Monolingual and Cross-Lingual Information Retrieval Models Based on (Bilingual) Word Embeddings

Ivan Vulić and Marie-Francine Moens
Department of Computer Science
KU Leuven, Belgium
{ivan.vulic, marie-francine.moens}@cs.kuleuven.be

ABSTRACT

We propose a new unified framework for monolingual (MoIR) and cross-lingual information retrieval (CLIR) which relies on the induction of dense real-valued word vectors known as word embeddings (WE) from comparable data. To this end, we make several important contributions: (1) We present a novel word representation learning model called Bilingual Word Embeddings Skip-Gram (BWESG) which is the first model able to learn bilingual word embeddings solely on the basis of document-aligned comparable data; (2) We demonstrate a simple yet effective approach to building document embeddings from single word embeddings by utilizing models from compositional distributional semantics. BWESG induces a shared cross-lingual embedding vector space in which both words, queries, and documents may be presented as dense real-valued vectors; (3) We build novel ad-hoc MoIR and CLIR models which rely on the induced word and document embeddings and the shared cross-lingual embedding space; (4) Experiments for English and Dutch MoIR, as well as for English-to-Dutch and Dutch-to-English CLIR using benchmarking CLEF 2001-2003 collections and queries demonstrate the utility of our WE-based MoIR and CLIR models. The best results on the CLEF collections are obtained by the combination of the WE-based approach and a unigram language model. We also report on significant improvements in ad-hoc IR tasks of our WE-based framework over the state-of-the-art framework for learning text representations from comparable data based on latent Dirichlet allocation (LDA).

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval-Information filtering

Keywords

Word embeddings; cross-lingual information retrieval; text representation learning; ad-hoc retrieval; comparable data; semantic composition; multilinguality; vector space retrieval models

1. INTRODUCTION

A revolution in text representation learning ignited by the recent success of neural network architectures and their variants is yet to find its breakthrough in the fundamental IR research. The current state-of-the-art text representation learning framework learns structured word representations known as *word embeddings* (WEs), which are simply (intelligently induced) dense real-valued continuous vectors. Besides improving computational efficiency, WEs lead to better generalizations, even allowing to generalize over the vocabularies observed in labelled data, and hence partially alleviating the ubiquitous problem of data sparsity. Their utility has been validated and proven in various semantic tasks such as semantic word similarity, synonymy detection or word analogy (e.g., [26, 1, 32, 22]). Moreover, the word embeddings have been proven to serve as useful unsupervised features for plenty of standard natural language processing tasks such as named entity recognition, chunking, semantic role labeling, part-of-speech tagging, etc. [40, 7].

Knowing the long-standing and firm relationship between vector space models, semantics modeling and IR [37, 16], *one goal of this paper is to establish a new link between the recent text representation learning methodology based on word embeddings and modeling in information retrieval*, with the focus on the fundamental ad-hoc retrieval task.

Moreover, following the recent trend of multilingual word embedding induction (e.g., [14, 11]), *in this paper we also focus on the induction of bilingual word embeddings (BWEs), and show how to use BWEs in cross-lingual information retrieval tasks*. We show that WE-based monolingual ad-hoc retrieval models may be considered as special and less general cases of the cross-lingual retrieval setting (i.e., operating with only one language instead of two), which results in a unified WE-based framework for monolingual and cross-lingual IR.

When operating in multilingual settings, it is highly desirable to learn embeddings for words denoting similar concepts that are very close in the *shared inter-lingual embedding space* (e.g., the representations for the English word *school* and the Spanish word *escuela* should be very similar). All prior work critically requires sentence-aligned parallel data and readily-available translation dictionaries [14, 11] to induce bilingual word embeddings (BWEs) that are consistent and closely aligned over languages. In this paper, we alleviate this strong requirement by using comparable data.

Contributions. Summarizing all this, in this paper, we investigate three key research questions:
(Q1) *Is it possible to induce high-quality bilingual word embeddings without the need of parallel data and any other readily available translation resources such as bilingual lexicons?* If the answer is positive, we will pave a way for bilingual word representation learning from comparable data for language pairs with limited and non-structured bilingual data resources.
(Q2) *Is it possible to combine word embeddings using the estab-*

SIGIR'15, August 09 - 13, 2015, Santiago, Chile.

DOI: http://dx.doi.org/10.1145/2766462.2767752.

lished models of semantic composition to construct similar representations for text units beyond the word level? If the answer is positive, we will be able to build structured representations of documents and queries to be used in various IR tasks.
(Q3) *Are word embeddings useful in ad-hoc information retrieval?* If the answer is positive, we will be able to build new models for monolingual and cross-lingual information retrieval which use the additional semantic information implicitly coded in the structured real-valued representations of words, documents and queries.

By aiming at providing the answers to Q1-Q3, this paper delivers several contributions as follows:
(C1) We present a novel word representation learning model called *Bilingual Word Embeddings Skip-Gram (BWESG)* which is the first model able to learn the bilingual word embeddings solely on the basis of document-aligned comparable data. BWESG is supported by recent advances in word representation learning.
(C2) We demonstrate a simple yet effective LSI-inspired approach to building document and query embeddings from single word embeddings by utilizing additive models from compositional distributional semantics. BWESG induces a *shared cross-lingual embedding vector space* in which words, queries, and documents may be presented in a uniform way as dense real-valued vectors.
(C3) We construct a novel unified framework for ad-hoc monolingual (MoIR) and cross-lingual information retrieval (CLIR) which relies on the induced word embeddings and constructed query and document embeddings. Experiments for English and Dutch MoIR, as well as for English-to-Dutch and Dutch-to-English CLIR using benchmarking CLEF 2001-2003 collections and queries demonstrate the utility of our novel MoIR and CLIR models based on word embeddings induced by the BWESG model.

2. BWESG: MODEL ARCHITECTURE

Before we describe our architecture for learning bilingual word embeddings (BWEs) from comparable data, we provide a short overview of the underlying skip-gram word representation learning model in monolingual settings. Our new bilingual representation learning model called BWESG is an extension of skip-gram (SG) to multilingual settings with multilingual comparable training data, and serves as the basis of our novel IR/CLIR framework introduced later in sect. 3.2.

2.1 Why Embeddings and Skip-Gram?

The idea of representing words as continuous real-valued vectors dates way back to mid-80s [36]. The idea met its resurgence in [2], where a neural language model learns word embeddings as part of a neural network architecture for statistical language modeling. This work inspired other approaches that learn word embeddings within the neural-network language modeling framework [6, 7]. Word embeddings are tailored to capture semantics and encode a continuous notion of semantic similarity (as opposed to semantically poorer discrete representations), necessary to share information between words and other text units.

Recently, the skip-gram and continuous bag-of-words (CBOW) model from [24, 25] revealed that the full neural-network structure is not needed at all to learn high-quality word embeddings (with extremely decreased training times compared to the full-fledged neural network models, see [24] for the full analysis of complexity of the models). These models are in fact simple single-layered architectures, where the objective is to predict a word's context given the word itself (skip-gram) or predict a word given its context (CBOW). Similar models called vector log-bilinear models were recently proposed in [30, 32].

Due to its simplicity, as well as its efficacy and consequent popularity in various tasks [25, 21, 22], in this paper we will focus on the adaptation of the skip-gram model with negative sampling from [25]. Our new model will operate in multilingual settings and use only document-aligned comparable data for training.

