A Resource-Free Evaluation Metric for Cross-Lingual Word Embeddings Based on Graph Modularity

Yoshinari Fujinuma
Computer Science
University of Colorado
fujinumay@gmail.com

Jordan Boyd-Graber
CS, iSchool, UMIACS, LSC
University of Maryland
jbg@umiacs.umd.edu

Michael J. Paul
Information Science
University of Colorado
mpaul@colorado.edu

Abstract

Cross-lingual word embeddings encode the meaning of words from different languages into a shared low-dimensional space. An important requirement for many downstream tasks is that word similarity should be independent of language—i.e., word vectors within one language should not be more similar to each other than to words in another language. We measure this characteristic using *modularity*, a network measurement that measures the strength of clusters in a graph. Modularity has a moderate to strong correlation with three downstream tasks, even though modularity is based only on the structure of embeddings and does not require any external resources. We show through experiments that modularity can serve as an intrinsic validation metric to improve unsupervised cross-lingual word embeddings, particularly on distant language pairs in low-resource settings.[1]

1 Introduction

The success of monolingual word embeddings in natural language processing (Mikolov et al., 2013b) has motivated extensions to cross-lingual settings. Cross-lingual word embeddings—where multiple languages share a single distributed representation—work well for classification (Klementiev et al., 2012; Ammar et al., 2016) and machine translation (Lample et al., 2018; Artetxe et al., 2018b), even with few bilingual pairs (Artetxe et al., 2017) or no supervision at all (Zhang et al., 2017; Conneau et al., 2018; Artetxe et al., 2018a).

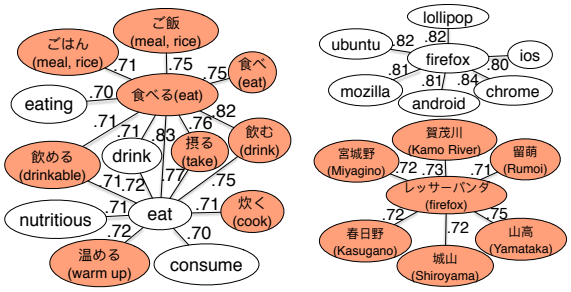
Typically the quality of cross-lingual word embeddings is measured with respect to how well they improve a downstream task. However, sometimes it is not possible to evaluate embeddings for a specific downstream task, for example a future task that does not yet have data or on a rare language that does not have resources to support traditional evaluation. In such settings, it is useful to have an *intrinsic* evaluation metric: a metric that looks at the embedding space itself to know whether the embedding is good *without* resorting to an extrinsic task. While extrinsic tasks are the ultimate arbiter of whether cross-lingual word embeddings work, intrinsic metrics are useful for low-resource languages where one often lacks the annotated data that would make an extrinsic evaluation possible.

However, few intrinsic measures exist for cross-lingual word embeddings, and those that do exist require external linguistic resources (e.g., sense-aligned corpora in Ammar et al. (2016)). The requirement of language resources makes this approach limited or impossible for low-resource languages, which are the languages where intrinsic evaluations are most needed. Moreover, requiring language resources can bias the evaluation toward words in the resources rather than evaluating the embedding space as a whole.

Our solution involves a graph-based metric that considers the characteristics of the embedding space without using linguistic resources. To sketch the idea, imagine a cross-lingual word embedding space where it is possible to draw a hyperplane that separates all word vectors in one language from all vectors in another. Without knowing anything about the languages, it is easy to see that this is a problematic embedding: the representations of the two languages are in distinct parts of the space rather than using a shared space. While this example is exaggerated, this characteristic where vectors are clustered by language often appears within smaller neighborhoods of the embedding space, we want to discover these clusters.

To measure how well word embeddings are mixed across languages, we draw on concepts from network science. Specifically, some cross-

[1] Our code is at https://github.com/akkikiki/modularity_metric



(a) low modularity (b) high modularity

Figure 1: An example of a low modularity (languages mixed) and high modularity cross-lingual word embedding lexical graph using k-nearest neighbors of "eat" (left) and "firefox" (right) in English and Japanese.

lingual word embeddings are *modular* by language: **vectors in one language are consistently closer to each other than vectors in another language** (Figure 1). When embeddings are modular, they often fail on downstream tasks (Section 2).

Modularity is a concept from network theory (Section 3); because network theory is applied to graphs, we turn our word embeddings into a graph by connecting nearest-neighbors—based on vector similarity—to each other. Our hypothesis is that *modularity will predict how useful the embedding is* in downstream tasks; low-modularity embeddings should work better.

We explore the relationship between modularity and three downstream tasks (Section 4) that use cross-lingual word embeddings differently: (i) cross-lingual document classification; (ii) bilingual lexical induction in Italian, Japanese, Spanish, and Danish; and (iii) low-resource document retrieval in Hungarian and Amharic, finding moderate to strong negative correlations between modularity and performance. Furthermore, using modularity as a validation metric (Section 5) makes MUSE (Conneau et al., 2018), an unsupervised model, more robust on distant language pairs. Compared to other existing intrinsic evaluation metrics, modularity captures complementary properties and is more predictive of downstream performance despite needing no external resources (Section 6).

