Noname manuscript No.
(will be inserted by the editor)

Cross-lingual document similarity estimation and dictionary generation with comparable corpora

Tadej Štajner · **Dunja Mladenić**

Received: date / Accepted: date

Abstract This paper proposes an approach for performing bilingual dictionary generation even when trained on widely available comparable bilingual corpora. We also show its capability to provide cross-lingual similarity estimates that correlate well with human judgments. We implement an approach using a nonlinear bilingual translation model that we train using comparable corpora. We propose a method using word embeddings and kernel approximation to train scalable non-linear transformations. We demonstrate that this novel method works better on a majority of evaluated language pairs.

Keywords cross-lingual text analysis · vector space machine translation · representation learning · comparable corpora · similarity learning · dictionary generation

1 Introduction

Having multi-lingual knowledge bases assumes that we can express the same concept in different languages. However, getting the correct meaning of a phrase in the first place is a hard problem even in a monolingual setting, without introducing the complication of translating it to another language. We use cross-linguality as a strategy of finding shortcuts to perform natural language processing tasks over multiple language without going to the effort of understanding or fully translating them.

For this paper, we focus on two such tasks: **bilingual dictionary generation** and **cross-lingual similarity estimation**. The purpose of the first is to build bilingual phrase dictionaries between languages that might not yet have such language resources. This task captures the capability of the system to capture (and translate) meanings of individual words. The

T. Štajner
Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Jamova ulica 39, 1000 Ljubljana, Slovenia
E-mail: tadej.stajner@ijs.si

D. Mladenić
Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Jamova ulica 39, 1000 Ljubljana, Slovenia
E-mail: dunja.mladenic@ijs.si

purpose of the second task is to estimate similarity of two documents, each being in a different language. This task captures to what degree the model can tell if two documents in different language express the same meaning, corresponding to its usefulness in cross-lingual information retrieval and recommender systems.

A common denominator in those two tasks (as well as others) is coming up with a representation of text that is amenable to a multi-lingual setting. The bag-of-words model is commonly used as a baseline feature representation. However, we cannot transfer meaning and knowledge easily when we consider words as independent units. This raises a question on whether we can use a more appropriate representation of content.

In this work, we are trying to exploit the following phenomena:

- **Availability of unaligned and weakly-aligned data.** While machine translation literature employs monolingual language models in the decoding phase, we use the local word co-occurrence patterns across large corpora as means to contextualize the data. Also, properly aligned bilingual corpora can be rare, but still have plentiful weakly aligned resources.
- **Non-linearities in cross-lingual relationships.** While a dictionary mapping, a simple example of a linear translation, can provide a reasonable baseline, it often fails to capture contextual nuances that translate poorly across languages. We hypothesize that since these nuances do not translate well when using linear mappings, since word meanings don't always have one-to-one correspondences across languages.

Comparable corpora are collections of documents that have certain common features, such as topic, domain, or genre, but don't have an explicit source-target relationship. In the case of Wikipedia, it contains cross-lingual links among documents of the same topic in different languages. Given these links, we work with the assumption that words that have the same meaning in different languages tend to appear in the same lexical contexts. In contrast with parallel corpora [7], comparable corpora have the following properties:

- Words have multiple senses per corpus
- Words have multiple translations per corpus
- Translations might not exist in the corpus
- Frequencies of occurrence not comparable
- Positions of occurrence not comparable

We consider multilingual corpora, such as Wikipedia as one such resource [23], since parallel corpora are relatively scarce, especially for technical and niche domains, and for language pairs not involving English. This additional relaxation on the input structure constraints allows us greater flexibility in translation or dictionary creation tasks, especially when bootstrapping under-resourced languages.

Besides learning a monolingual representation, cross-linguality also includes learning a cross-lingual mapping. Here, several possibilities exist on how to pose the problem: in the trivial case, if two languages are similar, a simple dictionary mapping already goes far. However, with language pairs with few similarities, the mapping may be more complex. This paper focuses on the latter example, showing that upon learning the right representation, we can use linear models to perform the actual mapping.

The main scientific contributions of this paper are the following:

- A novel method using kernel approximation on word embeddings that outperforms existing baselines on dictionary estimation even with comparable corpora on several language pairs.

- We demonstrate that this novel method works better on a majority of tested language pairs.
- We show that methods based on word embeddings can successfully be used to estimate cross-lingual document similarity.

Throughout the paper, we use the following notation: X and Y represent the vector space representations of source and target language corpora, which are assumed to be aligned at the document level. Conversely, x and y are used to represent individual documents from source and target languages, respectively. The problem is posed as finding a mapping $X \rightarrow Y$ that minimizes a certain error metric.