2.2 Skip-Gram Model

As already hinted in sect. 2.1, our departure point is the log-linear skip-gram model from [24] trained using the negative sampling procedure [25], as implemented in the `word2vec` package.[1] The skip-gram model learns word embeddings (WEs) in a similar way to neural language models [2, 6], but without a non-linear hidden layer.

In the monolingual setting, we assume one language L with vocabulary V, and a corpus of words $w \in V$, along with their contexts $c \in V^c$, where V^c is the context vocabulary. Contexts for each word w_n are typically neighboring words in a context window of size cs (i.e., $w_{n-cs}, \ldots, w_{n-1}, w_{n+1}, \ldots, w_{n+cs}$), so effectively it holds $V^c \equiv V$.[2]

Each word type $w \in V$ is associated with a vector $\vec{w} \in \mathbb{R}^{dim}$ (its pivot word representation or pivot word embedding, see fig. 1), and a vector $\vec{w_c} \in \mathbb{R}^{dim}$ (its context embedding). dim is the dimensionality of WEs. The entries in these vectors are latent, and treated as parameters θ to be learned by the model.

In short, the idea of the skip-gram model is to scan through the corpus (which is typically unannotated [24]) *word by word* in turn (i.e., these are the pivot words), and learn from the pairs *(word, context)*. The learning goal is to maximize the ability of predicting context words for each pivot word in the corpus. The probability of observing the context word v given the pivot word w is defined by the softmax function:

$$P(v|w) = \frac{1}{1 + \exp(-\vec{w} \cdot \vec{v_c})} \quad (1)$$

Each word token w in the corpus is treated in turn as the pivot and all pairs of word tokens $(w, w \pm 1)$,...,$(w, w \pm t(cs))$ are appended to the training dataset D, where $t(cs)$ is an integer sampled from a uniform distribution on $\{1, \ldots, cs\}$.[3] The global training objective J is then to maximize the probabilities that all pairs from D are indeed observed in the corpus:

$$J = \arg\max_{\theta} \sum_{(w,v) \in D} \log \frac{1}{1 + \exp(-\vec{w} \cdot \vec{v_c})} \quad (2)$$

where θ are the parameters of the model, that is, pivot and context word embeddings which have to be learned. One may see that this objective function has a trivial solution by setting $\vec{w} = \vec{v_c}$, and $\vec{w} \cdot \vec{v_c} = Val$, where Val is a large enough number [10]. In order to prevent this trivial training scenario, the *negative sampling* procedure comes into the picture [6, 25].

In short, the idea behind negative sampling is to present the model with a set D' of artificially created or sampled "negative pivot-context" pairs (w, v'), which by assumption serve as negative examples, that is, they do not occur as observed/positive *(word,*

[1] `https://code.google.com/p/word2vec/`

[2] Testing other options for context selection such as dependency-based contexts [19] is beyond the scope of this work, and it was recently shown that these contexts may not lead to any gains in the final WEs [13].

[3] The original skip-gram model utilizes dynamic window sizes, where cs denotes the maximum window size. Moreover, the model takes into account sentence boundaries in context selection, that is, it selects as context words only words occurring in the same sentence as the pivot word.

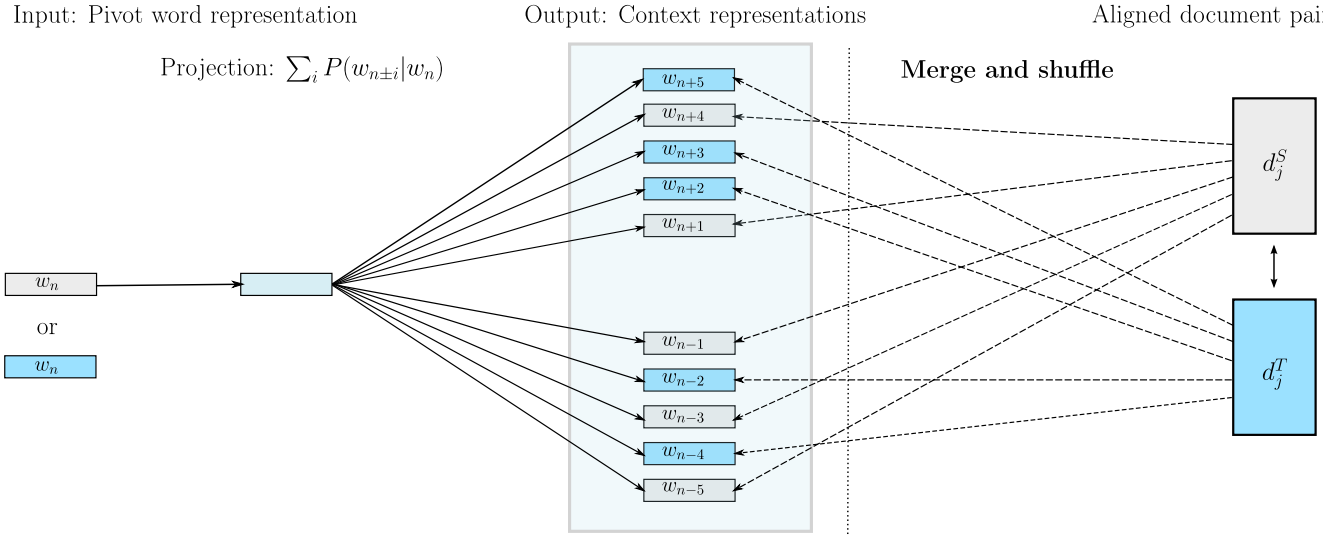


Figure 1: The architecture of our BWE Skip-Gram model for learning bilingual word embeddings from document-aligned comparable data. Source language words and documents are drawn as light gray/unshaded boxes, while target language words and documents as blue/shaded boxes. The right side of the figure (separated by a vertical dashed line) illustrates how a pseudo-bilingual document is constructed from a pair of aligned documents; two documents are first merged, and then words in the pseudo-bilingual document are randomly shuffled to ensure that both source language words and target language words occur as context words.

context) pairs in the training corpus. The model then has to adjust the parameters θ in such a way to also maximize the probability that these negative pairs will not occur in the corpus. The interested reader may found further details about the negative sampling procedure, and the new exact objective function along with its derivation in [21]. For illustrative purposes and simplicity, here we present the approximative objective function with negative sampling by Goldberg and Levy [10]:

$$J = \arg\max_{\theta} \sum_{(w,v)\in D} \log \frac{1}{1+\exp(-\vec{w}\cdot\vec{v_c})} + \sum_{(w,v')\in D'} \log \frac{1}{1+\exp(\vec{w}\cdot\vec{v_c'})} \quad (3)$$

The free parameters θ are updated using stochastic gradient descent and backpropagation, with global learning rate as defined in `word2vec` [24]. By optimizing the objective from eq. (3), the model incrementally pushes observed pivot WEs towards context WEs of their collocates in the corpus. After training, words that predict similar context words should end up having similar WEs.

2.3 Final Model: BWE Skip-Gram

In the next step, we extend the skip-gram model to work with bilingual document-aligned comparable data (e.g., Wikipedia articles in different languages aligned through inter-wiki links). An overview of our architecture for learning BWEs from comparable data is given in fig. 1.

