Lang Resources & Evaluation (2020) 54:1059–1078
https://doi.org/10.1007/s10579-020-09498-1

ORIGINAL PAPER

Evaluating cross-lingual textual similarity on dictionary alignment problem

Yiğit Sever[1] · Gönenç Ercan[2]

Published online: 29 June 2020

Abstract Bilingual or even polylingual word embeddings created many possibilities for tasks involving multiple languages. While some tasks like cross-lingual information retrieval aim to satisfy users' multilingual information needs, some enable transferring valuable information from resource-rich languages to resource-poor ones. In any case, it is important to build and evaluate methods that operate in a cross-lingual setting. In this paper, Wordnet definitions in 7 different languages are used to create a semantic textual similarity testbed to evaluate cross-lingual textual semantic similarity methods. A document alignment task is created to be used between Wordnet glosses of synsets in 7 different languages. Unsupervised textual similarity methods—Wasserstein distance, Sinkhorn distance and cosine similarity—are compared with a supervised Siamese deep learning model. The task is modeled both as a retrieval task and an alignment task to investigate the hubness of the semantic similarity functions. Our findings indicate that considering the problem as a retrieval and alignment problem has a detrimental effect on the results. Furthermore, we show that cross-lingual textual semantic similarity can be used as an automated Wordnet construction method.

Keywords Cross-lingual textual semantic similarity · Word embeddings · Wasserstein distance · Sinkhorn distance · Siamese neural network

✉ Gönenç Ercan
gonenc@cs.hacettepe.edu.tr

Yiğit Sever
yigit@ceng.metu.edu.tr

1 Department of Computer Engineering, Middle East Technical University, Ankara, Turkey

2 Institute of Informatics, Hacettepe University, Ankara, Turkey

1 Introduction

Recently proposed polylingual information retrieval methods are breaking the language barrier in many tasks. Today it is possible to search in one language to retrieve resources indexed in another language (Balikas et al. 2018). Tasks such as cross-lingual search (Vulić and Moens 2015; Litschko et al. 2018) and plagiarism detection (Barrón-Cedeño et al. 2010; Potthast et al. 2011; Franco-Salvador et al. 2016; Rupnik et al. 2016) are becoming more effective.

Furthermore, by building on these tools, it is possible to advance the state-of-the-art of core natural language processing tasks by cross-lingual training and transfer learning techniques (Johnson et al. 2019). Naturally, these methods require some representation such as multilingual word embeddings that can operate between languages. Thus, evaluation of both the word embeddings and the methods using these embeddings is an important endeavour.

Word embeddings are used to create an embedding space that encodes semantic relationships between words (Mikolov et al. 2013b). Methods for building polylingual word embeddings are proposed, extending these word embedding spaces to span more than one language (Artetxe et al. 2018a; Jawanpuria et al. 2019). A major contribution of these embeddings is the creation of unsupervised machine translation methods that do not require parallel corpora (Leng et al. 2019).

With the introduction of new polylingual word embedding methods, evaluation of their performance became an important task. Cross lingual document classification (Klementiev et al. 2012) and dependency parsing (Upadhyay et al. 2016) are two tasks used for evaluating the bilingual word embeddings. Most of the methods evaluate the built embeddings in word-level tasks like bilingual dictionary induction (BLI) (Mikolov et al. 2013a) that evaluates the embeddings based on the distance between word pairs from both languages. However, overfitting towards higher scores in BLI does not necessarily lead to similar performance for downstream tasks (Glavas et al. 2019). While some of these evaluation methods are based on high level tasks like cross-lingual dependency parsing, an evaluation of textual similarity metrics ranging from unsupervised to supervised is missing from the literature.

To fill this gap, we build a cross-lingual textual similarity evaluation dataset from wordnet concept definitions. English Princeton WordNet (PWN) (Fellbaum 1998) is a database of semantic and lexical relationships of senses. The same concepts defined for PWN along with their semantic relationships are used for other language editions of Wordnet (Vossen 1998). These concepts, referred to as synsets also contain definitions known as *glosses*. As synsets are shared among wordnets, it is possible to align different language glosses on sense-level. We use Open Multilingual Wordnet (OMW) (Bond and Paik 2012) to create a cross-lingual testbed for semantic textual similarity.

Another motivation is the alignment of monolingual dictionaries to the PWN definitions which can be considered as an automatic wordnet construction method. Although there are Wordnet construction methods based on different resources like word translations (Ercan and Haziyev 2019), word sense disambiguation (Taghizadeh and Faili 2016) and word embeddings (Khodak et al. 2017) they are focused on mapping words to synsets to associate the correct lemmas in the target language with the synset. Unfortunately, these translated synsets lack definitions in the target language but only lemmas are translated. The proposed method maps definitions possibly retrieved from a monolingual dictionary translating both the lemmas and the definitions to the target language.

The main contribution of this research is the construction of an evaluation methodology for cross-lingual textual similarity methods. The proposed evaluation methodology aims to investigate the following research questions.

- Comparison between unsupervised similarity functions, namely more traditionally used cosine similarity, Wasserstein distance and its entropy regularized version Sinkhorn.
- Comparison between supervised and unsupervised methods, to investigate what can be achieved with a learning paradigm.
- As dictionary alignment is one-to-one, it is possible to formulate the problem as both a similarity based retrieval method and a maximum weighted bipartite graph matching. These two approaches are used to investigate the homogeneity of the sentence-level similarity functions. Even when the similarity function is symmetric, the nearest neighbors of source and target language sentences can differ from each other, having different levels of similarity for the same sentence pair. The gap between the empirical results for the retrieval and alignment based methods indicate the noise created by this non-homogeneity.

The rest of the article is structured as follows; Sect. 2 presents the related work including methods for learning bilingual word embeddings and their evaluation, and wordnet construction. Section 3 presents the framework of the study, the two approaches to dictionary alignment and the details of our dataset. Section 4 introduces the evaluated cross-lingual textual semantic similarity metrics and Sect. 5 covers the details of the evaluation methodology. Results are discussed in Sect. 6.