2 Related Work

Scenarios for vector space machine translation have traditionally relied on machine translation (MT). The typical way of re-using NLP components across languages consists of translating the input content using an MT system, and using the translations as input to the NLP component [6]. This can be applied to problems such as cross-lingual entity linking [3] and sentiment analysis [4], as well as entity tracking, topic detection and cross-lingual recommendation [27].

However, using a full machine translation pipeline can be impractical, since the existence of such as system assumes that there are existing sentence-level alignments that were used to train such a system. However, many languages and vertical domains may not have sufficient aligned corpora for training machine translation systems, but may have access to comparable corpora. Therefore, we explore the consequences and possibilities of using comparable corpora for translation tasks. From the performance viewpoint, machine translation represents an upper bound that is unlikely to be outperformed by simpler approximations on a general domain.

Furthermore, as discussed by [18], Neural Machine Translation approaches implicitly learn a shared cross-lingual embedding space by optimizing for the MT objective, and whereas this thesis focuses on models that explicitly learn cross-lingual word representations also for other use cases. These methods generally do so at a lower cost than MT. In terms of speed and efficiency, they can be considered to be to MT what word embedding models are to language modeling.

2.1 Word embeddings

Recently, many natural language processing tasks started relying on using word embeddings as the underlying representation for solving their problem, such as multilingual classification using word embeddings [10] without requiring word-level alignments. The approach [14] poses the problem of translation as a combination of first learning a low-dimensional representation of the aligned corpora, and then calculating the translation matrix that translates one low-dimensional representation of text to another using a least-squares optimization problem, using a skip-gram model that learn a linear mapping from words to components using neural network learning techniques, maximizing the probability of a word given its neighborhood [15, 13]. Their experiments show that learning a linear mapping by solving a least-squares problem using stochastic gradient descent yields to good results for predicting word translations.

The skip-gram model for a given language is defined in the following fashion: let $w_1, w_2, \ldots, w_T$ represent a sequence of training words. The objective of the Skip-gram model is to maximize the average log-probability of words' contexts within the training corpus of size T given a training window of size k:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{j=-k}^{k}\log p(w_{t+j}|w_t) \tag{1}$$

The probability model uses two parameter vectors for every word: u_w and v_w, which represent the input and output vectors that map from the input to the hidden layer, and from the hidden to the output layer, represented as word-context and context-word matrices U and V. The probability, given a vocabulary of size V is defined as:

$$p(w_i|w_j) = \frac{\exp(u_{w_i}^T v_{w_j})}{\sum_{l=1}^{V}\exp(u_l^T v_{w_j})} \tag{2}$$

These models are trained using stochastic gradient descent, where the gradient is obtained from the propagation rule. The approach also uses a negative sampling technique to generate negative examples by randomly generating k word-context pairs for every real word-context pair w, c using a stochastic gradient descent approach. Literature [11] shows that this approach is equivalent to factorizing a shifted positive point-wise mutual information matrix (SPPMI) into U and V.

$$V^T U = \mathrm{SPPMI}_k(w,c) = \max(\mathrm{PMI}(w,c) - \log k, 0) \tag{3}$$

For our purpose, we are interested in the low-dimensional embedding V, which we can use to multiply with the input vector to obtain the representation the the embedding space. Subsequently, we denote the word-embedding representation of the query document x as $x_{low} = xV^T$.

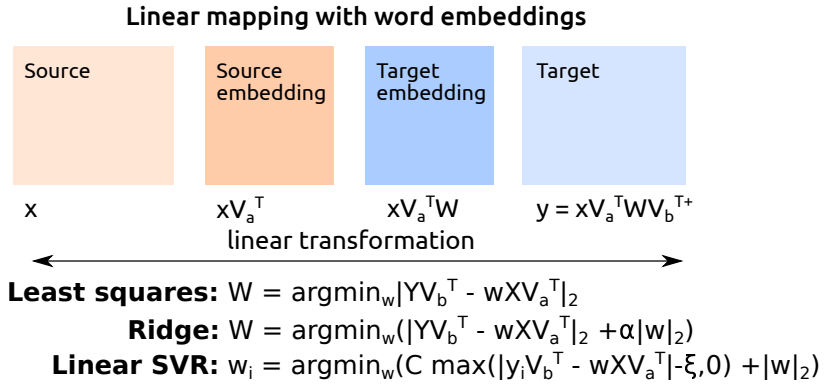

Fig. 1 Integration of distributed word representation and various cross-lingual linear mapping approaches

Let X_{low} and Y_{low} denote low-dimensional word embeddings of two language corpora, obtained as hidden layer representations from either Skip-gram or CBOW models. In the approach in [14], the authors propose to learn a linear mapping W that gives the least squares

solution by minimizing $|Y_{low} - WX_{low}|$. The approach uses a stochastic gradient solver to obtain a solution W, and is denoted with the LSQ-* prefix in our experiments. Figure 1 shows the architecture of this approach.