Let us assume that we possess a document-aligned comparable corpus which is defined as $\mathcal{C} = \{d_1, d_2, \ldots, d_N\} = \{(d_1^S, d_1^T), (d_2^S, d_2^T), \ldots, (d_N^S, d_D^T)\}$, where $d_j = (d_j^S, d_j^T)$ denotes a pair of aligned documents in the source language L_S and the target language L_T, respectively, and N is the number of documents in the corpus. V^S and V^T are vocabularies associated with languages L_S and L_T. The goal is to learn word embeddings for all words in both V^S and V^T which will be semantically coherent and closely aligned over languages.

In the first step, we *merge* two documents d_j^S and d_j^T from the aligned document pair d_j into a single "pseudo-bilingual" document d_j' and remove sentence boundaries. Following that, we randomly *shuffle* the newly constructed pseudo-bilingual document. The intuition behind this pre-training shuffling step (see fig. 1) is to assure that each word w, regardless of its actual language, obtains word collocates from both vocabularies. The idea of having bilingual contexts for each pivot word in each pseudo-bilingual document will steer the final model towards constructing a shared inter-lingual embedding space. Since the model depends on the alignment at the document level, in order to ensure the bilingual contexts instead of monolingual contexts, it is intuitive to assume that larger window sizes will lead to better bilingual embeddings. We test this hypothesis and the effect of window size in sect. 5.1.

The final model called BWE Skip-gram (BWESG) then relies on the monolingual variant of the skip-gram model trained on these shuffled pseudo-bilingual documents. The model learns word embeddings for source and target language words which are aligned over the dim embedding dimensions and may be represented in the same shared inter-lingual embedding space. The BWESG-based representation of word w, regardless of its actual language, is then a dim-dimensional vector:

$$\overrightarrow{w} = [f_{w,1}, \ldots, f_{w,k}, \ldots, f_{w,dim}] \quad (4)$$

where $f_{w,k} \in \mathbb{R}$ denotes the score for the k-th inter-lingual feature associated with w within the dim-dimensional shared inter-lingual embedding space. Since all words share the embedding space, semantic similarity between words may be computed both monolingually and across languages. We will extensively use this property during the construction of our MoIR and CLIR models.

3. BWESG-BASED IR: FROM WORD EMBEDDINGS TO RETRIEVAL MODELS

The BWESG model induces structured representations from comparable data for single words only. In this section, we show how to compute similar structured representations for text units beyond the level of single words. The compositional approach will be used to construct embeddings for *documents* and *queries*, and it will be the basic building block of our MoIR and CLIR models. To this

end, we rely on two key modeling assumptions: (1) We treat documents and queries as bags of words and do not impose any syntactic information to the document structure. This simplification is the standard practice in IR modeling, as in the ubiquitous unigram language model (e.g., [35, 3, 23]), relevance models (e.g., [18, 17]) or topic model based retrieval models (e.g., [44, 45, 12]; (2) We rely on the intuitions behind *semantic composition* models from the literature on distributional compositional semantics (e.g., [4, 27]).

For simplicity, we only discuss CLIR modeling in this section. The simpler MoIR models may be directly derived from the more general CLIR setting.

3.1 Constructing Document and Query Embeddings

Let us assume that a document d from a target document collection $\mathcal{DC}$ is a bag of word tokens $d = \{w_1, w_2, \ldots, w_{|N_d|}\}$, where $|N_d|$ denotes the length of the document d expressed by the number of word tokens.

In order to present the document d in the dim-dimensional embedding space induced by the BWESG model, we need to apply a model of *semantic composition* to learn its dim-dimensional vector representation $\overrightarrow{d}$. Formally, we may specify its vector representation as the dim-dimensional vector/embedding:

$$\overrightarrow{d} = \overrightarrow{w_1} \star \overrightarrow{w_2} \star \ldots \star \overrightarrow{w_{|N_d|}} \quad (5)$$

where $\overrightarrow{w_1}$,...,$\overrightarrow{w_{|N_d|}}$ are dim-dimensional WEs learned from the training data (see eq. (4)), and $\star$ is a compositional vector operator such as addition, point-wise multiplication, tensor product, matrix multiplication, etc.

A plethora of models for semantic composition have been proposed in the relevant literature, differing in their choice of vector operators, input structures and required knowledge (e.g., [29, 38, 4, 11, 27], to name only a few). In this work, driven by the observed linear linguistic regularities in the embedding spaces (e.g., [26, 20])[4], we opt for simple *addition* (denoted by +) from [29] as the compositional operator, due to its simplicity, the ease of applicability considering the bag-of-words assumption, as well as its solid performance in various compositional tasks [29, 27]. The dim-dimensional document embedding (DE) $\overrightarrow{d}$ is then a vector:

$$\overrightarrow{d} = \overrightarrow{w_1} + \overrightarrow{w_2} + \ldots + \overrightarrow{w_{|N_d|}} \quad (6)$$

We have constructed a document representation in the same embedding space where word embeddings live (see eq. (4) again):

$$\overrightarrow{d} = [f_{d,1}, \ldots, f_{d,k}, \ldots, f_{d,dim}] \quad (7)$$

We call this composition model for the DE induction *ADD-BASIC*.

In an extended approach, we may weigh the summands from eq. (6) using their *self-information* as weights. The idea is that the IDF-inspired weights will assign more importance to words bearing more information content during the compositional process. The property should then be reflected in a better document $\overrightarrow{d}$ in the dim-dimensional embedding space, and its "truer" semantics. We compute the self-information weight si_w of word w based on its occurrence in the target document collection $\mathcal{DC}$ using the standard formula [8] as follows:

$$si_w = -\ln \frac{freq(w, \mathcal{DC})}{|N_{\mathcal{DC}}|} \quad (8)$$

where $freq(w, \mathcal{DC})$ denotes the frequency of word w in the target document collection $\mathcal{DC}$, and $|N_{\mathcal{DC}}|$ is the size of the document collection measured by the total number of word tokens. A document embedding is then constructed out of its WEs as follows:

$$\overrightarrow{d} = si_{w_1} \cdot \overrightarrow{w_1} + si_{w_2} \cdot \overrightarrow{w_2} + \ldots + si_{w_{|N_d|}} \cdot \overrightarrow{w_{|N_d|}} \quad (9)$$

We call this composition model for the DE induction *ADD-SI*.

The construction of query embeddings (QEs) under the bag-of-words assumption proceeds in a completely analogous manner. As the knowledge of the target document collection may not be known from the query side, we restrict the construction of QEs to the ADD-BASIC composition model. A QE of a query $Q = \{q_1, q_2, \ldots, q_m\}$ containing m query terms in the embedding space is then:

$$\overrightarrow{Q} = \overrightarrow{q_1} + \overrightarrow{q_2} + \ldots + \overrightarrow{q_m} \quad (10)$$

$$\overrightarrow{Q} = [f_{Q,1}, \ldots, f_{Q,k}, \ldots, f_{Q,dim}] \quad (11)$$

3.2 Final MoIR and CLIR Models

In the previous section, we have demonstrated how to build embeddings for documents and queries, and how to represent them in the same embedding semantic space. This property makes the computation of semantic similarity between some document d from the target document collection and the issued query almost trivial. The similarity score $sim(d, Q)$ directly related to the relevance of d for Q is computed by applying a similarity function (SF) on their embeddings. In this paper, we use the standard cosine similarity measure as SF, the standard choice in the embedding induction literature (e.g., [25, 30, 21]):

$$sim(d, Q) = SF(d, Q) = \frac{\overrightarrow{d} \cdot \overrightarrow{Q}}{|\overrightarrow{d}| \cdot |\overrightarrow{Q}|} \quad (12)$$

The documents from $\mathcal{DC}$ are then ranked in descending order according to their similarity scores with query Q.