2 Related work

2.1 Related work on word embeddings

Word embeddings are high dimensional representations of words, able to encode semantic and syntactic information. They improved the state-of-the-art results in many language related tasks, making them essential resources and an influential research topic. Most word embedding methods model the context of words by accumulating statistics from large text corpora whereas the context is defined in terms of the local proximity in text, e.g. window of 5 words, sentences or paragraphs. Finally, the context statistics of a word is typically encoded with a single d-dimensional vector, creating a word embedding space.

The seminal work of Mikolov et al. , proposed neural language models (Bengio et al. 2003) to build the word vectors in lower dimensional space without building the full-dimensional co-occurrence matrices (Mikolov et al. 2013b). The GloVe model used the second order associations with other words to better model the similarity between the words (Pennington et al. 2014). FastText (Mikolov et al. 2018) built upon word2vec by adding position dependent features and subword information, achieving a robust model towards handling out of vocabulary words (Bojanowski et al. 2016).

Bilingual word embeddings extend this idea to get vector spaces that include representations for two languages. As text documents are usually written in one language, using the statistics of a corpora is not adequate and a bilingual signal is needed to infer the relationships across the languages. One such signal is parallel corpora, where aligned sentences in both languages are provided to the construction method (Gouws et al. 2015). Another signal is the use of bilingual lexicons, where word alignments exist for a small subset mostly referred as seed lexicon (Mikolov et al. 2013a). Some completely unsupervised methods work on either comparable corpora (Mogadala and Rettinger 2016; Luong et al. 2015) or identify the word alignments automatically (Alvarez-Melis and Jaakkola 2018). A recent survey article compared and reviewed different bilingual word embedding construction methods (Ruder et al. 2019).

Bilingual word embeddings are evaluated on downstream tasks where the model is built from one language and the evaluation is carried out in another, mainly cross lingual document classification (Klementiev et al. 2012; Gouws et al. 2015). Further approaches include dependency parsing (Upadhyay et al. 2016) and bilingual dictionary induction (BLI) (Gouws et al. 2015). Recently, Glavas et al. (2019) evaluated various approaches on cross lingual natural language inference and cross lingual document retrieval.

2.2 Related work on wordnet construction

Vossen (1998) broke down wordnet construction into *merge* and *expand* approaches. Merge approach first creates a wordnet in the target language and then maps the relationship between synsets of source and target languages. Expand approach uses machine translation or other unsupervised methods to induce a wordnet in the target language using PWN or other existing wordnet. Due to cost of creating a wordnet from ground up, expand approach is preferred.

Diab (2004) showed that expand approach can still be used for morphologically dissimilar languages such as Arabic and English. Also, they have identified that semantic relationships in a language's wordnet hold for the target language's wordnet as well.

Early work on automated wordnet construction use parallel corpora (Sagot and Fišer 2008), external knowledge sources like Wikipedia (Ruiz-Casado et al. 2005) or rely on machine translation (Lam et al. 2014).

More recent methods started using word embeddings in combination with other resources such as machine translation and bilingual dictionaries. Sand et al. (2017) trained skip-gram variant of word2vec (Mikolov et al. 2013b) to extend existing synsets of Norwegian wordnet using the nearest neighbours of the lemmas. Khodak et al. (2017) used machine translation for a target word in target language by querying the PWN with all possible candidates. They proposed two prominent approaches to filter from the candidate set; using sentence embeddings of glosses and word sense induction. Note that none of the methods use cross-lingual word embeddings, but rely on machine translation, bilingual dictionaries or parallel corpora instead.

3 General evaluation framework

In this section we describe our general evaluation framework. The dataset is comprised of wordnet glosses in 7 languages (Table 2). Each gloss is associated with a single synset per language. We assume that each gloss defines the sense of the synset. The OMW of Bond and Paik (2012) ensures that glosses are matched one-to-one across each language. The dataset is described in detail in Sect. 3.1.

Take wordnet definition or *gloss* g_q^p in language *p*. Given two sets of glosses for both source and target languages as $S = \{g_1^s, \dots g_n^s\}$ and $T = \{g_1^t, \dots g_n^t\}$, the task is to find a *mapping* $\phi(g_i^s) = g_j^t$ using only bilingual word embedding based similarity functions. This directly evaluates the cross-lingual short-text similarity function over concept definitions acquired from Wordnet.

The mapping function ϕ is defined in two distinct ways to evaluate the semantic similarities further. The function ϕ_r models the problem as a *pseudo-retrieval* task. A gloss in source language is mapped to the most similar gloss in the target language, multiple glosses can be mapped to a single target gloss.

$$\phi_r(g_i^s) = \underset{g^t \in T}{\operatorname{argmax}} \left(\operatorname{sim}(g_i^s, g^t)\right) \tag{1}$$

A common issue in high-dimensional vector spaces is the hubness problem (Ruder et al. 2019). That is, some of the word vectors (embeddings) in the vector space tend to be similar to many of the other vectors. Since the vector space is defined in terms of the word embeddings, sentence representations building on these embeddings can inherit the hubness problem. To investigate the role of hubs in our study, we cast the mapping of wordnet definitions problem to a maximum bipartite graph matching problem. New mapping function ϕ_m seeks a one-to-one mapping where a vertex participates only once in a similarity mapping and the sum of edges is maximum with respect to all possible mappings. This can be considered as finding a bijective mapping *M* between the sets *S* and *T* with the maximum sum of the similarities (Fig. 1).

$$\phi_m = \underset{M}{\operatorname{argmax}} \left(\sum_{(g^s, g^t) \in M} \operatorname{sim}(g^s, g^t) \right) \tag{2}$$

$$\begin{array}{c c} & \begin{array}{ccc} g_1^s & g_2^s & g_3^s \end{array} \\ \begin{array}{c} g_1^t \\ g_2^t \\ g_3^t \end{array} & \begin{pmatrix} 0.0181 & 0.0126 & 0.0221 \\ 0.0123 & 0.0089 & 0.0144 \\ 0.0100 & 0.0096 & 0.0114 \end{pmatrix} \end{array} \tag{3}$$