We can demonstrate this particular transformation by taking a word in the input space, transforming it to the embedding, and mapping it back to the word space. For instance, using the SKIP representation, the word *king* returns as *king, iii, queen, reign, throne* for its top ranking components, all of which are strong indicators of local context. When using representations using word embeddings in other approaches, we denote that using LSQ-SKIP for the skip-grams representation with least-squares mapping to translate between embeddings of different languages.

When discussing use of comparable corpora and embeddings in cross-lingual applications, the work described in [16] demonstrates cross-lingual named entity recognition by using a combination of two approaches: first, an annotation mapping based on comparable corpora, and second, a cross-lingual embedding mapping that is used when re-training the model for another language. The work uses a modified method of LSQ-SKIP as part of their model.

3 Proposed approach

This section presents several ways to improve on LSQ-SKIP. Given the described related work on word embeddings, we propose a novel method for performing fast non-linear approximations for mapping between language vector spaces. We evaluate it on multiple architectures with different representation learning models, as well as translation models.

While linear models can be efficiently estimated, that inherently limits the expressiveness of the model. We hypothesize that we can perform better transformation by first using representation learning with word embeddings on individual languages, but using a non-linear mapping from one language to another instead of a least-squares model.

However, we are still left with the constraint of being efficient on large corpora, scaling to hundreds of millions of tokens and millions of distinct words. Learning a non-linear mapping between two large vector spaces is computationally difficult, which is why we resort to approximations. To that end, we hypothesize that due to the embedding space being robust to small localized noise, we can use a faster, approximate cross-lingual mapping model.

This section describes hypotheses on algorithm design that we evaluated in order to improve on the basic least-squares mapping.

3.1 Monolingual embeddings with a kernel approximation mapping

In order to maintain a usable run time, approaches using embeddings for representation [14] rely on providing a linear projection for translation, potentially limiting performance. We introduce non-linearities among the embedding spaces either by using different kernels in the support vector regression, but the number of data points and output dimensions can often be too big for a quadratic training algorithm. In this subsection, we propose solving the task of mapping between word embeddings using an approach that can efficiently represent non-linearities. We use the Nyström method for kernel approximation [25,26] that reduces the cost of learning with large datasets by using an approximate kernel map. Instead of training a non-linear regression, we can apply an approximate kernel map to our input and use much

more efficient linear solvers downstream to produce the outputs, such as stochastic gradient descent [2].

The Nyström kernel approximation approximates the full kernel matrix K by first sampling m examples, denoted by $\hat{x}_1, \ldots \hat{x}_m$, and then constructs a low rank matrix $\hat{K}_r = K_b K^+ K_b^T$, where $K_b = [\kappa(x_i, \hat{x})]_{N \times m}$, $\hat{K} = [\kappa(\hat{x}_i, \hat{x}_j)]_{m \times m}$, and K^+ is the Moore-Penrose pseudo-inverse of $\hat{K}$ and r is the rank of $\hat{K}$. Having the approximation $\hat{K}$ and a kernel function κ, we define the transformation

$$z(x) = \hat{\Sigma}_r^{1/2} \hat{U}_r (\kappa(x, \hat{x}_1), \ldots, \kappa(x, \hat{x}_m))^T \tag{4}$$

that becomes an m-dimensional representation of the input x. Learning a linear machine that predicts Y using $z(X)$ is equivalent to using the kernel κ directly with a SVM regression that predicts Y directly from X.

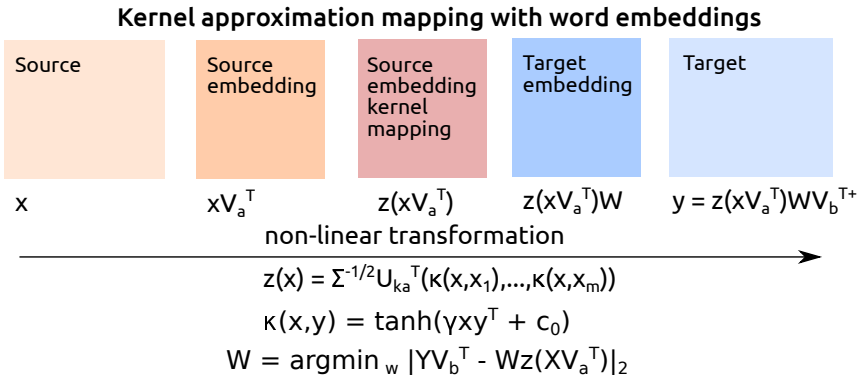

Fig. 2 Integration of kernel approximation as a non-linearity in the cross-lingual translation pipeline. Note that this architecture only supports one direction.