Since our new BWESG model learns to represent words from two different languages in the same shared cross-lingual embedding space, we may use exactly the same modeling approach to monolingual and cross-lingual information retrieval. All words in the embedding space retain their "language annotations"; although the words from two different languages are represented in the same semantic space, we still know whether a word belongs to language L_S (e.g., English) or language L_T (e.g., Dutch). We may now summarize the complete CLIR retrieval process as follows:

1. Given is a document-aligned comparable corpus in two languages L_S and L_T with vocabularies V^S and V^T. **Induce** the set of bilingual word embeddings $\mathcal{BWE}$ using the BWESG embedding learning model (see sect. 2.3). The set comprises dim-dimensional word embeddings $\overrightarrow{w}$ for each word $w \in V^S$ and each word $w \in V^T$.
2. Given is a target document collection $\mathcal{DC} = \{d'_1, \ldots, d'_{N'}\}$ in language L_T, where N' denotes the number of documents in the collection. **Compute** dim-dimensional document embeddings $\overrightarrow{d'}$ for each $d' \in \mathcal{DC}$ using the dim-dimensional WEs from the set $\mathcal{BWE}$ obtained in the previous step and a semantic composition model (ADD-BASIC or ADD-SI; see eq. (6) and eq. (9)).

[4]Recent work has shown that word embeddings are excellent at capturing semantic regularities in language, where the semantic relations are typically characterized by a relation-specific linear offset. The famous example from [26] states that the linear combination of word embeddings $\overrightarrow{king} - \overrightarrow{man} + \overrightarrow{woman}$ will result in a new vector closest to the embedding of $\overrightarrow{queen}$.

Table 1: Statistics of the experimental setup: (a) Monolingual EN→EN and NL→NL retrieval; (b) Cross-lingual EN→NL and NL→EN retrieval. In the "Query Set" column: for instance, EN'01:41-90 denotes the language of the query (EN) and the year of the campaign (2001), while 41-90 denotes the corresponding standard CLEF themes used to extract queries for that campaign.

Monolingual					Cross-lingual				
Direction	$\mathcal{DC}$	# Docs	Query Set	# Queries	*Direction*	$\mathcal{DC}$	# Docs	Query Set	# Queries
EN→EN 2001	LAT	110,861	EN'01: 41-90	47	NL→EN 2001	LAT	110,861	NL'01: 41-90	47
EN→EN 2002	LAT	110,861	EN'02: 91-140	42	NL→EN 2002	LAT	110,861	NL'01: 91-140	42
EN→EN 2003	LAT+GH	166,753	EN'03: 141-200	53	NL→EN 2003	LAT+GH	166,753	NL'03: 141-200	53
NL→NL 2001	NC+AD	190,604	NL'01: 41-90	50	EN→NL 2001	NC+AD	190,604	EN'01: 41-90	50
NL→NL 2002	NC+AD	190,604	NL'02: 91-140	50	EN→NL 2002	NC+AD	190,604	EN'02: 91-140	50
NL→NL 2003	NC+AD	190,604	NL'03: 141-200	56	EN→NL 2003	NC+AD	190,604	EN'03: 141-200	56

3. After the query $Q = \{q_1, \ldots, q_m\}$ is issued in language L_S, **compute** a dim-dimensional query embedding using the ADD-BASIC composition model (see eq. (11)).
4. **For each** $d' \in \mathcal{DC}$, **compute** the semantic similarity score $sim(d', Q)$ which quantifies each document's relevance to the query Q (see eq. (12)).
5. **Rank** all documents from $\mathcal{DC}$ according to their similarity scores from the previous step.

The only difference in the MoIR retrieval process is the fact that both the query and the target document collection are given in the same language. Strictly speaking, when performing monolingual retrieval, one does not require a comparable bilingual corpus any more, as the monolingual WEs may be simply learned on a large monolingual corpus (e.g., [25, 30]).[5] However, for the clarity of presentation, we have decided to stress the complete modeling analogy between the monolingual and cross-lingual approach to IR. In summary, we have created a unified framework for MoIR and CLIR which relies solely on word embeddings induced in an unsupervised fashion from document-aligned comparable data.

4. EXPERIMENTAL SETUP

Training Collections. We use the same training data collection as in [41, 42]. The training dataset comprises 7,612 document-aligned English-Dutch Wikipedia article pairs together with 6,206 Europarl English-Dutch document pairs [15]. We do not exploit alignments at the sentence level in Europarl and treat it as a corpus aligned only at the document level. After the stop words removal, the final vocabularies consist of 76,555 words in English and 71,168 words in Dutch [41].

Test Collections and Queries. All our experiments were performed on the standard datasets used in the cross-lingual evaluation of the CLEF 2001-2003 campaigns (e.g., [33, 34]). The target collection in Dutch (NL) comprises 190,604 Dutch news articles from the NRC Handelsblad 94-95 and the Algemeen Dagblad 94-95 newspapers (NC+AD). In English (EN), one target collection comprises 110,861 news articles from the 1994 LA times (LAT), and another target EN collection additionally includes the Glasgow Herald 1995 and comprises 166,753 news articles in total (LAT+GH).

As a standard practice [17, 44, 41], both English and Dutch queries have been extracted from the *title* and *description* fields of the CLEF themes for the years 2001-2003. Stop words were removed from queries, and queries without relevant documents were removed from the query sets in both languages.

[5]Investigating the differences in the MoIR performance when embeddings are trained on a larger monolingual corpus (as opposed to only the monolingual part from a comparable dataset) is left for future work.

We evaluate all retrieval models within the CLEF 2001-2003 evaluation campaigns in (1) *monolingual ad-hoc retrieval*: (1a) English queries and English target documents ($EN \rightarrow EN$); (1b) Dutch queries and Dutch target documents ($NL \rightarrow NL$), and (2) *cross-lingual ad-hoc retrieval*: (2a) English queries and Dutch target documents ($EN \rightarrow NL$); (2b) Dutch queries and English target documents ($NL \rightarrow EN$). Tab. 1 provides an overview of the complete experimental setup.

Mean Average Precision (MAP) scores are used as the main evaluation metric for all experiments. We have also experimented with recall-precision curves, but for brevity we omit these in the final presentation, as the main findings were in line with the analysis based on the MAP scores.

Retrieval Models in Comparison. We use the same families of models for both MoIR and CLIR.

WE-VS. Our new retrieval model which relies on the induction of word embeddings and their usage in the construction of query and document embeddings is described in sect. 3.2. We investigate the retrieval ability of our new vector space retrieval model based on (bilingual) word embeddings by comparing it to the set of standard MoIR and CLIR models. We opt for ADD-BASIC as the composition model unless noted otherwise. Later we show the comparison of ADD-BASIC and ADD-SI (see sect. 5.3).