For example, Table 1 list three wordnet definitions from PWN and Romanian Wordnet. The similarity matrix formed using sentence embeddings (Sect. 4) is

Table 1 The example of how matching can improve the results over retrieval

	English	Romanian
headspace.n.01	The volume left at the top of a filled container (bottle or jar or tin) before sealing	Volumul rămas într-un recipient (sticlă, borcan sau bidon) înainte de sigilare
glass.v.04	Put in a glass container	A umple (sub presiune) o butelie, o sticlă, î nchizând-o apoi ermetic
trunk_lid.n.01	Hinged lid for a trunk	capac prin în balamale de cufăr

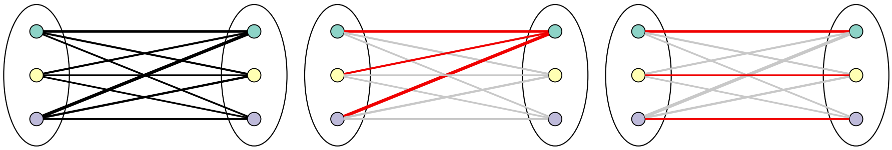

Fig. 1 The comparison of retrieval and matching approaches. The two sets denote two languages and nodes denote the definitions. The nodes on the same level denote the same sense across languages. The stroke widths are proportional to the textual similarity between two nodes. The topmost node acts as a hub for the retrieval approach and each definition is mapped to it. Matching approach optimizes the overall similarity

shown in Table 3. With the retrieval scheme ϕ_r, all 3 target definitions are mapped to `headspace.n.01` since it has the highest cosine similarity to all Romanian glosses. This simply leads to a 0.33 accuracy. The matching case ϕ_m solves the linear assignment for this cost matrix to maximize the cosine similarity and yields a score of 0.044 with all 3 glosses aligned correctly. We hypothesize that the difference between pseudo-retrieval (ϕ_r) and matching (ϕ_m) strategies reveal how much the similarity function is prone to the hubness problem.

3.1 Wordnet Dataset

The Open Multilingual Wordnet (OMW)[1] project has aggregated free, open access wordnets. They serve wordnets in different languages, unifying them under consistent format and common PWN 3.0 identifiers. A subset of languages from OMW is selected to create a cross-lingual textual similarity dataset. The majority of research on creating and extending wordnets focuses on linking synsets across languages, giving less emphasis to glosses. Since our study uses wordnet definitions, we are constrained only to the resources that include gloss information. So, among the 34 wordnets collected by OMW, 6 wordnets that contain definitions are used.

We present the wordnets used to build the dataset on Table 2. Among these wordnets the number of available definitions vary significantly; from as large as 50×10^3 to as low as 3×10^3. Further details of these wordnets are given in Bond and Paik (2012). Although the number of wordnets can be increased by including other languages wordnets like Danish (Pedersen et al. 2009), Russian (Balkova et al. 2004) and German (Hamp and Feldweg 1997), as we intend to publish the dataset publicly, we opted to limit the dataset to only wordnets with permissive licenses.

Each wordnet we use were built by manual or semi-automatic methods, the glosses are mostly taken from ready resources like dictionaries. The definitions for the Bulgarian wordnet BulTreeBank (Simov and Osenova 2010) has been acquired from two large lexical sources and has been constructed using a merge approach. Greek wordnet (Stamou et al. 2004) has been developed under the BalkaNet project and went through the validation process for gloss completeness (Grigoriadou et al.

[1] http://compling.hss.ntu.edu.sg/omw/.

2004). Similarly, since it's inception under the BalkaNet project (Tufiş et al. 2004), the Romanian Wordnet (Tufiş et al. 2008) have opted to reference comprehensive Romanian lexicographic resources and have manually checked when necessary and continued to develop their resources in the following years. The Italian wordnet ItalWordnet (Toral et al. 2010) has built upon the Italian wordnet that was part of the EuroWordNet project (Alonge et al. 1999). The initial paper states that multiple dictionaries and a reference corpus has been used for the construction of the resource. The Slovene wordnet SloWNet (Fišer et al. 2012) sources their definitions automatically from Wikipedia. At the time of writing, the details of Albanet (Ruci 2008) are not available. With the addition of PWN, we have resources that encompass 7 languages, all of which are free and open access. Wordnets will be referred by their corresponding ISO 639-1 code in the rest of the article.

As all the glosses are associated with a PWN 3.0 synset, they are aligned and each definition with the same synset id refers to the same concept. Our evaluation framework maps from source language wordnet to a target language, so pairs of these 7 wordnets can be used to create the datasets. Although there are 21 possible pairs, the pairs that contain less than 2000 common synsets are discarded. Table 3 shows the statistics of the 15 language pairs used in the experiments. The reported numbers reveal that the available synsets do not always overlap. For instance, although Greek and Italian wordnets both contain more than 10K definitions, their intersection is 4801.

As wordnet lists related concepts tied to each other with "is a type of" (hypernymy) and "part of" (meronymy) relationships, some of the concepts are very similar to each other. Wordnet can be encoded as a graph, where the synsets are vertices and edges are these relationships. The *sparsity* of the wordnet graph quantifies the number of synsets that are connected with a hypernym/hyponym or meronym/holonym relationships. The graph density is defined as in Equation 4.

$$D(G) = \frac{2 * E(G)}{V(G) * (V(G) - 1)} \tag{4}$$

We present the density of wordnet pairs in Table 3. Some pairs such as el-sq and bg-el are more dense than others such as en-ro.

Finally, our dataset, evaluation framework and implementation are publicly available online[2].

3.2 Bilingual word embeddings

Different methods for building bilingual word embeddings have been proposed in the literature (Artetxe et al. 2018b; Gouws et al. 2015; Ruder et al. 2019). As we focus on textual similarity of definitions rather than individual words, evaluating and comparing these embeddings are not in the scope of this article. Thus, the embeddings are built using a single method for all language pairs to make a fair comparison.

[2] https://yigitsever.github.io/Evaluating-Dictionary-Alignment/.