2 Background: Cross-Lingual Word Embeddings and their Evaluation

There are many approaches to training cross-lingual word embeddings. This section reviews the embeddings we consider in this paper, along with existing work on evaluating those embeddings.

2.1 Cross-Lingual Word Embeddings

We focus on methods that learn a cross-lingual vector space through a post-hoc mapping between independently constructed monolingual embeddings (Mikolov et al., 2013a; Vulić and Korhonen, 2016). Given two separate monolingual embeddings and a bilingual seed lexicon, a projection matrix can map translation pairs in a given bilingual lexicon to be near each other in a shared embedding space. A key assumption is that cross-lingually coherent words have "similar geometric arrangements" (Mikolov et al., 2013a) in the embedding space, enabling "knowledge transfer between languages" (Ruder et al., 2017).

We focus on mapping-based approaches for two reasons. First, these approaches are applicable to low-resource languages because they do not requiring large bilingual dictionaries or parallel corpora (Artetxe et al., 2017; Conneau et al., 2018).[2] Second, this focus separates the word embedding task from the cross-lingual mapping, which allows us to focus on evaluating the specific multilingual component in Section 4.

2.2 Evaluating Cross-Lingual Embeddings

Most work on evaluating cross-lingual embeddings focuses on extrinsic evaluation of downstream tasks (Upadhyay et al., 2016; Glavas et al., 2019). However, intrinsic evaluations are crucial since many low-resource languages lack annotations needed for downstream tasks. Thus, our goal is to develop an intrinsic measure that correlates with downstream tasks without using any external resources. This section summarizes existing work on intrinsic methods of evaluation for cross-lingual embeddings.

One widely used intrinsic measure for evaluating the coherence of monolingual embeddings is QVEC (Tsvetkov et al., 2015). Ammar et al. (2016) extend QVEC by using canonical correlation analysis (QVEC-CCA) to make the scores comparable across embeddings with different dimensions. However, while both QVEC and QVEC-CCA can be extended to cross-lingual word embeddings, they are limited: they require external annotated corpora. This is problematic in cross-lingual settings since this requires annotation to be consistent across languages (Ammar et al., 2016).

Other internal metrics do not require external

[2] Ruder et al. (2017) offers detailed discussion on alternative approaches.

resources, but those consider only part of the embeddings. Conneau et al. (2018) and Artetxe et al. (2018a) use a validation metric that calculates similarities of cross-lingual neighbors to conduct model selection. Our approach differs in that we consider whether cross-lingual nearest neighbors are *relatively closer* than intra-lingual nearest neighbors.

Søgaard et al. (2018) use the similarities of intra-lingual neighbors and compute graph similarity between two monolingual lexical subgraphs built by subsampled words in a bilingual lexicon. They further show that the resulting graph similarity has a high correlation with bilingual lexical induction on MUSE (Conneau et al., 2018). However, their graph similarity still only uses intra-lingual similarities but not cross-lingual similarities.

These existing metrics are limited by either requiring external resources or considering only part of the embedding structure (e.g., intra-lingual but not cross-lingual neighbors). In contrast, our work develops an intrinsic metric which is highly correlated with multiple downstream tasks but does not require external resources, and considers both intra- and cross-lingual neighbors.

Related Work A related line of work is the intrinsic evaluation measures of probabilistic topic models, which are another low-dimensional representation of words similar to word embeddings. Metrics based on word co-occurrences have been developed for measuring the monolingual coherence of topics (Newman et al., 2010; Mimno et al., 2011; Lau et al., 2014). Less work has studied evaluation of cross-lingual topics (Mimno et al., 2009). Some researchers have measured the overlap of direct translations across topics (Boyd-Graber and Blei, 2009), while Hao et al. (2018) propose a metric based on co-occurrences across languages that is more general than direct translations.

3 Approach: Graph-Based Diagnostics for Detecting Clustering by Language

This section describes our graph-based approach to measure the intrinsic quality of a cross-lingual embedding space.

3.1 Embeddings as Lexical Graphs

We posit that we can understand the quality of cross-lingual embeddings by analyzing characteristics of a lexical graph (Pelevina et al., 2016; Hamilton et al., 2016). The lexical graph has words as nodes and edges weighted by their similarity in the embedding space. Given a pair of words (i, j) and associated word vectors (v_i, v_j), we compute the similarity between two words by their vector similarity. We encode this similarity in a weighted adjacency matrix A: $A_{ij} \equiv \max(0, \text{cos\_sim}(v_i, v_j))$. However, nodes are only connected to their k-nearest neighbors (Section 6.2 examines the sensitivity to k); all other edges become zero. Finally, each node i has a label g_i indicating the word's language.

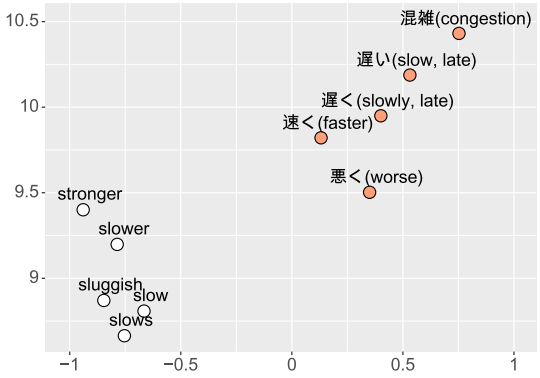


Figure 2: Local t-SNE (van der Maaten and Hinton, 2008) of an EN-JA cross-lingual word embedding, which shows an example of "clustering by language".