LM-UNI. The first baseline model in comparison is the omnipresent standard query likelihood model from Ponte and Croft [35] which generates the query Q under the bag-of-words assumption that terms are independent given the documents. It is a unigram language model specified by the following formula with standard Dirichlet smoothing [46]:

$$P(Q|d) = \prod_{i=1}^{m} P(q_i|d)$$
$$= \prod_{i=1}^{m} \frac{N_d}{N_d + \mu} P(q_i|d) + (1 - \frac{N_d}{N_d + \mu}) P(q_i|\mathcal{DC}) \quad (13)$$

where N_d is the length of document d from the target collection $\mathcal{DC}$, and μ is the parameter of Dirichlet smoothing. $P(q_i|d)$ is the maximum likelihood estimate of the query term q_i in d, while $P(q_i|\mathcal{DC})$ is its maximum likelihood estimate in the entire target collection. In the actual implementation, we operate with log probabilities. It is intuitive that the LM-UNI model will lead to much better results in the monolingual setting, as the amount of shared words between different languages is typically very limited, and therefore other representations for CLIR are sought [41] (see next).

LDA-IR. Another baseline retrieval model similar to WE-VS, relies

on structured semantic-representations of words and documents obtained by the latent Dirichlet allocation (LDA) model [5] for the monolingual setting, and its bilingual variant called bilingual LDA [28, 31] for the cross-lingual setting. The LDA-based MoIR model was introduced in [44], while the description of the BiLDA-based CLIR model is available in [41]. In short, query likelihood is computed as follows:

$$P(Q|d) = \prod_{i=1}^{m} P(q_i|d) = \prod_{i=1}^{m} \sum_{k=1}^{K} P(q_i|z_k)P(z_k|d) \quad (14)$$

where z_k denotes the k-th (cross-lingual) latent topic induced from training data (out of K topics), $P(q_i|z_k)$ is a probability score learned from the training data, and $P(z_k|d)$ are probability scores after running the trained LDA/BiLDA model on each document of the target collection. In the cross-lingual setting the latent topics act as a bridge over the lexical chasm between two different languages [41, 43].

LM-UNI+LDA-IR. It was shown both for the monolingual [44] and cross-lingual retrieval [41] that *combining* the basic unigram language model with a semantically aware model such as LDA-IR leads to improved retrieval models, where the combined model typically scores significantly higher than each of the single models. Since LM-UNI and LDA-IR are probabilistic models, it is straightforward to combine them using the following formula [44, 41]:

$$P(q_i|d) = \lambda P_{lda}(q_i|d) + (1-\lambda)P_{lm}(q_i|d) \quad (15)$$

where $P_{lda}(q_i|d)$ is the probability score of the LDA-IR model acquired following eq. (14), and $P_{lm}(q_i|d)$ is the probability score of LM-UNI following eq. (13). λ is the linear interpolation parameter which assigns the weight balance between the two constituent models. The combined LM-UNI+LDA-IR is a very strong monolingual baseline [44], and is a state-of-the-art CLIR model that requires only comparable data for training [41], the same setup as for our BWESG model and the corresponding WE-VS retrieval model.

LM-UNI+WE-VS. Following the same line of thinking, we hypothesize that the combination of the unigram model with our new semantically aware WE-VS model will also lead to improved retrieval scores. While the comparison of single LDA-IR and WE-VS models will directly analyze which of the two semantic representations (LDA-based or embeddings-based) is better fit for the ad-hoc MoIR and CLIR retrieval tasks, the comparison of the combined LM-UNI+LDA-IR and LM-UNI+WE-VS will analyze whether we profit from introducing the WE-based semantic representations into the combined model. We combine the two models using a simple approach as follows: (1) separately rank the documents from $\mathcal{DC}$ using LM-UNI ($score_{lm}(d, Q)$, see eq. (13)) and WE-VS (see sect. 3.2, the score is $score_{we}(d, Q)$); (2) normalize the scores in both ranked lists to the interval $[0, 1]$; (3) linearly combine the normalized scores, with λ as the interpolation parameter:

$$score(d, Q) = \lambda score_{we}(d, Q) + (1-\lambda) score_{lm}(d, Q) \quad (16)$$

GT+LM+LDA: An additional baseline in the CLIR setting uses an SMT system (i.e., *Google Translate* (GT)) to translate the query from the source to target language, and then the actual retrieval is performed monolingually using the LM-UNI+LDA-IR monolingual model as described in [44, 41]. Unlike the LDA-based and WE-based CLIR models, this (SMT-based) model does not construct a shared inter-lingual semantic space, and in addition it uses an external translation resource as an extra source of knowledge.

Parameters. Since each of the retrieval models comes with the burden of its own parameters, we list the parameters by model. The parameter values from the single models are also used in the combined models unless stated otherwise.

WE-VS: We have trained the BWESG model with random shuffling on 10 random corpora shuffles of our training data with all other parameters set to the default parameters for skip-gram from the `word2vec` package [25]. We have varied the number of dimensions d from 100 to 800 in steps of 100. Moreover, in order to test the effect of window size on final results, we have varied the maximum window size cs from 10 to 100 in steps of 10.

LM-UNI: The Dirichlet interpolation parameter is set to the standard suggested value: $\mu = 1000$ [46, 44].

LDA-IR: We have trained LDA and BiLDA using Gibbs sampling [39, 43] with the suggested value for the number of topics: $K = 1000$ which yielded optimal or near-optimal results in MoIR and CLIR as reported in the literature [44, 39, 41]. All other parameters of the topic models (i.e., hyper-parameters, the number of iterations of the Gibbs sampler) have also been set to the standard values reported before in the relevant literature.

LM-UNI+LDA-IR: In order to ensure the optimal performance of the baseline combined model, the grid search over λ values $\{0.0, \ldots, 1.0\}$ (in steps of 0.1) has been performed, and the results with the optimal λ value are reported.

LM-UNI+WE-VS: We have tested several typical values for the interpolation parameter λ: 0.3 (more weight is assigned to LM-UNI); 0.5 (equal importance); 0.7 (more weight is assigned to WE-VS).

5. RESULTS AND ANALYSIS

In this section, we evaluate our novel WE-based framework in the ad-hoc monolingual and cross-lingual tasks. We again stress that for CLIR we operate in a difficult setting which requires only comparable data to induce cross-lingual structured semantic representations, and does not rely on any parallel sentence-aligned data or readily available bilingual lexicons.

5.1 Experiment I: Monolingual Retrieval

Test I: Single Models. The results of the single MoIR models (WE-VS, LM-UNI and LDA-IR) are available in tab. 2. We show the results with $dim = 300$ and $dim = 600$, and the maximum window size cs set to 60. We investigate the influence of the dimensionality of representation and maximum window size later in Test III. All MAP scores for WE-VS are computed as averages over 10 different random corpora shuffles.[6] Based on the results, we may observe several interesting phenomena:

(i) Across all evaluation runs, our new WE-VS retrieval model which relies on the induction of word embeddings scores higher than LDA-IR, the other model which relies on structured semantic representations of words and documents. It is the first evidence that combining the BWESG model for inducing word embeddings plus a semantic composition model to induce document and query embeddings may be more beneficial in IR applications than previously widely used LDA- and BiLDA-based representations and approaches. All differences in scores are highly statistically significant (see the results of statistical significance tests in tab. 2).

[6]We have conducted a small experiment testing the influence of the random shuffling procedure, and have detected that the fluctuation of results for all models relying on the BWESG embeddings caused by the random shuffling procedure is typically non-significant. Moreover, the fluctuation is lesser than the fluctuation caused by the inherent randomness of the original skip-gram model (due to negative sampling and window size sampling) [25], and may be further decreased by relying on larger window sizes.