Table 2 Details on wordnets used for the study

References	Language	Language code	Number of definitions
Ruci (2008)	Albanian	`sq`	4681
Simov and Osenova (2010)	Bulgarian	`bg`	4959
Fellbaum (1998)	English	`en`	117,659
Stamou et al. (2004)	Greek	`el`	18,136
Toral et al. (2010)	Italian	`it`	12,688
Tufiş et al. (2008)	Romanian	`ro`	58,754
Fišer et al. (2012)	Slovene	`sl`	3144

Table 3 The number of matching synsets in language pairs and the corresponding relationship densities in these synsets

Language pair	Common synsets	Density $*10^5$
en-bg	4959	1.219
en-el	18,136	1.094
en-it	12,688	0.549
en-ro	58,754	0.290
en-sl	3144	0.634
en-sq	4681	3.847
bg-el	2817	2.289
bg-it	2115	1.566
bg-ro	4701	1.323
el-it	4801	1.724
el-ro	2144	1.317
el-sq	4681	3.854
it-ro	10,353	0.725
ro-sl	2085	1.123
ro-sq	4646	3.868

FastText (Mikolov et al. 2018) provides embeddings for 157 different languages trained on Wikipedia and Common Crawl[3]. For efficiency concerns, each embedding set of the 7 languages are pruned to the the most frequent 500×10^3 words in each language. These word embeddings of different languages are not on the same embedding space thus cannot be used directly to measure cross-lingual word similarity.

VecMap[4] (Artetxe et al. 2018a) is used to align the embeddings to a shared space. VecMap can align embeddings using a seed lexicon or using an unsupervised method. We opted to use the supervised mode and acquired the top 10, 000 word alignments, sorted by confidence, from OpenSubtitles 2018 (Lison and Tiedemann

[3] https://fasttext.cc/.

[4] https://github.com/artetxem/vecmap.

2016)[5] as the seed lexicon for each language. After the alignment, a cross-lingual embedding space for each of the 15 language pairs are created.

4 Cross-lingual textual similarity

In this section we introduce the textual similarity functions that we evaluated in this study.

4.1 Unsupervised textual similarity methods

Machine translation monolingual baseline (`MT`) As the baseline, the task at hand is cast into monolingual retrieval by translating the glosses of 6 wordnets (excluding PWN) into English using machine translation. Google Translate is used as the MT engine. A simple monolingual word document matrix is constructed using a corpora built from both definition sets and weighted using smoothed tf-idf measure such that for the word x in the matrix, $x_{t,d} = \mathrm{tf}_{td} \cdot \log \frac{N+1}{df_w+1}$ where tf_{td} is the frequency of the word t in definition d and df_w is the number of definitions that include w. Cosine similarity is used as the similarity metric between the vector representations of two definitions. Only the retrieval approach ϕ_r is used such that the most similar vector is selected per source vector.

Cosine similarity between sentence embeddings (`SEMB`) Arora et al. (2016) has shown that weighed average of word vectors as sentence representations perform well. In their proposed approach, word embeddings that make up a sentence is weighed with *smooth inverse frequency* (SIF) and averaged across the sentence. Using smooth inverse frequency weighting, word embeddings $v_w \in R^d$ where word w is in a vocabulary V can be averaged over a sentence S such that $S \subset V$ to get sentence embedding v_S in the same dimensionality R^d.

$$v_S = \frac{1}{|S|} \sum_{w \in S} \mathrm{SIF}\ (\mathrm{w}) v_w \tag{5}$$

The authors point out that the metric is similar to tf-idf weighting scheme if "one treats a "sentence" as a "document" and make the reasonable assumption that the sentence doesn't typically contain repeated words" (Arora et al. 2016). These assumptions hold for us so we scaled our word embeddings using tf-idf weights to get *sentence embeddings*; $v_S = \frac{1}{|S|} \sum_{w \in S} \mathrm{tf\text{-}idf}_{w,S} v_w$. The sentence embedding representation is used with both the matching ϕ_m and retrieval ϕ_r schemes.

Word mover's distance (`WMD`) In order to incorporate a distance metric that handles individual similarities between words via word embeddings instead of relying upon a binary overlap metric, Kusner et al. (2015) has proposed to show the distances between documents as a special case of Earth Mover's Distance (Rubner et al. 1998). Given two documents d_1 and d_2, the distance between two documents

[5] http://opus.nlpl.eu/index.php.

can be shown as the cost of transferring individual words of one document into the other.

First off, for a vocabulary with n many words, the documents are represented as n dimensional normalized bag-of-words vectors. For two documents d^1 and d^2, the words that occur in the document are specified as d_i^1 and d_j^2. The word embeddings of words are x_i and x_j, Kusner et al. (2015) uses Eucledian distances so that $c(i,j) = \|x_i - x_j\|_2$. Finally, the WMD aims to optimize the transport matrix T for the given constraints;

$$min \sum_{i,j=1}^{n} T_{i,j} c(i,j) \text{subject to} \sum_{j=1}^{n} T_{ij} = d_i^1 \text{and} \sum_{i=1}^{n} T_{ij} = d_j^2 \tag{6}$$

Sinkhorn (SNK) Cuturi (2013) extended the WMD metric by introducing a smoothed version that uses *entropic regularization* which allowed them to use Sinkhorn-Knopp matrix scaling algorithm (Sinkhorn and Knopp 1967). Balikas et al. (2018) used Sinkhorn distance in cross-lingual document retrieval problem. On top of the WMD metric, instead of using normalized bag of words representation for the documents, they used term frequency and tf-idf to weigh the document representations.

Sinkhorn distance is similar to WMD with the addition of regularization term λ;

$$min \sum_{i,j=1} T_{i,j} c(i,j) - \frac{1}{\lambda} E(T) \text{subject to} \sum_{j=1}^{n} T_{i,j} = d_i^1 \text{and} \sum_{i=1}^{n} T_{i,j} = d_j^2 \tag{7}$$

where $E(T)$ is the entropy of the transport matrix T.