3.2 Clustering by Language

We focus on a phenomenon that we call "clustering by language", when word vectors in the embedding space tend to be more similar to words in the same language than words in the other. For example in Figure 2, the intra-lingual nearest neighbors of "slow" have higher similarity in the embedding space than semantically related cross-lingual words. This indicates that words are represented differently across the two languages, thus our hypothesis is that clustering by language degrades the quality of cross-lingual embeddings when used in downstream tasks.

3.3 Modularity of Lexical Graphs

With a labeled graph, we can now ask whether the graph is *modular* (Newman, 2010). In a cross-lingual lexical graph, modularity is the degree to which words are more similar to words in the *same* language than to words in a *different* language. This is undesirable, because the representation of words is not transferred across languages. If the nearest neighbors of the words are instead within the same language, then the languages are not mapped into the cross-lingual space consis-

tently. In our setting, the language l of each word defines its group, and *high* modularity indicates embeddings are more similar *within* languages than *across* languages (Newman, 2003; Newman and Girvan, 2004). In other words, good embeddings should have *low* modularity.

Conceptually, the modularity of a lexical graph is the difference between the proportion of edges in the graph that connect two nodes from the same language and the *expected* proportion of such edges in a randomly connected lexical graph. If edges were random, the number of edges starting from node i within the same language would be the degree of node i, $d_i = \sum_j A_{ij}$ for a weighted graph, following Newman (2004), times the proportion of words in that language. Summing over all nodes gives the expected number of edges within a language,

$$a_l = \frac{1}{2m} \sum_i d_i \mathbb{1}\left[g_i = l\right], \quad (1)$$

where m is the number of edges, g_i is the label of node i, and $\mathbb{1}\left[\cdot\right]$ is an indicator function that evaluates to 1 if the argument is true and 0 otherwise.

Next, we count the fraction of edges e_{ll} that connect words of the same language:

$$e_{ll} = \frac{1}{2m} \sum_{ij} A_{ij} \mathbb{1}\left[g_i = l\right] \mathbb{1}\left[g_j = l\right]. \quad (2)$$

Given L different languages, we calculate overall modularity Q by taking the difference between e_{ll} and a_l^2 for all languages:

$$Q = \sum_{l=1}^{L} (e_{ll} - a_l^2). \quad (3)$$

Since Q does not necessarily have a maximum value of 1, we normalize modularity:

$$Q_{norm} = \frac{Q}{Q_{max}}, \text{where } Q_{max} = 1 - \sum_{l=1}^{L} (a_l^2). \quad (4)$$

The higher the modularity, the more words from the same language appear as nearest neighbors. Figure 1 shows the example of a lexical subgraph with low modularity (left, $Q_{norm} = 0.143$) and high modularity (right, $Q_{norm} = 0.672$). In Figure 1b, the lexical graph is modular since "firefox" does not encode same sense in both languages.

Our hypothesis is that cross-lingual word embeddings with lower modularity will be more successful in downstream tasks. If this hypothesis holds, then modularity could be a useful metric for cross-lingual evaluation.

Language	Corpus	Tokens
English (EN)	News	23M
Spanish (ES)	News	25M
Italian (IT)	News	23M
Danish (DA)	News	20M
Japanese (JA)	News	28M
Hungarian (HU)	News	20M
Amharic (AM)	LORELEI	28M

Table 1: Dataset statistics (source and number of tokens) for each language including both Indo-European and non-Indo-European languages.

4 Experiments: Correlation of Modularity with Downstream Success

We now investigate whether modularity can predict the effectiveness of cross-lingual word embeddings on three downstream tasks: (i) cross-lingual document classification, (ii) bilingual lexical induction, and (iii) document retrieval in low-resource languages. If modularity correlates with task performance, it can characterize embedding quality.

4.1 Data

To investigate the relationship between embedding effectiveness and modularity, we explore five different cross-lingual word embeddings on six language pairs (Table 1).

Monolingual Word Embeddings All monolingual embeddings are trained using a skip-gram model with negative sampling (Mikolov et al., 2013b). The dimension size is 100 or 200. All other hyperparameters are default in Gensim (Řehůřek and Sojka, 2010). News articles except for Amharic are from Leipzig Corpora (Goldhahn et al., 2012). For Amharic, we use documents from LORELEI (Strassel and Tracey, 2016). MeCab (Kudo et al., 2004) tokenizes Japanese sentences.

Bilingual Seed Lexicon For supervised methods, bilingual lexicons from Rolston and Kirchhoff (2016) induce all cross-lingual embeddings except for Danish, which uses Wiktionary.[3]

4.2 Cross-Lingual Mapping Algorithms

We use three supervised (MSE, MSE+Orth, CCA) and two unsupervised (MUSE, VECMAP) cross-lingual mappings:[4]

[3] https://en.wiktionary.org/

[4] We use the implementations from original authors with default parameters unless otherwise noted.