Table 2: A comparison of single and combined MoIR models. All results are given as MAP scores. LM+LDA is an abbreviation denoting the LM-UNI+LDA-IR model, while LM+WE refers to LM-UNI+WE-VS model. y and x denote statistically significant improvements ($p < 0.05$ and $p < 0.01$, respectively) of the WE-VS model over LDA-IR using a two-tailed Wilcoxon signed rank test. They also denote statistically significant improvements ($p < 0.05$ and $p < 0.01$, respectively) of the combined LM+WE model over another combined LM+LDA model (which is by design an upper bound of the LM-UNI model) using the same test.

	EN→EN			NL→NL		
Model	2001	2002	2003	2001	2002	2003
LM-UNI	.381	.360	.359	.256	.323	.357
LDA-IR	.279	.216	.241	.131	.143	.130
dim:300; cs:60						
WE-VS	.324x	.258x	.257y	.203x	.237x	.224x
dim:600; cs:60						
WE-VS	.329x	.281x	.262y	.204x	.262x	.231x
LM+LDA	.399	.360	.379	.260	.326	.357
dim:300; cs:60						
LM+WE (λ=0.3)	.412y	.381x	.401y	.271x	.349x	.372x
LM+WE (λ=0.5)	.429x	.394x	.407x	.279x	.370x	.382x
LM+WE (λ=0.7)	.451x	.392y	.389	.270	.364x	.373y
dim:600; cs:60						
LM+WE (λ=0.3)	.419y	.382x	.403y	.274x	.350x	.373x
LM+WE (λ=0.5)	.436x	.391x	.408x	.282x	.371x	.383x
LM+WE (λ=0.7)	.430x	.392y	.381	.268	.367x	.374y

Table 3: A comparison of single and combined CLIR models. All results are given as MAP scores. LM+LDA+WE denotes a combination of all three single models. y and x denote statistically significant improvements ($p < 0.05$ and $p < 0.01$, respectively) of the WE-VS model over LDA-IR using a two-tailed Wilcoxon signed rank test. They also denote statistically significant improvements ($p < 0.05$ and $p < 0.01$, respectively) of the combined LM+WE and LM+LDA+WE models over baseline LM+LDA using the same test.

	NL→EN			EN→NL		
Model	2001	2002	2003	2001	2002	2003
LM-UNI	.094	.108	.092	.078	.125	.112
LDA-IR	.197	.139	.123	.145	.137	.171
dim:300; cs:60						
WE-VS	.187	.204x	.120	.174	.185y	.157
dim:600; cs:60						
WE-VS	.222y	.230x	.127	.178y	.219x	.181
LM+LDA	.267	.225	.199	.225	.268	.278
GT+LM+LDA	.307	.275	.248	.230	.240	.244
dim:300; cs:60						
LM+WE (λ=0.3)	.189	.273	.197	.101	.159	.150
LM+WE (λ=0.5)	.218	.283y	.220	.113	.184	.167
LM+WE (λ=0.7)	.255	.307x	.219	.180	.209	.208
dim:600; cs:60						
LM+WE (λ=0.3)	.205	.281y	.198	.107	.167	.154
LM+WE (λ=0.5)	.236	.299x	.215	.123	.203	.183
LM+WE (λ=0.7)	.286	.317x	.222	.190	.249	.225
dim:600; cs:60						
LM+LDA+WE (λ=0.3)	.277	.263	.210	.229	.288	.283
LM+LDA+WE (λ=0.5)	.281y	.281y	.214	.240	.297y	.290
LM+LDA+WE (λ=0.7)	.302x	.302x	.227	.244y	.311x	.302y

(ii) Due to the larger number of dimensions, WE-VS with $dim = 600$ should provide more semantic expressiveness than WE-VS with $dim = 300$. The difference in results seems to be consistent across all evaluation runs. However, the difference is small and statistically insignificant for most of the evaluation runs.

(iii) As expected, LM-UNI is the best scoring single MoIR model and it scores better even then our new WE-VS model. This is consistent with previous findings (e.g., [44, 45]) for the monolingual setting, where it has been stated that structured semantic representations such as LDA-based representations (LDA-IR) (or now embeddings) are too coarse-grained to produce retrieval scores on a par with the scores obtained by a query likelihood unigram language model. However, the prior work also demonstrates that structured semantic representations are very useful when fused with the unigram language model (see next).

Test II: Combined Models. Another test aimed at detecting which of the semantic representations (i.e., LDA-based or WE-based) is more helpful when combined with the LM-UNI model. The MAP scores presented in tab. 2 reveal the following:

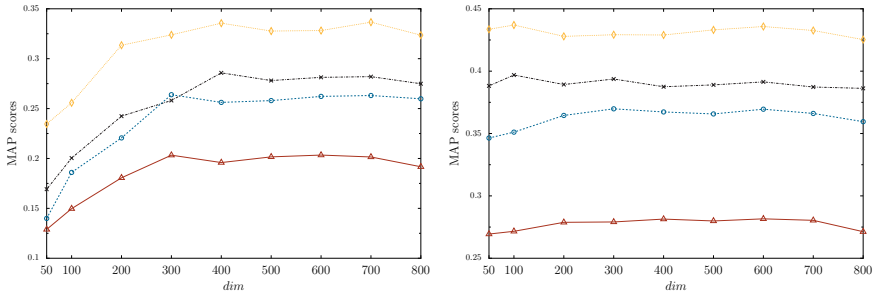
(i) Our new combined LM-UNI+WE-VS model which combined the semantic knowledge coded in document embeddings with a unigram language model produces the highest overall MAP scores over all evaluation runs. The MAP scores obtained by LM-UNI+WE-VS are also higher than the scores obtained by the other combined model in comparison, which fuses LM-UNI and LDA-IR (LM-UNI+LDA-IR), and the improvements are largely significant.

(ii) Since we have used the optimal value for λ when reporting the scores for LM-UNI+LDA-IR, all results of that combined model are by design higher than LM-UNI alone. It leads to this general conclusion: Including additional semantic knowledge (in the form of document embeddings composed from the previously induced word embeddings) is useful for MoIR modeling as it captures complementary pieces of information and encodes the document semantics that a simple unigram model cannot capture. The semantic knowledge is better coded in word and document embeddings than in the previously widely used LDA-based representation models for IR.

Test III: Influence of Dimensionality and Window Size. In order to test how the dimensionality of embeddings and the size of the local context cs influence the final retrieval scores, we conducted two experiments. In the first experiment, we fixed the parameter cs to 60 and then varied the number of dimensions dim from 50 to 800. The MAP scores across different settings for dim for the WE-VS and LM-UNI+WE-VS MoIR retrieval models for a selection of evaluation runs (2001 and 2002 campaigns) are presented in fig. 2(a) and fig. 2(b). In another experiment, we fixed the parameter dim to 300 and varied the number of dimensions cs from 10 to 100. The results for WE-VS and LM-UNI+WE-VS are displayed in fig. 3(a) and fig. 3(b). We may summarize the findings here:

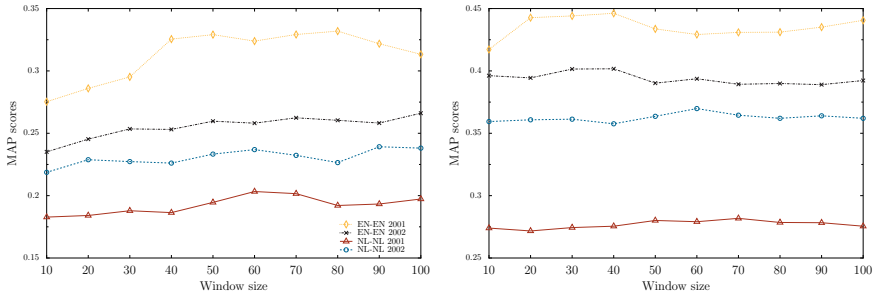
(i) The performance of WE-VS is stable across different settings for parameter dim for larger dim-s (≥ 300). Smaller vectors ($d = 50, 100, 200$) are simply not semantically expressive, and do not capture and encode all semantic properties of words and documents. On the other hand, it is interesting to notice that the combined LM-UNI+WE-VS is fairly robust to the change of the parameter dim controling the dimensionality of representation, and all embedding models (even with $d = 50$) help boost the scores in the combined model.

