concept Models for Cross-Language Information Retrieval, in *Proceedings of the 21st International Jont Conference on Artifical Intelligence, IJCAI 2009*, Pasadena, California, USA, 1513-1518, 2009.
[8] R. Das, M. Zaheer, and C. Dyer, 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 of the Asian Federation of Natural Language Processing, ACL 2015*, Beijing, China, 795-804, 2015.
[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, Minneapolis, Minnesota, 4171-4186, 2019.
[10] A.B. Dieng, F.J.R. Ruiz, D.M. Blei, Topic Modeling in Embedding Spaces, Transactions of the Association for Computational Linguistics 8 (2020) 439–453.
[11] Y. Doval, J. Camacho-Collados, L. Espinosa-Anke, and S. Schockaert, On the Robustness of Unsupervised and Semi-supervised Cross-lingual Word Embedding Learning, *arXiv:1908.07742 [cs],* 2020.
[12] S. T. Dumais, Latent Semantic Indexing (LSI) and TREC-2, in *Proceedings of The Second Text REtrieval Conference, TREC 1993*, Gaithersburg, Maryland, USA, 105-115, 1993.
[13] X. Fu, T. Wang, J. Li, C. Yu, and W. Liu, Improving Distributed Word Representation and Topic Model by Word-Topic Mixture Model, in *Proceedings of The 8th Asian Conference on Machine Learning, ACML 2016*, Hamilton, New Zealand, 190–205, 2016.
[14] G. Glavaš, R. Litschko, S. Ruder, and I. Vulić, 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, ACL 2019*, Florence, Italy, 710-721, 2019.
[15] Y. Goldberg and O. Levy, word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method, *arXiv:1402.3722,* 2014.
[16] A. Gretton, K.M. Borgwardt, M.J. Rasch, B. Schölkopf, A. Smola, A Kernel Two-Sample Test, Journal of Machine Learning Research 13 (2012) 723–773.
[17] Z. Guan, X. Liu, L. Wu, J. Wu, R. Xu, J. Zhang, Y. Li, Cross-lingual multi-keyword rank search with semantic extension over encrypted data, Information Sciences 514 (2020) 523–540.
[18] T. Hofmann, Probabilistic Latent Semantic Indexing, in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR 1999*, Berkeley, California, USA, 50–57, 1999.
[19] Y. Hoshen and L. Wolf, Non-Adversarial Unsupervised Word Translation, *arXiv:1801.06126 [cs],* 2018.
[20] A. M. Hoyle, P. Goel, and P. Resnik, Improving Neural Topic Models using Knowledge Distillation, in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, Online, 1752-1771, 2020.
[21] Z. Jiang, A. El-Jaroudi, W. Hartmann, D. Karakos, and L. Zhao, Cross-lingual Information Retrieval with BERT, *arXiv:2004.13005 [cs, stat],* 2020.
[22] G. Klein, Y. Kim, Y. Deng, V. Nguyen, J. Senellart, and A. Rush, OpenNMT: Neural Machine Translation Toolkit, in *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas, AMTA 2018*, Boston, MA, 177-184, 2018.
[23] J. D. Lafferty and G. Lebanon, Information Diffusion Kernels, in *Proceedings of the Advances in Neural Information Processing Systems 15, NIPS 2002*, Vancouver, British Columbia, Canada, 375-382, 2002.
[24] B. Li, X. Du, M. Chen, Cross-language question retrieval with multi-layer representation and layer-wise adversary, Information Sciences 527 (2020) 241–252.
[25] C. Li, H. Wang, Z. Zhang, A. Sun, and Z. Ma, 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, SIGIR 2016*, Pisa, Italy, 165–174, 2016.
[26] R. Litschko, G. Glavaš, S. P. Ponzetto, and I. Vulić, Unsupervised Cross-Lingual Information Retrieval using Monolingual Data Only, in *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018*, Ann Arbor, MI, USA, 1253-1256, 2018.
[27] R. Litschko, G. Glavaš, I. Vulic, and L. Dietz, 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, SIGIR 2019*, Paris, France, 1109–1112, 2019.
[28] R. Litschko, I. Vulić, S. P. Ponzetto, and G. Glavaš, Evaluating Multilingual Text Encoders for Unsupervised Cross-Lingual Retrieval, in *Proceedings of the 43rd European Conference on IR Research, ECIR 2021*, Virtual Event, 342-358, 2021.