Mean-squared error (MSE) Mikolov et al. (2013a) minimize the mean-squared error of bilingual entries in a seed lexicon to learn a projection between two embeddings. We use the implementation by Artetxe et al. (2016).

MSE with orthogonal constraints (MSE+Orth) Xing et al. (2015) add length normalization and orthogonal constraints to preserve the cosine similarities in the original monolingual embeddings. Artetxe et al. (2016) further preprocess monolingual embeddings by mean centering.[5]

Canonical Correlation Analysis (CCA) Faruqui and Dyer (2014) maps two monolingual embeddings into a shared space by maximizing the correlation between translation pairs in a seed lexicon.

Conneau et al. (2018, MUSE) use language-adversarial learning (Ganin et al., 2016) to induce the initial bilingual seed lexicon, followed by a refinement step, which iteratively solves the orthogonal Procrustes problem (Schönemann, 1966; Artetxe et al., 2017), aligning embeddings without an external bilingual lexicon. Like MSE+Orth, vectors are unit length and mean centered. Since MUSE is unstable (Artetxe et al., 2018a; Søgaard et al., 2018), we report the best of five runs.

Artetxe et al. (2018a, VECMAP) induce an initial bilingual seed lexicon by aligning intra-lingual similarity matrices computed from each monolingual embedding. We report the best of five runs to address uncertainty from the initial dictionary.

4.3 Modularity Implementation

We implement modularity using random projection trees (Dasgupta and Freund, 2008) to speed up the extraction of k-nearest neighbors,[6] tuning $k = 3$ on the German RCV2 dataset (Section 6.2).

4.4 Task 1: Document Classification

We now explore the correlation of modularity and accuracy on cross-lingual document classification. We classify documents from the Reuters RCV1 and RCV2 corpora (Lewis et al., 2004). Documents have one of four labels (Corporate/Industrial, Economics, Government/Social, Markets). We follow Klementiev et al. (2012), except we use all EN training documents and documents in each target language (DA, ES, IT, and JA) as tuning and test data. After removing out-of-vocabulary words, we split documents in target languages into 10% tuning data and 90% test data. Test data are 10,067 documents for DA, 25,566 for IT, 58,950 for JA, and 16,790 for ES. We exclude languages Reuters lacks: HU and AM. We use deep averaging networks (Iyyer et al., 2015, DAN) with three layers, 100 hidden states, and 15 epochs as our classifier. The DAN had better accuracy than averaged perceptron (Collins, 2002) in Klementiev et al. (2012).

[5] One round of iterative normalization (Zhang et al., 2019)

[6] https://github.com/spotify/annoy

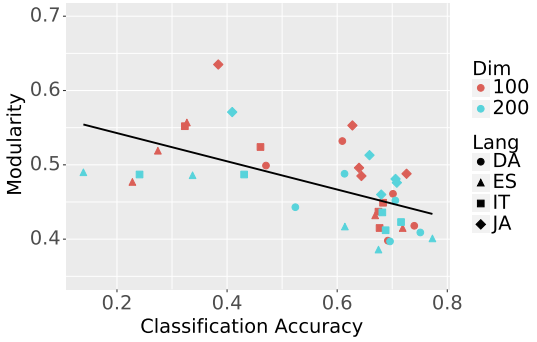


Figure 3: Classification accuracy and modularity of cross-lingual word embeddings ($\rho = -0.665$): less modular cross-lingual mappings have higher accuracy.

	Method	Acc.	Modularity
Supervised	MSE	0.399	0.529
	CCA	0.502	0.513
	MSE+Orth	0.628	0.452
Unsupervised	MUSE	0.711	0.431
	VECMAP	0.643	0.432

Table 2: Average classification accuracy on (EN → DA, ES, IT, JA) along with the average modularity of five cross-lingual word embeddings. MUSE has the best accuracy, captured by its low modularity.

Results We report the correlation value computed from the data points in Figure 3. Spearman's correlation between modularity and classification accuracy on all languages is $\rho = -0.665$. Within each language pair, modularity has a strong correlation within EN-ES embeddings ($\rho = -0.806$), EN-JA ($\rho = -0.794$), EN-IT ($\rho = -0.784$), and a moderate correlation within EN-DA embeddings ($\rho = -0.515$). MUSE has the best classification accuracy (Table 2), reflected by its low modularity.

Error Analysis A common error in EN → JA classification is predicting Corporate/Industrial for documents labeled Markets. One cause is documents with 終値 "closing price"; this has few market-based English neighbors (Table 3). As a result, the model fails to transfer across languages.

市場 “market”	終値 “closing price”
新興 “new coming”	上げ幅 “gains”
market	株価 “stock price”
markets	年初来 “yearly”
軟調 “bearish”	続落 “continued fall”
マーケット “market”	月限 “contract month”
活況 “activity”	安値 “low price”
相場 “market price”	続伸 “continuous rise”
底入 “bottoming”	前日 “previous day”
為替 “exchange”	先物 “futures”
ctoc	小幅 “narrow range”

Table 3: Nearest neighbors in an EN-JA embedding. Unlike the JA word “market”, the JA word “closing price” has no EN vector nearby.