4.2 Supervised textual similarity method

Siamese long-short term memory (LSTM) We used a similarity function that relies on cosine similarity over the outputs of the LSTM units.

$$\exp(\sum ||y^s \cdot y^t||_2) \tag{8}$$

where y^s is the output of the LSTM that encodes the source definitions and y^t is the output of the LSTM that encodes the target definitions. We use the AdaDelta optimizer as suggested by Zeiler (2012) and our loss function is the mean squared error. The LSTM weights are uniformly initialized (Glorot and Bengio 2010). Furthermore, gradient clipping is employed in order to avoid the exploding gradient problem (Pascanu et al. 2012). We train our model using positive and negative definition pairs; with the aligned golden corpora that was created by simply bringing the definitions that have the same synset ids together for a wordnet pair, we randomly shuffle half of the definitions and label them as a negative pair.

The definitions are encoded as a sequence of words, one-hot-coded to map to the pretrained cross-lingual word embeddings. Following the input layer and the embedding layer, a Siamese LSTM network with shared weights encode the definitions in both languages. The cosine similarity between the two encoded definitions are calculated, forcing the network to learn a projection matrix in the LSTM layer. To avoid overfitting dropout is used in the LSTM layer. The objective for the network is the mean squared error between the learned similarity and the synset alignment. Given a source definition and a candidate translation, the neural network calculates a supervised similarity between the two definitions. The hyperparameters of the model, including dropouts and the number of units in the LSTM layer are tuned using grid search for each language, optimized with a held-out validation data for each language.

5 Evaluation methodology

All language pairs are considered for the evaluation. The contrary would have been revolving around English as a pivot language and creating pairs that always included English. In order to have a representative evaluation, the language pairings that do not offer at least 2000 pairs of definitions are ignored. The unsupervised approaches are evaluated by calculating the tf-idf scores using all available corpus then randomly selecting 1000 pairs of definitions. For `WMD` and `SNK`, we have extended the open source implementation of Balikas et al. (2018).[6] The linear assignment solver for matching approaches ϕ_m is the Jonker-Volgenant algorithm, available online.[7]

Evaluation is done on precision at one (P@1) metric for both the retrieval ϕ_r and the matching ϕ_m approaches. If the task of finding the closest target definition given the source definition is thought as a query, over a set of queries Q, P@1 is defined as;

$$P@1 = \frac{1}{Q}\sum_{q\in Q}\delta(q) \tag{9}$$

Where $\delta(q)$ is 1 if the highest ranked definition in the result has the same identifier as the source definition and 0 otherwise. Since there is no ranking but a single match for the ϕ_m, the measure becomes a binary indicator function for whether two definitions have the same identifier or not. The results will be presented in percentage (P@1 $\times$ 100).

6 Results and discussion

The same test sets are used for both supervised and unsupervised algorithms for a fair comparison. All available instances are used for training the LSTM model. We will start by discussing the pseudo-retrieval ϕ_r approach for all the applicable similarity metrics. Then the results of the matching approach ϕ_m are discussed. Finally we perform a case-study to investigate the performance of using cross-

6 https://github.com/balikasg/WassersteinRetrieval.

7 https://github.com/src-d/lapjv.

Table 4 P@1% results of the retrieval (ϕ_r) approaches

Source language	Target language	SEMB	SNK	WMD	Supervised	MT
bg	el	7.1	28	28.8	9.7	10.1
bg	it	5.8	24.4	25.1	7.6	0.1
bg	ro	8.4	29.6	28.9	9.3	3.6
el	it	4.7	27.4	26.4	12.8	0.2
el	ro	6.4	34.6	34.5	19.5	2.4
el	sq	2.1	27.1	28.6	10.9	1.2
en	bg	8.8	45.3	46.2	10.4	4.6
en	el	9.1	57.8	57.4	24.8	3.2
en	it	5.6	36.7	36.5	24.2	0.4
en	ro	10.8	57.1	58.6	52	2.4
en	sl	5.7	22.2	23.3	4.2	1.1
en	sq	3.7	66.6	67.5	29.9	1.8
it	ro	5.3	31.1	30.7	18.1	0.7
ro	sl	5	19.9	20.4	6.2	8.3
ro	sq	3.9	25.7	25.8	9.7	1.5
		6.16	**35.57**	**35.91**	**16.62**	**2.77**

lingual textual semantic similarity methods for constructing a wordnet by including all available definitions.

6.1 Pseudo-retrieval

Table 4 presents the precision at one percentage scores of the retrieval approaches. The machine translation baseline indicates that the task is not trivially solved using a machine translation engine. As the definitions are constructed independently from PWN, machine translation fails to find the correct senses accurately. The weighted word embedding representation for sentences, SEMB is only able to retrieve the correct definitions less than 10% of the time. A significant improvement is observed when WMD is used. For example, while only 3.7% of the definitions are correctly retrieved in the first rank with SEMB for en-sq, it is improved to 67.5% when WMD is used as the similarity metric. This shows the shortcoming of the weighted average sentence representation clearly. The Sinkhorn similarity (SNK) is an extension of WMD. As a retrieval method SNK achieves slightly worse scores compared to WMD.

In order to investigate the effectiveness of the ranking algorithms, we considered the whole retrieval results of the queries instead of taking only the top answer into consideration. Figure 2 shows the number of correct matches and the top k ranking results for Sinkhorn (SNK). The results indicate that most of the correct matches are gathered around higher ranks. Note that a similar pattern is observed for three languages with different ranges of accuracy. Although the $P@1$ score of bg-it is as