[29] L. Liu, H. Huang, Y. Gao, Y. Zhang, and X. Wei, Neural Variational Correlated Topic Modeling, in *Proceedings of the 2019 World Wide Web Conference, WWW 2019*, San Francisco, CA, USA, 1142-1152, 2019.
[30] Y. Liu, Z. Liu, T.-S. Chua, and M. Sun, Topical Word Embeddings, in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, AAAI 2015*, Austin, Texas, USA, 2418-2424, 2015.
[31] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, Distributed representations of words and phrases and their compositionality, in *Proceedings of the Advances in Neural Information Processing Systems 26, NeurIPS 2013*, Lake Tahoe, Nevada, USA, 3111-3119, 2013.
[32] F. Nan, R. Ding, R. Nallapati, and B. Xiang, Topic Modeling with Wasserstein Autoencoders, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, ACL 2019*, Florence, Italy, 6345-6381, 2019.
[33] J. Pennington, R. Socher, and C. Manning, Glove: Global Vectors for Word Representation, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, Doha, Qatar, 1532-1543, 2014.
[34] T. Ruas, C.H.P. Ferreira, W. Grosky, F.O. de França, D.M.R. de Medeiros, Enhanced word embeddings using multi-semantic representation through lexical chains, Information Sciences 532 (2020) 16–32.
[35] S. Ruder, I. Vulić, A. Søgaard, A Survey of Cross-lingual Word Embedding Models, Journal of Artificial Intelligence Research 65 (2019) 569–631.
[36] P.H. Schönemann, A Generalized Solution of the Orthogonal Procrustes Problem, Psychometrika 31 (1) (1966) 1–10.
[37] T. Schuster, O. Ram, R. Barzilay, and A. Globerson, Cross-Lingual Alignment of Contextual Word Embeddings, with Applications to Zero-shot Dependency Parsing, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, Minneapolis, MN, USA, 1599-1613, 2019.
[38] B. Shi, W. Lam, S. Jameel, S. Schockaert, and K. P. Lai, Jointly Learning Word Embeddings and Latent Topics, in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2017*, Shinjuku, Tokyo, Japan, 375–384, 2017.
[39] S. L. Smith, D. H. P. Turban, S. Hamblin, and N. Y. Hammerla, Offline bilingual word vectors, orthogonal transformations and the inverted softmax, in *Proceedings of the 5th International Conference on Learning Representations, ICLR 2017*, Toulon, France, 1-10, 2017.
[40] A. Søgaard, S. Ruder, and I. Vulić, On the Limitations of Unsupervised Bilingual Dictionary Induction, in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018*, Melbourne, Australia,778-788, 2018.
[41] J. Su, S. Wu, B. Zhang, C. Wu, Y. Qin, D. Xiong, A neural generative autoencoder for bilingual word embeddings, Information Sciences 424 (2018) 287–300.
[42] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf, Wasserstein Auto-Encoders, *arXiv:1711.01558 [cs, stat],* 2019.
[43] I. Vulić, W. De Smet, M.-F. Moens, Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora, Information Retrieval 16 (3) (2013) 331–368.
[44] I. Vulić and M.-F. Moens, 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, SIGIR 2015*, Santiago, Chile, 363–372, 2015.
[45] R. Wang, X. Hu, D. Zhou, Y. He, Y. Xiong, C. Ye, and H. Xu, Neural Topic Modeling with Bidirectional Adversarial Training, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, Online, 340-350, 2020.
[46] Z. Wang, J. Xie, R. Xu, Y. Yang, G. Neubig, and J. G. Carbonell, Cross-lingual Alignment vs Joint Training: A Comparative Study and A Simple Unified Framework, in *Proceedings of the 8th International Conference on Learning Representations, ICLR 2020*, Addis Ababa, Ethiopia, 1-15, 2020.
[47] C. Xing, D. Wang, C. Liu, and Y. Lin, 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, NAACL-HLT 2015*, Denver, Colorado, 1006-1011, 2015.

[48] P. Yu and J. Allan, A Study of Neural Matching Models for Cross-lingual IR, in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2020*, Virtual Event, China, 1637–1640, 2020.
[49] D. Zhou, M. Truran, T. Brailsford, V. Wade, H. Ashman, Translation Techniques in Cross-Language Information Retrieval, ACM Computing Surveys 45 (1) (2012) 1–44.
[50] D. Zhou, W. Zhao, X. Wu, S. Lawless, J. Liu, An iterative method for personalized results adaptation in cross-language search, Information Sciences 430–431 (2018) 200–215.