The approximation is achieved by subsampling the data on which the kernel is evaluated, allowing us to make predictions by calculating sigmoid kernel values to those samples, and running a fast SGD-based multivariate ridge regression on the output [24]. Since the running time for kernel approximation is proportional to the number of samples m and number of dimensions d, this allows for linear scalability with respect to corpus size.

Figure 2 shows the approach: first, we translate the training set of document vectors in both languages into their respective embedding representations. Second, using the word embeddings representation of the source language $X_{low} = XV_a^T$, we generate a kernel approximation model around random documents, denoted as $X_{ka} = z(X_{low})$. Finally, we fit a multivariate least squares regression from the first language kernel space to the target language embedding, solving: $|Y_{low} - WX_{ka}|_2$. Given V_a, the non-linear function z, and V_b, we can then produce translations of individual documents vectors, as shown in the equation in 2.

3.2 Monolingual embeddings with a regression tree mapping

In order to assess the compromise of kernel approximation, we also consider other regression methods to achieve the same goals. For instance, using a multivariate regression, composed

of Extremely randomized trees (Extra-Trees) [8], which we denote by EXT-*. While this method has been proposed for being fast to train compared to other decision tree learning approaches, it still has a higher run-time compared to kernel-approximation based methods, we hypothesize that it can also generate improvement over the linear methods. In essence, we train a separate regression tree for every embedding dimension of the target language, given the embeddings of the source language.

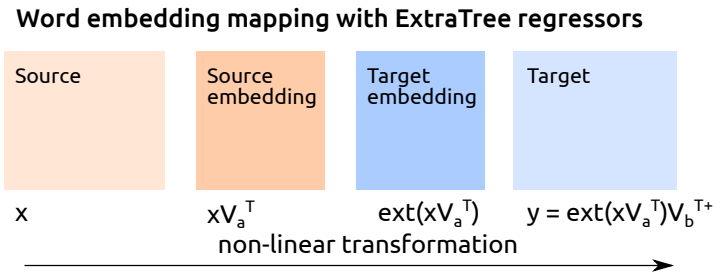

Fig. 3 Multivariate ExtraTree regressors as a non-linear cross-lingual mapping. The *ext* operator represents the multivariate ExtraTree model, which is obtained by concatenating predictors for individual dimensions in the distributional target space.

4 Baselines

For the purpose of comparison with existing methods, we give a short survey of other approaches typically used for similar tasks and that we include in our experiments.

4.1 Cross-lingual Latent Semantic Indexing

One of the early approaches that deal with obtaining a low-dimensional text representation for use in multilingual contexts is Cross-lingual Latent Semantic Indexing (denoted as CLLSI) [5], a method that known to work well for cross-lingual information retrieval [12], leveraging the idea of latent concepts. It is a supervised approach that performs latent semantic indexing on aligned document pairs by concatenating the feature spaces of both languages, weighted using the TF-IDF scheme. It simultaneously works as a translation model as well as a representation model.

CLLSI solves the polysemy and synonymy problems by applying SVD to compute the approximation matrix to the term-document matrix by decomposing the approximation matrix into three matrices, as shown in Figure 4. In this process, the high-dimensional data set is reduced to a lower dimensional vector space, preserving the substructure of the original data while reducing the amount of variations.

4.2 Low-rank Canonical Correlation Analysis

Another family of methods that is frequently encountered in cross-lingual text analysis scenarios is Canonical correlation analysis (CCA) that addresses the following optimization

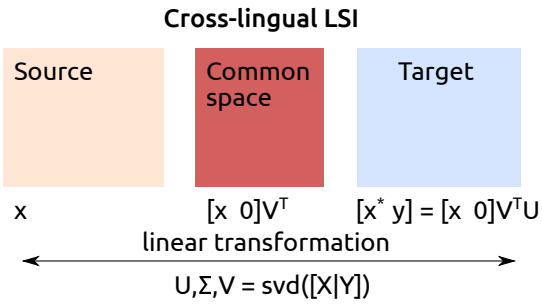

Fig. 4 Architecture of Cross-lingual LSI, showing how the target y is obtained from the input x

problem: given two comparable corpora matrices X and Y, find a pair of linear mappings that maximize the correlation between variables of X and Y. We use those mappings to translate examples from both spaces into a common subspace. The intuition behind Low-rank CCA Translation (LR-CCA) [21] is that the cross-lingual mappings can be more compactly represented in a low-dimensional basis instead of the high-dimensional word spaces, and that simply concatenating the document pairs, as done in approaches, such as CLLSI, can lead to loss of information. Training a LR-CCA model consists of two steps: first, a low-rank SVD decomposition of the co-variance matrix of both aligned corpora, acting as representation learning. Given these transforms, we perform canonical correlation analysis on the low-rank representations of both languages to learn a translation model between them. The input data uses the TF-IDF weighing scheme.