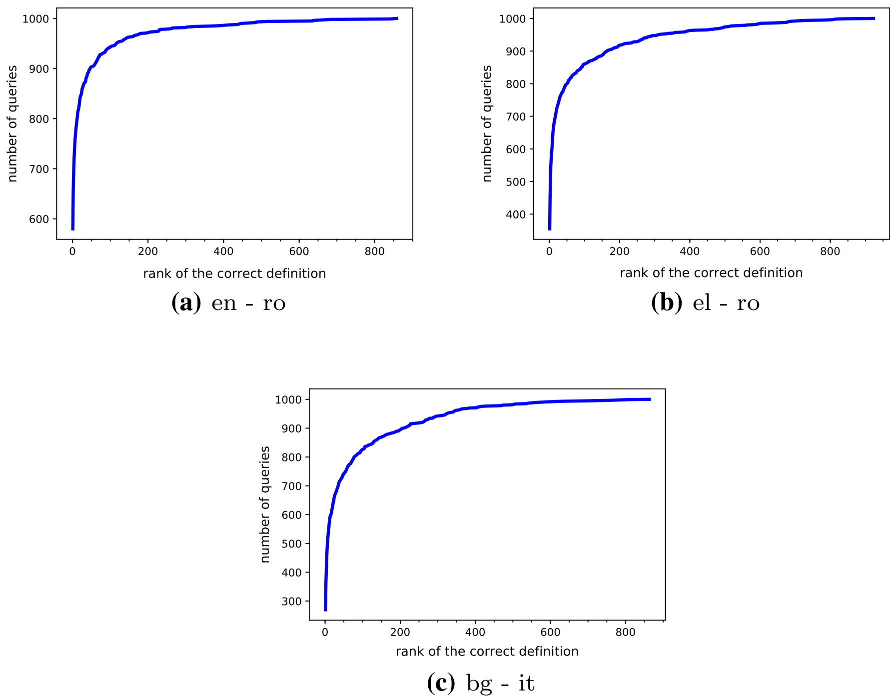

Fig. 2 The plots for the ranks of the correct definitions when retrieved using `SNK` approach for English - Romanian, Greek - Romanian and Bulgarian - Italian

low as 24.4, the correct definitions are gathered in the top 100 ranks. This figure seems to hold for the other similarity measures as well.

When the supervised Siamese LSTM model is used for retrieval, although it achieves higher scores compared to `SEMB` and the MT baseline, it is consistently lower than `WMD` and `SNK`. It should be noted that the Supervised model is trained with limited number of instances available. Table 5 shows the training set size and $P@1$ scores. The performance of the supervised method improves for en-ro, which can be attributed to the 43315 training instances for the language pair. This is also evident from language pairs en-it, en-el and el-ro. One outlier result is between en-sq, which achieves a higher score than other language pairs with similar number of instances. This we believe is related to the construction methodology of Albanian wordnet rather than the number of instances. To overcome the training set size problem, we have performed an additional experiment with the supervised learning method to investigate transfer learning. First, we used multilingual word embeddings (i.e. more than two word embeddings that have been mapped to a shared space). First, we trained our model on a language pair, e.g. en-ro and then tuned the LSTM model for another pair (e.g. en-it). This was to resolve the problem of having less training instances for resource poor languages. As we were not able to improve the results with transfer learning we are not reporting the results of this experiment. We believe that, as LSTM depends on the word order which varies from language to

Table 5 Supervised results

Language		Training samples	Validation samples	P@1%	
Source	Target			Matching	Retrieval
en	bg	2969	990	11.9	10
en	el	12,831	4277	31.5	26.7
en	it	8766	2922	23.8	22.3
en	ro	43,315	14,439	59.6	51.7
en	sl	1632	545	5.8	4.6
en	sq	2756	919	40.1	33.1
bg	el	1362	455	10.1	8.7
bg	it	836	279	10.3	9.1
bg	ro	2775	926	10	8.7
el	it	2850	951	13.9	12.4
el	ro	10,416	3473	18	17.6
el	sq	2748	917	10.2	8
it	ro	7014	2339	21.5	19.5
ro	sl	825	275	7.9	6.7
ro	sq	2736	913	11.5	11.4
				19.07	**16.7**

language, it is not suitable for transfer learning. However, we believe that this idea can be further pursued with different deep learning architectures and further guide methods in pursuit of an effective supervised universal textual similarity method.

6.2 Matching

While most tasks like cross-lingual plagiarism detection and retrieval use the similarity metrics for ranking the documents, the nature of the dictionary alignment allows us to use a matching framework. We believe that this can point to an important shortcoming of the embedding based retrieval methods, known as the hubness problem. Table 6 shows the effectiveness of the methods in a matching framework. The `SEMB` method improves significantly from 6.16% to 38.33% while Supervised method achieves an increase from 16.62% to 19.07%. The same similarity metric when used within a matching framework, achieves significantly higher scores. This indicates to the hubness problem inherited from the word embeddings, *fooling* the algorithm to select definitions that are made of words that are exhibit strong similarity to many other words. Another interesting finding is related to `SNK` and `WMD`, while the latter is more effective in a pseudo-retrieval setting, `SNK` appears to be more effective in matching. We attribute this to the same hubness problem. Considering the entropic regularization in Sinkhorn, it encourages the metric to use more word similarities compared to WMD. This increases the

Table 6 P@1% results of the matching (ϕ_m) approaches

Source Lang	Target Lang	SEMB	SNK	WMD	Supervised
bg	el	31.3	38.5	38	10.1
bg	it	30.2	37.2	35.9	10.3
bg	ro	34.9	38.8	38.1	10
el	it	29.5	37	36.2	13.9
el	ro	37.3	45.9	46.1	18
el	sq	28.6	39.1	37.5	10.2
en	bg	47.3	58.8	57.6	11.9
en	el	58.3	71.7	69.8	31.5
en	it	36.7	43.3	43.4	23.8
en	ro	61.9	71.9	71.3	59.6
en	sl	26.4	34.9	35.4	5.8
en	sq	69.4	80.1	78.1	40.1
it	ro	34.6	42.3	41.4	21.5
ro	sl	22.2	27.1	25.8	7.9
ro	sq	26.4	30.8	29.9	11.5
		38.33	**46.49**	**45.63**	**19.07**

chance of having more noise for hub words in the textual similarity function. These extra similarities not accounted in WMD improve the results.