(ii) The performance of WE-VS is also relatively stable across different settings for cs for larger values of the parameter (≥ 30), but comparable results are obtained even with smaller window sizes. Again, we may notice that the combined model is robust to the fluctuation of cs.



(a) WE-VS, $cs = 60$ (b) LM-UNI+WE-VS, $cs = 60$

Figure 2: An analysis of the behavior of (a) WE-VS and (b) LM-UNI+WE-VS across different parameter settings for dimensionality dim of WEs, QEs, and DEs constructed using the BWESG model. The maximum window size cs is set to 60 for all models.



(a) WE-VS, $dim = 300$ (b) LM-UNI+WE-VS, $dim = 300$

Figure 3: An analysis of the behavior of (a) WE-VS and (b) LM-UNI+WE-VS across different settings for the maximum window size parameter cs of WEs, QEs, and DEs constructed using the BWESG model. Dimensionality of the representation dim is set to 300.

Based on these analyses, we may conclude that the inclusion of embeddings in the retrieval process is invariantly useful. The same conclusion is valid for the cross-lingual retrieval scenarios, but we do not show these results due to space constraints.

5.2 Experiment II: Cross-Lingual Retrieval

Test I: Single Models. The results of the single CLIR models (WE-VS, LM-UNI and LDA-IR) are available in tab. 3. We again show the results with $dim = 300$ and $dim = 600$, and the maximum window size cs set to 60, and all MAP scores for WE-VS are computed as averages over 10 different random corpora shuffles.

(i) WE-VS is now the best scoring single CLIR model across all evaluation runs. LM-UNI, which was the best scoring MoIR model, is now outscored by the other two models which rely on structured semantic representations. While the words shared between a query and a target document are very strong indicators for the retrieval process in the monolingual setting, their importance is heavily diminished in the cross-lingual setting, as the amount of shared words between two different languages is very limited. In the setting without any bilingual dictionaries or parallel data, the shared cross-lingual semantic representations become increasingly important. (ii) LDA-IR model which relies on the BiLDA topic model [28] is the current state-of-the-art single model for CLIR with comparable data [41]. However, we may see that our new WE-VS CLIR model outperforms LDA-IR across all evaluation runs.

Test II: Combined Models. Similar as for MoIR, the combined CLIR models are also compared. The results are available in tab. 3.

(i) The combined LM-UNI+LDA-IR model for CLIR, which was the top scoring query likelihood model in [41, 42], is outperformed by LM-UNI+WE-VS for the NL-to-EN retrieval direction. However, it is not the case for the EN-to-NL retrieval, where the LM+LDA combination significantly outperforms the new LM+WE model. Although the single WE-VS model compares favorably to LDA-IR, the advantage of WE-VS is lost in the combined model. We have investigated the causes of such behavior and the difference in results for NL-to-EN and EN-to-NL retrieval directions, and have come to the following conclusion: (1) In the NL-to-EN retrieval direction, many queries do not contain any NL terms to be found in the EN target collectiosn. Therefore, the LM-UNI model is unable to provide any ranking of the documents there, and the ranking is provided solely on the basis of the semantic models such as LDA-IR or WE-VS. Moreover, the NL terms which are found in both NL queries and EN target collections are typically very relevant to the information need (e.g., named entities such as *Alberto Tomba*, *Miguel Indurain*), so the combined model may profit from the information from both single models. However, in the EN-to-NL direction, many irrelevant and general terms from EN queries are found in the NL target collection (e.g., words like *words*, *document*, *telling*) which lead LM-UNI and the combined model astray; (2) Since WE-VS is a vector space model, while LM-UNI is a probabilistic language model, their combination is executed post-hoc, after obtaining the rankings separately and then normalizing the scores (see eq. (16) in sect. 4). This way, the errors from the LM-UNI model are assigned too much weight in the combined model. This negative effect is removed when combining LM-UNI and LDA-IR (see eq. (15) in sect. 4), since both models are probabilistic and their combination is executed at the level of single query terms. This way, the errors from LM-UNI may be immediately corrected by smoothing it with semantically aware LDA-IR.

Test III: Yet Another Combined Model. Knowing all this, in order to test whether embeddings may be useful even for the EN-to-NL retrieval direction, we have constructed a new combined model which combines the evidence from all three single models. The new combined LM-UNI+LDA-IR+WE-VS model first computes

Table 4: A comparison (MAP scores) of two different semantic composition models from sect. 3.2 (ADD-BASIC and ADD-SI) utilized to construct document embeddings from the word embeddings induced by BWESG. The retrieval model is WE-VS. dim is set to the value 300 and 600 (in brackets), while cs is 60 for all models. y and x denote statistically significant improvements ($p < 0.05$ and $p < 0.01$, respectively) of ADD-SI over ADD-BASIC using a two-tailed Wilcoxon signed rank test.

	Monolingual						Cross-lingual					
	EN→EN			NL→NL			NL→EN			EN→NL		
Composition	2001	2002	2003	2001	2002	2003	2001	2002	2003	2001	2002	2003
ADD-BASIC (300)	.324	.258	.257	.203	.237	.224	.187	.204	.120	.174	.185	.157
ADD-SI (300)	.338	.278y	.255	.212	.253y	.227	.216x	.213y	.122	.189y	.208x	.161
ADD-BASIC (600)	.329	.281	.262	.204	.262	.231	.221	.230	.127	.178	.219	.181
ADD-SI (600)	.344y	.301y	.263	.215	.275y	.234	.237y	.233	.130	.189	.229x	.184

separate retrieval scores using WE-VS and LM-UNI+LDA-IR (this way, we diminish the effect of the errors caused by using LM-UNI alone), normalizes the scores to the interval $[0, 1]$, and then computes the combined scores using eq. (16) again. The results for the new combined model are given in tab. 3.

(i) We may observe that the new combined model now significantly outperforms the LM-UNI+LDA-IR model [41] over all evaluation runs with English queries and Dutch target collections. The LDA-IR model decreases the errors from LM-UNI, and then the WE-VS model introduces additional semantic knowledge that is coded in document embeddings. Although both LDA-IR and WE-VS rely on structured semantic representations of documents and query terms, it seems that the two models capture partially complementary pieces of information, and combining them is beneficial for the complete retrieval process.

(ii) As already detected in [44, 41], fusing complementary retrieval evidence in the combined models leads to better overall models. The fusion of different retrieval clues is especially important in the minimalist CLIR setting which relies only on comparable training data for the induction of additional semantic knowledge in the form of structured text representations.