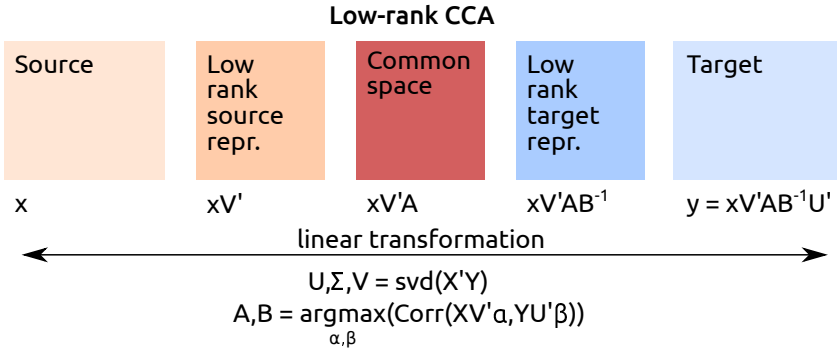

Fig. 5 Architecture of the low rank CCA bidirectional translation pipeline.

Figure 5 illustrates the various representation spaces we can project a document x to. We denote the transformation from the source language word space to the low-rank space with V and U for the source and target languages respectively, and A and B to represent the mapping from both low-rank spaces to the shared representation space. While this method has a stronger model for cross-lingual correlations than CLLSI, it is still limited for tackling polysemy.

4.3 Regression CCA

We also consider other text representations using CCA-based approaches for translation, such as Regression Canonical Correlation Analysis (CCAR). Regression CCA is a method that specializes in computing an accurate mapping for a single input example x_i [19], showing good results on the task of cross-lingual information retrieval.

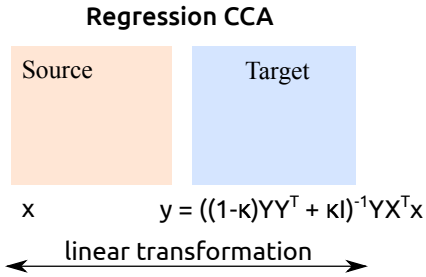

Fig. 6 Illustration of Regression CCA, showing the direct linear transformation expression.

Since the method works on a per-query basis, it works well also on infrequent words, due to the fact that it iteratively optimizes a solution around a given word. On the other hand, the per-query conjugate gradient solver is computationally demanding at prediction time. When mapping the same word across and back, the reconstruction contains terms that are more indicative of the topic. For example, *king* reconstructs as *king, kingdom, royal, reign, son, crown*. While the topical indicators can be beneficial for certain tasks, such as information retrieval, we hypothesize that it may be better to model a finer grained context instead.

5 Experiments

Since the goal of this paper is to provide a recipe for learning an optimal representation and mapping strategy, we evaluate the performance of the methods on the task of dictionary generation on multiple language pairs.

To summarize, we want to test the following hypotheses:

- Do approaches using word embeddings outperform baselines?
- Do non-linear approaches outperform linear approaches?
- Can we can get the same level of performance when using approximate non-linearities approaches compared to exact non-linear regressors?

5.1 Dataset description

We train our models on multiple pairs of language versions of Wikipedia. Since documents in Wikipedia have cross-lingual links that determine correspondence of the same article in other languages, we use those cross-lingual links as alignments of individual data points between two languages. This is known as document-level alignment, which assumes that

the pair of documents has corresponding content. We use the English, Spanish, German, Catalan, Slovene and Croatian versions of Wikipedia. We enumerate the approaches in Table 5.1. The ordering and segmentation of the approaches corresponds to the hypotheses and experiments we wish to perform.

Method name	Representation learning	Cross-lingual mapping
Baselines		
CLLSI [5]	LSI on concatenated language matrices	
LRCCA[21]	SVD on cross-cov. matrix	CCA
CCAR [19]	CCA via conjugate gradient solver	
Word embeddings		
LSQ-SKIP [14]	Skip-gram	Least squares SGD
Approximate non-linear (proposed approach)		
SIGLSQ-SKIP	Skip-gram	Kernel approximation, least squares
Non-approximate non-linear		
EXT-SKIP	Skip-gram	Extremely randomized tree regression

Table 1: An overview of all methods used in this evaluation, showing the differences in representation learning an cross-lingual mapping components, where applicable. The methods' segments represent the hypothesis that they are designed to test.