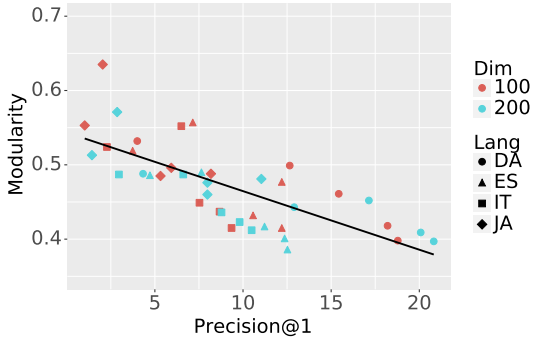


Figure 4: Bilingual lexical induction results and modularity of cross-lingual word embeddings ($\rho = -0.789$): lower modularity means higher precision@1.

4.5 Task 2: Bilingual Lexical Induction (BLI)

Our second downstream task explores the correlation between modularity and bilingual lexical induction (BLI). We evaluate on the test set from Conneau et al. (2018), but we remove pairs in the seed lexicon from Rolston and Kirchhoff (2016). The result is 2,099 translation pairs for ES, 1,358 for IT, 450 for DA, and 973 for JA. We report precision@1 (P@1) for retrieving cross-lingual nearest neighbors by cross-domain similarity local scaling (Conneau et al., 2018, CSLS).

Results Although this task ignores intra-lingual nearest neighbors when retrieving translations, modularity still has a high correlation ($\rho = -0.785$) with P@1 (Figure 4). MUSE and VECMAP beat the three supervised methods, which have the lowest modularity (Table 4). P@1 is low compared to other work on the MUSE test set (e.g., Conneau et al. (2018)) because we filter out translation pairs which appeared in the large training lexicon compiled by Rolston and Kirchhoff (2016), and the raw corpora used to train monolingual embeddings (Table 1) are relatively small compared to Wikipedia.

	Method	P@1	Modularity
Supervised	MSE	7.30	0.529
	CCA	3.06	0.513
	MSE+Orth	10.57	0.452
Unsupervised	MUSE	11.83	0.431
	VECMAP	12.92	0.432

Table 4: Average precision@1 on (EN → DA, ES, IT, JA) along with the average modularity of the cross-lingual word embeddings trained with different methods. VECMAP scores the best P@1, which is captured by its low modularity.

4.6 Task 3: Document Retrieval in Low-Resource Languages

As a third downstream task, we turn to an important task for low-resource languages: lexicon expansion (Gupta and Manning, 2015; Hamilton et al., 2016) for document retrieval. Specifically, we start with a set of EN seed words relevant to a particular concept, then find related words in a target language for which a comprehensive bilingual lexicon does not exist. We focus on the disaster domain, where events may require immediate NLP analysis (e.g., sorting SMS messages to first responders).

We induce keywords in a target language by taking the n nearest neighbors of the English seed words in a cross-lingual word embedding. We manually select sixteen disaster-related English seed words from Wikipedia articles, “*Natural hazard*” and “*Anthropogenic hazard*”. Examples of seed terms include “earthquake” and “flood”. Using the extracted terms, we retrieve disaster-related documents by keyword matching and assess the coverage and relevance of terms by area under the precision-recall curve (AUC) with varying n.

Test Corpora As positively labeled documents, we use documents from the LORELEI project (Strassel and Tracey, 2016) containing any disaster-related annotation. There are 64 disaster-related documents in Amharic, and 117 in Hungarian. We construct a set of negatively labeled documents from the Bible; because the LORELEI corpus does not include negative documents and the Bible is available in all our languages (Christodouloupoulos and Steedman, 2015), we take the chapters of the gospels (89 documents), which do not discuss disasters, and treat these as non-disaster-related documents.

Results Modularity has a moderate correlation with AUC ($\rho = -0.378$, Table 5). While modularity focuses on the entire vocabulary of cross-lingual

Lang.	Method	AUC	Mod.
AM	MSE	0.578	0.628
	CCA	0.345	0.501
	MSE+Orth	0.606	0.480
	MUSE	0.555	0.475
	VECMAP	0.592	0.506
HU	MSE	0.561	0.598
	CCA	0.675	0.506
	MSE+Orth	0.612	0.447
	MUSE	0.664	0.445
	VECMAP	0.612	0.432
Spearman Correlation ρ		-0.378	

Table 5: Correlation between modularity and AUC on document retrieval. It shows a moderate correlation to this task.

word embeddings, this task focuses on a small, specific subset—disaster-relevant words—which may explain the low correlation compared to BLI or document classification.

5 Use Case: Model Selection for MUSE

A common use case of intrinsic measures is model selection. We focus on MUSE (Conneau et al., 2018) since it is unstable, especially on distant language pairs (Artetxe et al., 2018a; Søgaard et al., 2018; Hoshen and Wolf, 2018) and therefore requires an effective metric for model selection. MUSE uses a validation metric in its two steps: (1) the language-adversarial step, and (2) the refinement step. First the algorithm selects an optimal mapping W using a validation metric, obtained from language-adversarial learning (Ganin et al., 2016). Then the selected mapping W from the language-adversarial step is passed on to the refinement step (Artetxe et al., 2017) to re-select the optimal mapping W using the same validation metric after each epoch of solving the orthogonal Procrustes problem (Schönemann, 1966).