5.3 Experiment III: Composition Models

In the final experiment, we test whether better semantic composition models lead to higher-quality document embeddings, and consequently to higher overall retrieval scores. A more informed and better composition function should better reflect how the composite meaning of a document is constructed from the meanings of its constituents, that is, single words. We compare our ADD-BASIC and the ADD-SI approach to semantic composition from sect. 3.1 using the WE-VS model for MoIR and CLIR. The results are reported in tab. 4.

The results clearly reveal that ADD-SI which is an additive model with additional weighting based on self-information outperforms the simpler ADD-BASIC model. The improvements are small, but consistent across all evaluation runs both for MoIR and CLIR. Encouraged by the findings, we plan to study more elaborate composition models and their influence on the IR models in future work.

6. CONCLUSIONS AND FUTURE WORK

We have presented a novel unified framework for monolingual and cross-lingual information retrieval that relies on the induction of dense real-valued word vectors known as word embeddings. This paper is a pilot study focusing on the importance and ability of the new neural-net inspired semantically-aware text representations in the IR/CLIR setting.

We have introduced a new model for learning structured monolingual and bilingual word embeddings known from comparable data. Our Bilingual Word Embeddings Skip-Gram (BWESG) model is the first model that is able to learn the shared cross-lingual semantic embedding space without the need for parallel data or readily available bilingual dictionaries. These minimal requirements make the model easily applicable to plenty of language pairs.

Further, we have presented how to compose query and document embeddings from single word embeddings induced by the BWESG model, and how to use these embeddings in the construction of novel monolingual and cross-lingual retrieval models. The new models which rely on the semantic knowledge coming from the embeddings demonstrate better results than previous state-of-the-art models in the monolingual and cross-lingual retrieval tasks conducted on the benchmarking CLEF datasets.

We believe that the proposed framework is only a start, as it ignites a series of new research questions and perspectives. (i) A straightforward path of future research is to use the BWESG model for other language pairs, and apply to other tasks such as (cross-lingual) document classification and clustering [14, 11]. (ii) Another straightforward extension is to replace the simple pre-training shuffling procedure with a more systematic context selection method. (iii) In this paper, we have relied only on simple additive semantic composition models, but based on the results from sect. 5.3, we believe that better results may be achieved by introducing more complex composition models available in the literature (e.g., [4, 27]). (iii) Finally, we have only investigated query likelihood approaches to ad-hoc retrieval, but one path of future work leads to studying the potential of the embedding-based approaches with the pseudo-relevance feedback modeling frameworks such as [17, 9].

Acknowledgments

We would like to thank the reviewers for their comments and suggestions. This research has been carried out in the frameworks of the Smart Computer-Aided Translation Environment (SCATE) project (IWT-SBO 130041) and the Personalised AdveRtisements buIlt from web Sources (PARIS) project (IWT-SBO 110067).

7. REFERENCES

[1] M. Baroni, G. Dinu, and G. Kruszewski. Don't count, predict! A systematic comparison of context-counting vs context-predicting semantic vectors. In *ACL*, pages 238–247, 2014.

[2] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155, 2003.

[3] A. Berger and J. Lafferty. Information retrieval as statistical translation. In *SIGIR*, pages 222–229, 1999.

[4] W. Blacoe and M. Lapata. A comparison of vector-based representations for semantic composition. In *EMNLP-CoNLL*, pages 546–556, 2012.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[6] R. Collobert and J. Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *ICML*, pages 160–167, 2008.

[7] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. P. Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537, 2011.

[8] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 1991.

[9] F. Diaz and D. Metzler. Improving the estimation of relevance models using large external corpora. In *SIGIR*, pages 154–161, 2006.

[10] Y. Goldberg and O. Levy. Word2vec explained: Deriving Mikolov et al.'s negative-sampling word-embedding method. *CoRR*, abs/1402.3722, 2014.

[11] K. M. Hermann and P. Blunsom. Multilingual models for compositional distributed semantics. In *ACL*, pages 58–68, 2014.

[12] J. Jagarlamudi and H. Daumé III. Extracting multilingual topics from unaligned comparable corpora. In *ECIR*, pages 444–456, 2010.

[13] D. Kiela and L. Bottou. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *EMNLP*, pages 36–45, 2014.

[14] A. Klementiev, I. Titov, and B. Bhattarai. Inducing crosslingual distributed representations of words. In *COLING*, pages 1459–1474, 2012.

[15] P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT SUMMIT*, pages 79–86, 2005.

[16] T. K. Landauer and S. T. Dumais. Solutions to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240, 1997.

[17] V. Lavrenko, M. Choquette, and W. B. Croft. Cross-lingual relevance models. In *SIGIR*, pages 175–182, 2002.

[18] V. Lavrenko and W. B. Croft. Relevance-based language models. In *SIGIR*, pages 120–127, 2001.

[19] O. Levy and Y. Goldberg. Dependency-based word embeddings. In *ACL*, pages 302–308, 2014.

[20] O. Levy and Y. Goldberg. Linguistic regularities in sparse and explicit word representations. In *CoNLL*, pages 171–180, 2014.

[21] O. Levy and Y. Goldberg. Neural word embedding as implicit matrix factorization. In *NIPS*, pages 2177–2185, 2014.

[22] O. Levy, Y. Goldberg, and I. Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of ACL,* to appear, 2015.

[23] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[24] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *ICLR Workshop Papers*, 2013.

[25] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119, 2013.

[26] T. Mikolov, W. Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In *NAACL-HLT*, pages 746–751, 2013.

[27] D. Milajevs, D. Kartsaklis, M. Sadrzadeh, and M. Purver. Evaluating neural word representations in tensor-based compositional settings. In *EMNLP*, pages 708–719, 2014.

[28] D. Mimno, H. Wallach, J. Naradowsky, D. A. Smith, and A. McCallum. Polylingual topic models. In *EMNLP*, pages 880–889, 2009.

[29] J. Mitchell and M. Lapata. Vector-based models of semantic composition. In *ACL*, pages 236–244, 2008.

[30] A. Mnih and K. Kavukcuoglu. Learning word embeddings efficiently with noise-contrastive estimation. In *NIPS*, pages 2265–2273, 2013.

[31] X. Ni, J.-T. Sun, J. Hu, and Z. Chen. Mining multilingual topics from Wikipedia. In *WWW*, pages 1155–1156, 2009.

[32] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.

[33] C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, editors. *CLEF 2001, Revised Papers*, 2002.

[34] C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, editors. *CLEF 2002, Revised Papers*, 2003.

[35] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *SIGIR*, pages 275–281, 1998.

[36] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.

[37] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[38] R. Socher, B. Huval, C. D. Manning, and A. Y. Ng. Semantic compositionality through recursive matrix-vector spaces. In *EMNLP-CoNLL*, pages 1201–1211, 2012.

[39] M. Steyvers and T. Griffiths. Probabilistic topic models. *Handbook of Latent Semantic Analysis*, 427(7):424–440, 2007.

[40] J. P. Turian, L. Ratinov, and Y. Bengio. Word representations: A simple and general method for semi-supervised learning. In *ACL*, pages 384–394, 2010.

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

[42] I. Vulić and M. Moens. A unified framework for monolingual and cross-lingual relevance modeling based on probabilistic topic models. In *ECIR*, pages 98–109, 2013.

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

[44] X. Wei and W. B. Croft. LDA-based document models for ad-hoc retrieval. In *SIGIR*, pages 178–185, 2006.

[45] X. Yi and J. Allan. A comparative study of utilizing topic models for information retrieval. In *ECIR*, pages 29–41, 2009.

[46] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems*, 22(2):179–214, 2004.