5.2 Experiments on bilingual dictionary estimation

As a multilingual dictionary resource to evaluate against we consider Wiktionary[1] in its RDF form [9] that has been further linked via crowd-sourcing. For evaluation scenarios using Wiktionary, we use only the language pairs that had more than 500 shared terms in their bilingual dictionary. We use up to 2000 words for every language pair. When pre-processing the data for learning, we cut off low-frequent words, so that the remaining words represent 95% of the weight of the matrix in order to make the training and pre-processing less compute intensive. Changing this ratio did not significantly improve performance. We tuned hyper-parameters of the methods using cross-validation on a separate tuning set of 200 Wiktionary word pairs.

We use the dictionaries as benchmark for our vector space machine translation models. We treat dictionary entry pair as a pair of document vectors. After training the translation models on a comparable bilingual corpus, we ask it to predict a translation of a given word. To evaluate, we measure precision at top-10.

Method	es-ca	es-de	es-en	ca-es	ca-de	de-es	de-ca	de-en	en-es	en-ca	en-de	en-hr
CCAR	0.89	0.03	0.08	0.89	0.01	0.02	0.01	0.02	0.04	0.02	0.01	0.10
CLLSI	0.26	0.13	0.11	0.15	0.05	0.08	0.04	0.08	0.14	0.15	0.04	0.06
LRCCA	0.48	0.31	0.18	0.35	0.11	0.25	0.10	0.20	0.47	0.32	0.10	0.20
LSQ-SKIP	0.62	0.18	0.18	0.50	0.12	0.29	0.09	0.10	0.52	0.21	0.08	0.08
SIGLSQ-SKIP	0.69	0.30	0.20	0.50	0.10	0.30	0.10	0.20	0.60	0.20	0.10	0.20
EXT-SKIP	0.69	0.30	0.20	0.50	0.10	0.30	0.10	0.20	0.60	0.21	0.10	0.19

Table 2: Precision at 10 on dictionary estimation

[1] http://dbpedia.org/Wiktionary

The results in Table 2 exhibit an upper bound with non-linear approaches given that particular combination of training and evaluation datasets. We also observe that performance depends on the density of cross-lingual connections among the language pairs, as well as their corpus size. Languages with smaller datasets and fewer alignments, such as Catalan or Slovenian, performed well when they were the source language, but worse when they were the target language. Some language pairs exhibit poor performance in both directions, suggesting that many the tested words were not successfully captured by the alignment, regardless of the method, which can be a consequence of inaccurate alignments. While computationally expensive, CCAR is the best approach when translating rare terms between language pairs where a sufficiently good linear mapping suffices, like among Spanish and Catalan. Also, this particular language pair has relatively dense cross-lingual links in Wikipedia.

Given our original hypotheses, we observe the following: while LSQ-SKIP shows more robust performance on all language pairs, their performance is not statistically significantly better than the baselines (outside of CLLSI, which is clearly outperformed at $p = 0.001$ on a paired T-test).

Within other experiments that have been omitted for brevity, we observe that there is not any statistically measurable difference by using other regularization strategies or loss functions compared to LSQ-SKIP.

However, when observing non-linear approaches SIGLSQ-SKIP and EXT-SKIP, we observe that they're either equal or better than the linear approaches. While EXT-SKIP does perform better on one language pair, there is no statistically distinguishable difference among them. However, the price in computational requirements is higher: for instance, on the German-English pair with 250 latent components, SIGLSQ-SKIP uses 277 seconds for training, while EXT-SKIP uses 9168 seconds.

5.3 Experiments on cross-lingual similarity estimation

Since many of the tasks involving multiple languages typically involve operating with a similarity metric that works across languages, we also evaluate our methods on the task of approximating cross-lingual similarity. Cross-lingual similarity is a metric in use cases such linking news stories across different languages [20]. For this purpose, we use a comparable corpus across multiple language pairs, annotated with human judgments about the similarity of a given document pair [17]. Recent work on estimating similarity metrics using this particular dataset [1] shows that Wikipedia-specific metrics that also use the link structure perform best. However, they don't generalize to cross-lingual similarity between two arbitrary documents. The authors evaluate their performance by observing the correlation between their approaches and the actual judgment scores. The relevant MT-based approaches in the paper that are comparable to our setup are reported to obtain Spearman-rank correlations of $\rho_{de-en} = 0.47$ and $\rho_{hr-en} = 0.48$.

The dataset has three language pairs that can be tested by our infrastructure: German-English, Croatian-English and Slovene-English. Every language pair has a 100 documents, each having two judgments. Within every judgment, there are four questions, the relevant one being the assessment on how similar a given pair of documents is on a five point scale.