Normally, MUSE uses an intrinsic metric, CSLS of the top 10K frequent words (Conneau et al., 2018, CSLS-10K). Given word vectors $s, t \in \mathbb{R}^n$ from a source and a target embedding, CSLS is a cross-lingual similarity metric,

$$\text{CSLS}(Ws, t) = 2\cos(Ws, t) - r(Ws) - r(t) \quad (5)$$

where W is the trained mapping after each epoch, and $r(x)$ is the average cosine similarity of the top 10 cross-lingual nearest neighbors of a word x.

What if we use modularity instead? To test modularity as a validation metric for MUSE, we compute modularity on the lexical graph of 10K most frequent words (Mod-10K; we use 10K for consistency with CSLS on the same words) after each epoch of the adversarial step and the refinement step and select the best mapping.

Family	Lang.	CSLS-10K		Mod-10K	
		Avg.	Best	Avg.	Best
Germanic	DA	**52.62**	**60.27**	52.18	60.13
	DE	**75.27**	**75.60**	75.16	75.53
Romance	ES	**74.35**	83.00	74.32	83.00
	IT	78.41	78.80	**78.43**	78.80
Indo-Iranian	FA	**27.79**	33.40	27.77	33.40
	HI	25.71	33.73	**26.39**	**34.20**
	BN	0.00	0.00	**0.09**	**0.87**
Others	FI	4.71	47.07	4.71	47.07
	HU	**52.55**	54.27	52.35	**54.73**
	JA	18.13	49.69	**36.13**	49.69
	ZH	5.01	37.20	**10.75**	37.20
	KO	16.98	20.68	**17.34**	**22.53**
	AR	15.43	33.33	**15.71**	**33.67**
	ID	67.69	68.40	**67.82**	68.40
	VI	0.01	0.07	0.01	0.07

Table 6: BLI results (P@1 $\times 100\%$) from EN to each target language with different validation metrics for MUSE: default (CSLS-10K) and modularity (Mod-10K). We report the average (Avg.) and the best (Best) from ten runs with ten random seeds for each validation metric. **Bold** values are mappings that are not shared between the two validation metrics. Mod-10K improves the robustness of MUSE on distant language pairs.

The important difference between these two metrics is that Mod-10K considers the relative similarities between intra- and cross-lingual neighbors, while CSLS-10K only considers the similarities of cross-lingual nearest neighbors.[7]

Experiment Setup We use the pre-trained fastText vectors (Bojanowski et al., 2017) to be comparable with the prior work. Following Artetxe et al. (2018a), all vectors are unit length normalized, mean centered, and then unit length normalized. We use the test lexicon by Conneau et al. (2018). We run ten times with the same random seeds and hyperparameters but with different validation metrics. Since MUSE is unstable on distant language pairs (Artetxe et al., 2018a; Søgaard et al., 2018; Hoshen and Wolf, 2018), we test it on English to languages from diverse language families: Indo-European languages such as Danish (DA), German (DE), Spanish (ES), Farsi (FA), Italian (IT), Hindi (HI), Bengali (BN), and non-Indo-European languages such as Finnish (FI), Hungarian (HU), Japanese (JA), Chinese (ZH), Korean (KO), Arabic (AR), Indonesian (ID), and Vietnamese (VI).

[7] Another difference is that k-nearest neighbors for CSLS-10K is $k = 10$, whereas Mod-10K uses $k = 3$. However, using $k = 3$ for CSLS-10K leads to worse results; we therefore only report the result on the default metric i.e., $k = 10$.

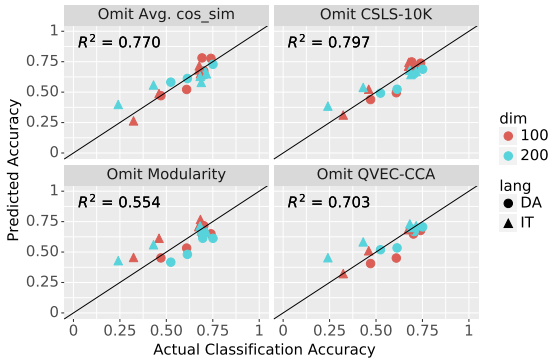


Figure 5: We predict the cross-lingual document classification results for DA and IT from Figure 3 using three out of four evaluation metrics. Ablating modularity causes by far the largest decrease ($R^2 = 0.814$ when using all four features) in R^2, showing that it captures information complementary to the other metrics.

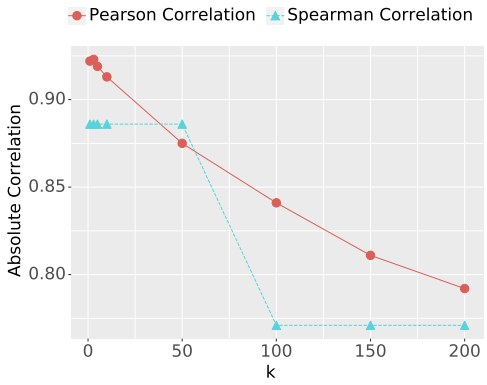


Figure 6: Correlation between modularity and classification performance (EN→DE) with different numbers of neighbors k. Correlations are computed on the same setting as Figure 3 using supervised methods. We use this to set $k = 3$.