6.3 Automated wordnet construction

The methods used in this study can be applied to a wordnet creation task. In order to demonstrate this, we have taken all of the available definitions for Bulgarian (bg) and English (en) pairs and used the matching approach. The intuition behind this is that given a set of dictionary definitions, best performing method SNK can align the new dictionary definitions to existing PWN definitions. For 3971 definition pairs available between Bulgarian and English, 1677 (42.23%) have been aligned correctly. While this result is below state-of-the-art wordnet construction methods Khodak et al. (2017); Ercan and Haziyev (2019), it should be noted that it only uses the definitions while the others build on additional features. We believe that these results can be improved by using word-level translations as it is done Khodak et al. (2017), since in our method, to map a single synset all definitions available in wordnet are compared to the target language definition, creating noise for the lemma translation task. This can be reduced to only few candidate definitions, if only possible translations defined in the bilingual dictionary are considered. Furthermore, for a method similar to Ercan and Haziyev (2019), the created wordnets lack definitions in the local language, but are able to map the lemmas with high precision. A complete wordnet with local language glosses is possible when these two methods are used in combination.

7 Conclusions and future work

In this article we have introduced a new dataset to evaluate cross-lingual textual semantic similarity methods. Our dataset has the potential to investigate different measures including supervised and unsupervised methods. Given this dataset state-of-the-art unsupervised methods are evaluated along with a supervised Siamese LSTM network. As the task can be considered both as a retrieval and a matching task, insights about the evaluated methods can be obtained.

Our experiments show that Word Mover's Distance achieves superior results than weighted average of word embeddings and supervised learning algorithm. Furthermore, we show that methods are adversely effected by the hubness problem. WMD appears to be less prone to these effects as it aligns the words and only uses the significant similarities (i.e. with low transportation cost). On the other hand, the Sinkhorn algorithm achieves better results in the matching task showing that this extra information can be valuable in certain settings. We believe that further research can yield better strategies to avoid the hubness problem.

Although our experiments with transfer learning did not improve the effectiveness of the supervised algorithm, we believe that our dataset can be used to build methods that can leverage multi-lingual information better. This evaluation dataset can guide researchers for building universal textual similarity metrics that can be trained with different language pairs and applied to a different language pair with fewer resources.

Acknowledgements This study was supported in part by The Scientific and Technological Research Council of Turkey (TUBITAK), with award no. 114E776.

References

Alonge, A., Bertagna, F., Calzolari, N., Roventini, A. (1999). The italian wordnet. Deliverable D032D033, EuroWordNet.

Alvarez-Melis, D., Jaakkola, T. (2018). Gromov-Wasserstein Alignment of Word Embedding Spaces. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp 1881–1890.

Arora, S., Liang, Y., Ma, T. (2016). A Simple but Tough-to-Beat Baseline for Sentence Embeddings.

Artetxe, M., Labaka, G., Agirre, E. (2018a). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp 789–798.

Artetxe, M., Labaka, G., Agirre, E. (2018b). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. arXiv preprint arXiv:180506297.

Balikas, G., Laclau, C., Redko, I., Amini, M.R. (2018). Cross-lingual Document Retrieval using Regularized Wasserstein Distance. In: Proceedings of the 40th European Conference ECIR conference on Information Retrieval, ECIR 2018, Grenoble, France, March 26-29, 2018.

Balkova, V., Sukhonogov, A., Yablonsky, S. (2004). Russian wordnet. In: Proceedings of the Second Global Wordnet Conference.

Barrón-Cedeño, A., Rosso, P., Agirre, E., Labaka, G. (2010). Plagiarism Detection Across Distant Language Pairs. In: Proceedings of the 23rd International Conference on Computational Linguistics, Association for Computational Linguistics, Stroudsburg, PA, USA, COLING '10, pp 37–45, http://dl.acm.org/citation.cfm?id=1873781.1873786, event-place: Beijing, China.

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, *3*, 1137–1155.

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T. (2016). Enriching word vectors with subword information. arXiv preprint arXiv:160704606.

Bond, F., & Paik, K. (2012). A Survey of WordNets and their Licenses. *GWC 2012 6th International Global Wordnet Conference*, *8*, 64.

Cuturi, M. (2013). Sinkhorn Distances: Lightspeed Computation of Optimal Transportation Distances. arXiv:13060895 [stat] 1306.0895.

Diab, M. (2004). The Feasibility of Bootstrapping an Arabic Wordnet Leveraging Parallel Corpora and an English Wordnet. In: Proceedings of the Arabic Language Technologies and Resources, NEMLAR, Cairo.

Ercan, G., & Haziyev, F. (2019). Synset expansion on translation graph for automatic wordnet construction. *Information Processing & Management*, *56*(1), 130–150.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database. Language, Speech, and Communication*. New York: MIT Press.

Fišer, D., Novak, J., Erjavec, T. (2012). sloWNet 3.0: Development, extension and cleaning. In: Proceedings of 6th International Global Wordnet Conference (GWC 2012), pp. 113–117.

Franco-Salvador, M., Rosso, P., & Montes-y Gómez, M. (2016). A systematic study of knowledge graph analysis for cross-language plagiarism detection. *Information Processing & Management*, *52*(4), 550–570.

Glavas, G., Litschko, R., Ruder, S., Vulic, I. (2019). How to (Properly) Evaluate Cross-Lingual Word Embeddings: On Strong Baselines, Comparative Analyses, and Some Misconceptions. arXiv:190200508 [cs] 1902.00508.

Glorot, X., Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp 249–256.

Gouws, S., Bengio, Y., Corrado, G. (2015). Bilbowa: Fast bilingual distributed representations without word alignments. In: Proceedings of the 32nd International Conference on Machine Learning, http://jmlr.org/proceedings/papers/v37/gouws15.pdf.

Grigoriadou, M., Kornilakis, H., Galiotou, E., Stamou, S., & Papakitsos, E. (2004). The software infrastructure for the development and validation of the Greek WordNet. *Romanian Journal of Information Science and Technology*, *7*(1–2), 89–105.

Hamp, B., Feldweg, H. (1997). Germanet-a lexical-semantic net for german. In: Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications.