We evaluate the approaches, presented in the previous sections on the task of estimating cross-lingual similarity. The goal is to produce a metric that correlates best with human judgments. For this purpose, we take evaluation document pairs x_a, x_b from the test set, and translate them into $x_a b = f_a b(x_a)$ and $x_b a = f_b a(x_b)$, respectively, where f is the translation

	ρ_{Spearman}			ρ_{Pearson}		
Method	$de-en$	$hr-en$	$sl-en$	$de-en$	$hr-en$	$sl-en$
trans_2 [1]	0.47	0.48				
CCAR	0.516	0.196	0.446	0.467	0.133	0.361
CLLSI	0.506	0.228	0.440	0.472	0.174	0.355
LRCCA	0.471	0.247	0.362	0.452	0.167	0.338
LSQ-SKIP	0.533	0.204	0.450	0.472	0.158	0.355
SIGLSQ-SKIP	0.526	0.206	0.450	0.474	0.159	0.356
EXT-SKIP	0.537	0.208	0.450	0.473	0.159	0.356

Table 3: Spearman and Pearsons's correlation coefficients between the predictions and ground truth similarity judgments

method. Since we can measure similarity in both spaces, we define our similarity metric as the average of both mean absolute errors between the translations and the original:

$$\text{similarity}(x_a, x_b) = \frac{\frac{|x_a - f_{b\to a}(x_a)|}{|x_a|} + \frac{|x_b - f_{a\to b}(x_b)|}{|x_b|}}{2} \tag{5}$$

In Table 3, we observe that a similar ranking among methods holds also for the task of similarity estimation. While the Croatian-English similarity estimates correlate less than the MT-based system in [1] (0.25 versus 0.48 Spearman correlation), the German-English ones correlate better (0.53 compared to 0.47). While the difference between the approaches can be explained by different translation training data, it also suggests that having a comparable corpus of sufficient size (such as the German-English Wikipedia mapping) can perform even better than the trans_2 system that uses an MT system when estimating cross-lingual document similarity.

On the other hand, while we observe that all the approaches using word embeddings represent similarity better than the baselines, there is no significant difference among different linear and non-linear models for translating among word embeddings, suggesting that estimating similarity is a simpler task than translation. Thus, we conclude that for estimating similarity, there is little need to employ non-linear estimators, with both the Spearman and Pearson correlation scores pointing to the same conclusion.

6 Conclusions

We proposed a novel cross-lingual representation learning method for using comparable corpora that uses word embeddings words combined with a kernel approximation mapping that outperforms several baselines while keeping its scalability characteristics.

We show that for the task of dictionary generation, we can not only outperform linear approaches for translating across word embeddings, the approximate methods reach the same level of performance than regression tree translators. Given the results, the most efficient method among the dominating group is SIGLSQ-SKIP, having predictable learning times that are linear with regard to the number of data points, performing comparably to the computationally more expensive EXT-SKIP.

We also show that it performed comparably to the computationally more expensive EXT-SKIP. None of the approaches among those using word embedding approach proved statistically significantly better than the other, but overall, they still outperform the linear approaches on a majority of language pairs. A notable exception is the Catalan-Spanish

language pair, where the CCAR proved to be highly precise when measuring with the Wiktionary bilingual dictionary due to the high relative connectedness among that language pair, and the ability of CCAR to translate rare words.

We show that given the additional constraint of having parallel corpora, we can still perform good dictionary estimation in certain cases, especially where the target language has a sufficiently rich monolingual corpus to train word embeddings on.

We have also demonstrated the presented approaches on the task of similarity estimation. While all approaches using word embeddings outperform baseline approaches, we don't observe comparative benefits when using non-linear mappings to estimate similarity. The results demonstrate that for the similarity estimation task, such a system trained on a comparable corpus can even perform at an equal level of performance to a machine translation system provided that the alignments in the comparable corpus have sufficient density.

Future work in this direction can either be towards applying these approaches on other tasks involving natural language processing across multiple languages, such as cross-lingual textual entailment, as well as tracking news articles across languages. On the other hand, this research opens up new possibilities in using weakly aligned corpora for machine translation in the face of scarce parallel resources [22], as well as investigating the causes behind the discrepancies of performance among different language pairs outside of superficial similarity of languages.

Acknowledgements This work was supported by the Slovenian Research Agency and the IST Programme of the EC under XLike (ICT-STREP-288342), LT-Web (ICT-287815-CSA) and RENDER (ICT-257790-STREP).