Results Table 6 shows P@1 on BLI for each target language using English as the source language. Mod-10K improves P@1 over the default validation metric in diverse languages, especially on the average P@1 for non-Germanic languages such as JA (+18.00%) and ZH (+5.74%), and the best P@1 for KO (+1.85%). These language pairs include pairs (EN-JA and EN-HI), which are difficult for MUSE (Hoshen and Wolf, 2018). Improvements in JA come from selecting a better mapping during the refinement step, which the default validation misses. For ZH, HI, and KO, the improvement comes from selecting better mappings during the adversarial step. However, modularity does not improve on all languages (e.g., VI) that are reported to fail by Hoshen and Wolf (2018).

6 Analysis: Understanding Modularity as an Evaluation Metric

The experiments so far show that modularity captures whether an embedding is useful, which suggests that modularity could be used as an intrinsic evaluation or validation metric. Here, we investigate whether modularity can capture *distinct* information compared to existing evaluation measures: QVEC-CCA (Ammar et al., 2016), CSLS (Conneau et al., 2018), and cosine similarity between translation pairs (Section 6.1). We also analyze the effect of the number of nearest neighbors k (Section 6.2).

6.1 Ablation Study Using Linear Regression

We fit a linear regression model to predict the classification accuracy given four intrinsic measures: QVEC-CCA, CSLS, average cosine similarity of translations, and modularity. We ablate each of the four measures, fitting linear regression with standardized feature values, for two target languages (IT and DA) on the task of cross-lingual document classification (Figure 3). We limit to IT and DA because aligned supersense annotations to EN ones (Miller et al., 1993), required for QVEC-CCA are only available in those languages (Montemagni et al., 2003; Martínez Alonso et al., 2015; Martınez Alonso et al., 2016; Ammar et al., 2016). We standardize the values of the four features before training the regression model.

Omitting modularity hurts accuracy prediction on cross-lingual document classification substantially, while omitting the other three measures has smaller effects (Figure 5). Thus, modularity complements the other measures and is more predictive of classification accuracy.

6.2 Hyperparameter Sensitivity

While modularity itself does not have any adjustable hyperparameters, our approach to constructing the lexical graph has two hyperparameters: the number of nearest neighbors (k) and the number of trees (t) for approximating the k-nearest neighbors using random projection trees. We conduct a grid search for $k \in \{1, 3, 5, 10, 50, 100, 150, 200\}$ and $t \in \{50, 100, 150, 200, 250, 300, 350, 400, 450, 500\}$ using the German RCV2 corpus as the held-out language to tune hyperparameters.

The nearest neighbor k has a much larger effect on modularity than t, so we focus on analyzing the effect of k, using the optimal $t = 450$. Our

earlier experiments all use $k = 3$ since it gives the highest Pearson's and Spearman's correlation on the tuning dataset (Figure 6). The absolute correlation between the downstream task decreases when setting $k > 3$, indicating nearest neighbors beyond $k = 3$ are only contributing noise.

7 Discussion: What Modularity Can and Cannot Do

This work focuses on modularity as a diagnostic tool: it is cheap and effective at discovering which embeddings are likely to falter on downstream tasks. Thus, practitioners should consider including it as a metric for evaluating the quality of their embeddings. Additionally, we believe that modularity could serve as a useful prior for the algorithms that *learn* cross-lingual word embeddings: during learning prefer updates that avoid increasing modularity if all else is equal.

Nevertheless, we recognize limitations of modularity. Consider the following cross-lingual word embedding "algorithm": for each word, select a random point on the unit hypersphere. This is a horrible distributed representation: the position of words' embedding has no relationship to the underlying meaning. Nevertheless, this representation will have very low modularity. Thus, while modularity can identify bad embeddings, once vectors are well mixed, this metric—unlike QVEC or QVEC-CCA—cannot identify whether the meanings make sense. Future work should investigate how to combine techniques that use both word meaning and nearest neighbors for a more robust, semi-supervised cross-lingual evaluation.

Acknowledgments

This work was supported by NSF grant IIS-1564275 and by DARPA award HR0011-15-C-0113 under subcontract to Raytheon BBN Technologies. The authors would like to thank Sebastian Ruder, Akiko Aizawa, the members of the CLIP lab at the University of Maryland, the members of the CLEAR lab at the University of Colorado, and the anonymous reviewers for their feedback. The authors would like to also thank Mozhi Zhang for providing the deep averaging network code. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the view of the sponsor.

References

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith. 2016. Massively multilingual word embeddings. *Computing Research Repository*, arXiv:1602.01925. Version 2.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of Empirical Methods in Natural Language Processing*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the Association for Computational Linguistics*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the Association for Computational Linguistics*.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018b. Unsupervised neural machine translation. In *Proceedings of the International Conference on Learning Representations*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*.