Jawanpuria, P., Balgovind, A., Kunchukuttan, A., & Mishra, B. (2019). Learning multilingual word embeddings in latent metric space: A geometric approach. *Transactions of the Association for Computational Linguistics*, *7*, 107–120. https://doi.org/10.1162/tacl_a_00257.

Johnson, A., Karanasou, P., Gaspers, J., Klakow, D. (2019). Cross-lingual transfer learning for japanese named entity recognition. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers), pp. 182–189.

Khodak, M., Risteski, A., Fellbaum, C., Arora, S. (2017). Automated WordNet Construction Using Word Embeddings. In: Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications, pp. 12–23.

Klementiev, A., Titov, I., & Bhattarai, B. (2012). Inducing crosslingual distributed representations of words. *Proceedings of COLING*, *2012*, 1459–1474.

Kusner, M.J., Sun, Y., Kolkin, N.I., Weinberger, K.Q. (2015). From Word Embeddings to Document Distances. In: Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37, JMLR.org, ICML'15, pp. 957–966.

Lam, K.N., Tarouti, F.A., Kalita, J. (2014). Automatically constructing Wordnet Synsets. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp 106–111, https://doi.org/10.3115/v1/P14-2018

Leng, Y., Tan, X., Qin, T., Li, X.Y., Liu, T.Y. (2019). Unsupervised Pivot Translation for Distant Languages. arXiv:190602461 [cs] 1906.02461.

Lison, P., Tiedemann, J. (2016). Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. International Conference on Language Resources and Evaluation.

Litschko, R., Glavaš, G., Ponzetto, S.P., Vulić, I. (2018). Unsupervised cross-lingual information retrieval using monolingual data only. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, ACM, pp 1253–1256.

Luong, T., Pham, H., Manning, C.D. (2015). Bilingual word representations with monolingual quality in mind. In: Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, pp. 151–159.

Mikolov, T., Le, Q.V., Sutskever, I. (2013a). Exploiting Similarities among Languages for Machine Translation. arXiv:13094168 [cs] 1309.4168.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J. (2013b). Distributed Representations of Words and Phrases and their Compositionality. arXiv:13104546 [cs, stat] http://arxiv.org/abs/1310.4546, arXiv: 1310.4546.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., Joulin, A. (2018). Advances in pre-training distributed word representations. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018).

Mogadala, A., Rettinger, A. (2016). Bilingual word embeddings from parallel and non-parallel corpora for cross-language text classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 692–702.

Pascanu, R., Mikolov, T., Bengio, Y. (2012). On the difficulty of training Recurrent Neural Networks. arXiv:12115063 [cs] 1211.5063.

Pedersen, B. S., Nimb, S., Asmussen, J., Sørensen, N. H., Trap-Jensen, L., & Lorentzen, H. (2009). DanNet: The challenge of compiling a wordnet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation*, *43*(3), 269–299. https://doi.org/10.1007/s10579-009-9092-1.

Pennington, J., Socher, R., Manning, C. (2014). Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532–1543.

Potthast, M., Barrón-Cedeño, A., Stein, B., & Rosso, P. (2011). Cross-language plagiarism detection. *Language Resources and Evaluation*, *45*(1), 45–62.

Rubner, Y., Tomasi, C., Guibas, L.J. (1998). A metric for distributions with applications to image databases. In: Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271), pp. 59–66 https://doi.org/10.1109/ICCV.1998.710701.

Ruci, E. (2008). On the current state of Albanet and related applications. Tech. rep., Technical report, University of Vlora.(http://fjalnet. com...).

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

Ruiz-Casado, M., Alfonseca, E., Castells, P. (2005). Automatic assignment of wikipedia encyclopedic entries to wordnet synsets. In: International Atlantic Web Intelligence Conference, Springer, pp 380–386.

Rupnik, J., Muhic, A., Leban, G., Skraba, P., Fortuna, B., & Grobelnik, M. (2016). News across languages-cross-lingual document similarity and event tracking. *Journal of Artificial Intelligence Research*, *55*, 283–316.

Sagot, B., Fišer, D. (2008). Building a free French wordnet from multilingual resources. In: OntoLex.

Sand, H., Velldal, E., Øvrelid, L. (2017). Wordnet extension via word embeddings: Experiments on the Norwegian Wordnet. In: Proceedings of the 21st Nordic Conference on Computational Linguistics, pp. 298–302.

Simov, K.I., Osenova, P. (2010). Constructing of an Ontology-based Lexicon for Bulgarian. In: LREC, Citeseer.

Sinkhorn, R., & Knopp, P. (1967). Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, *21*(2), 343–348.

Stamou, S., Nenadic, G., Christodoulakis, D. (2004). Exploring Balkanet Shared Ontology for Multilingual Conceptual Indexing. In: LREC

Taghizadeh, N., & Faili, H. (2016). Automatic wordnet development for low-resource languages using cross-lingual wsd. *Journal of Artificial Intelligence Research*, *56*, 61–87.

Toral, A., Bracale, S., Monachini, M., Soria, C. (2010). Rejuvenating the Italian WordNet: Upgrading, standardising, extending. In: Proceedings of the 5th International Conference of the Global WordNet Association (GWC-2010), Mumbai

Tufiş, D., Barbu, E., Mititelu, V. B., Ion, R., & Bozianu, L. (2004). The romanian wordnet. *Romanian Journal on Information Science and Technology*, *7*(2–3), 105–122.

Tufiş, D., Ion, R., Bozianu, L., Ceauşu, A., Ştefănescu, D. (2008). Romanian wordnet: Current state, new applications and prospects. In: Proceedings of 4th Global WordNet Conference, GWC, pp. 441–452.

Upadhyay, S., Faruqui, M., Dyer, C., Roth, D. (2016). Cross-lingual Models of Word Embeddings: An Empirical Comparison. arXiv:160400425 [cs] 1604.00425.

Vossen, P. (1998). Introduction to EuroWordNet. *Computers and the Humanities*, *32*(2/3), 73–89.

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

Zeiler, M.D. (2012). ADADELTA: An Adaptive Learning Rate Method. arXiv:12125701 [cs] 1212.5701.