References

1. Barrón-Cedeno, A., Paramita, M.L., Clough, P., Rosso, P.: A comparison of approaches for measuring cross-lingual similarity of wikipedia articles. In: ECIR, pp. 424–429 (2014)
2. Bottou, L.: Large-scale machine learning with stochastic gradient descent. In: Y. Lechevallier, G. Saporta (eds.) Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010), pp. 177–187. Springer, Paris, France (2010). URL `http://leon.bottou.org/papers/bottou-2010`
3. Cassidy, T., Ji, H., Deng, H., Zheng, J., Han, J.: Analysis and refinement of cross-lingual entity linking. In: Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics, pp. 1–12. Springer (2012)
4. Duh, K., Fujino, A., Nagata, M.: Is machine translation ripe for cross-lingual sentiment classification? In: ACL (Short Papers), pp. 429–433 (2011)
5. Dumais, S.T., Letsche, T.A., Littman, M.L., Landauer, T.K.: Automatic cross-language retrieval using latent semantic indexing. In: AAAI spring symposium on cross-language text and speech retrieval, vol. 15, p. 21 (1997)
6. Fortuna, B., Shawe-Taylor, J.: The use of machine translation tools for cross-lingual text mining. Learning With Multiple Views, workshop at the ICML (2005)
7. Fung, P.: A statistical view on bilingual lexicon extraction: from parallel corpora to non-parallel corpora. In: Machine Translation and the Information Soup, pp. 1–17. Springer (1998)
8. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Machine learning **63**(1), 3–42 (2006)
9. Hellmann, S., Brekle, J., Auer, S.: Leveraging the crowdsourcing of lexical resources for bootstrapping a linguistic data cloud. In: Semantic Technology, pp. 191–206. Springer (2013)
10. Lauly, S., Boulanger, A., Larochelle, H.: Learning multilingual word representations using a bag-of-words autoencoder. arXiv preprint arXiv:1401.1803 (2014)
11. Levy, O., Goldberg, Y.: Neural word embedding as implicit matrix factorization. In: Advances in Neural Information Processing Systems, pp. 2177–2185 (2014)
12. Littman, M.L., Dumais, S.T., Landauer, T.K.: Automatic cross-language information retrieval using latent semantic indexing. In: Cross-language information retrieval, pp. 51–62. Springer (1998)

13. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
14. Mikolov, T., Le, Q.V., Sutskever, I.: Exploiting similarities among languages for machine translation. CoRR **abs/1309.4168** (2013)
15. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. arXiv preprint arXiv:1310.4546 (2013)
16. Ni, J., Dinu, G., Florian, R.: Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. arXiv preprint arXiv:1707.02483 (2017)
17. Paramita, M.L., Clough, P., Aker, A., Gaizauskas, R.J.: Correlation between similarity measures for inter-language linked wikipedia articles. In: LREC, pp. 790–797 (2012)
18. Ruder, S.: A survey of cross-lingual embedding models. arXiv preprint arXiv:1706.04902 (2017)
19. Rupnik, J., Fortuna, B.: Regression canonical correlation analysis. Learning from (2008)
20. Rupnik, J., Muhic, A., Leban, G., Skraba, P., Fortuna, B., Grobelnik, M.: News across languages-cross-lingual document similarity and event tracking. Journal of Artificial Intelligence Research **55**, 283–316 (2016)
21. Rupnik, J., Muhic, A., Škraba, P.: Low-rank approximations for large, multi-lingual data. Avaliable at http://ailab. ijs. si/primoz_skraba/papers/nips_full. pdf (2011)
22. Skadiņa, I., Aker, A., Mastropavlos, N., Su, F., Tufis, D., Verlic, M., Vasiļjevs, A., Babych, B., Clough, P., Gaizauskas, R., et al.: Collecting and using comparable corpora for statistical machine translation. In: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey (2012)
23. Sorg, P., Cimiano, P.: Exploiting wikipedia for cross-lingual and multilingual information retrieval. Data & Knowledge Engineering **74**, 26–45 (2012)
24. Vedaldi, A., Zisserman, A.: Efficient additive kernels via explicit feature maps. Pattern Analysis and Machine Intelligence, IEEE Transactions on **34**(3), 480–492 (2012)
25. Williams, C., Seeger, M.: Using the nyström method to speed up kernel machines. In: Proceedings of the 14th Annual Conference on Neural Information Processing Systems, EPFL-CONF-161322, pp. 682–688 (2001)
26. Yang, T., Li, Y.F., Mahdavi, M., Jin, R., Zhou, Z.H.: Nyström method vs random fourier features: A theoretical and empirical comparison. In: NIPS, pp. 485–493 (2012)
27. Zhang, L., Rettinger, A., Färber, M., Tadić, M.: A comparative evaluation of cross-lingual text annotation techniques. In: Information Access Evaluation. Multilinguality, Multimodality, and Visualization, pp. 124–135. Springer (2013)