Jordan Boyd-Graber and David M. Blei. 2009. Multilingual topic models for unaligned text. In *Proceedings of Uncertainty in Artificial Intelligence*.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: The Bible in 100 languages. *Proceedings of the Language Resources and Evaluation Conference*.

Michael Collins. 2002. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In *Proceedings of Empirical Methods in Natural Language Processing*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proceedings of the International Conference on Learning Representations*.

Sanjoy Dasgupta and Yoav Freund. 2008. Random projection trees and low dimensional manifolds. In *Proceedings of the annual ACM symposium on Theory of computing*.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the European Chapter of the Association for Computational Linguistics*.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59).

Goran Glavas, Robert Litschko, Sebastian Ruder, and Ivan Vulic. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. *Computing Research Repository*, arXiv:1902.00508. Version 1.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Language Resources and Evaluation Conference*.

Sonal Gupta and Christopher D. Manning. 2015. Distributed representations of words to guide bootstrapped entity classifiers. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of Empirical Methods in Natural Language Processing*.

Shudong Hao, Jordan Boyd-Graber, and Michael J. Paul. 2018. From the Bible to Wikipedia: adapting topic model evaluation to multilingual and low-resource settings. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Yedid Hoshen and Lior Wolf. 2018. Non-adversarial unsupervised word translation. In *Proceedings of Empirical Methods in Natural Language Processing*.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of International Conference on Computational Linguistics*.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of Empirical Methods in Natural Language Processing*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *Proceedings of the International Conference on Learning Representations*.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the European Chapter of the Association for Computational Linguistics*.

David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. 2004. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*.

Héctor Martínez Alonso, Anders Johannsen, Sussi Olsen, Sanni Nimb, Nicolai Hartvig Sørensen, Anna Braasch, Anders Søgaard, and Bolette Sandford Pedersen. 2015. Supersense tagging for Danish. In *Proceedings of the Nordic Conference of Computational Linguistics*.

Héctor Martınez Alonso, Anders Johannsen, Sussi Olsen, Sanni Nimb, and Bolette Sandford Pedersen. 2016. An empirically grounded expansion of the supersense inventory. In *Proceedings of the Global Wordnet Conference*.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation. *Computing Research Repository*, arXiv:1309.4168. Version 1.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of Advances in Neural Information Processing Systems*.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *Proceedings of the Human Language Technology Conference*.

David M. Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Proceedings of Empirical Methods in Natural Language Processing*.

David M. Mimno, Hanna M. Wallach, Edmund M. Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of Empirical Methods in Natural Language Processing*.

Simonetta Montemagni, Francesco Barsotti, Marco Battista, Nicoletta Calzolari, Ornella Corazzari, Alessandro Lenci, Antonio Zampolli, Francesca Fanciulli, Maria Massetani, Remo Raffaelli, Roberto Basili, Maria Teresa Pazienza, Dario Saracino, Fabio Zanzotto, Nadia Mana, Fabio Pianesi, and Rodolfo Delmonte. 2003. Building the Italian syntactic-semantic treebank. In *Treebanks: Building and Using Parsed Corpora*. Springer.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Mark E. J. Newman. 2003. Mixing patterns in networks. *Physical Review E*, 67(2).

Mark E. J. Newman. 2004. Analysis of weighted networks. *Physical Review E*, 70(5).

Mark E. J. Newman. 2010. *Networks: An introduction*. Oxford university press.

Mark E. J. Newman and Michelle Girvan. 2004. Finding and evaluating community structure in networks. *Physical Review E*, 69(2).

Maria Pelevina, Nikolay Arefiev, Chris Biemann, and Alexander Panchenko. 2016. Making sense of word embeddings. In *Proceedings of the 1st Workshop on Representation Learning for NLP*.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.

Leanne Rolston and Katrin Kirchhoff. 2016. Collection of bilingual data for lexicon transfer learning. *UWEE Technical Report*.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2017. A survey of cross-lingual embedding models. *Computing Research Repository*, arXiv:1706.04902. Version 2.

Peter H. Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1).

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the Association for Computational Linguistics*.

Stephanie Strassel and Jennifer Tracey. 2016. LORELEI language packs: Data, tools, and resources for technology development in low resource languages. In *Proceedings of the Language Resources and Evaluation Conference*.

Yulia Tsvetkov, Manaal Faruqui, Wang Ling, Guillaume Lample, and Chris Dyer. 2015. Evaluation of word vector representations by subspace alignment. In *Proceedings of Empirical Methods in Natural Language Processing*.

Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. 2016. Cross-lingual models of word embeddings: An empirical comparison. In *Proceedings of the Association for Computational Linguistics*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Ivan Vulić and Anna Korhonen. 2016. On the role of seed lexicons in learning bilingual word embeddings. In *Proceedings of the Association for Computational Linguistics*.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the Association for Computational Linguistics*.

Mozhi Zhang, Keyulu Xu, Ken-ichi Kawarabayashi, Stefanie Jegelka, and Jordan Boyd-Graber. 2019. Are girls neko or shōjo? Cross-lingual alignment of non-isomorphic embeddings with iterative normalization. In *Proceedings of the Association for Computational Linguistics*.