DIAGNOSING LANGUAGE INCONSISTENCY IN CROSS-LINGUAL WORD EMBEDDINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-lingual embeddings encode meaning of words from different languages into a shared low-dimensional space. However, despite numerous applications, evaluation of such embeddings is limited. We focus on diagnosing the problem of words segregated by languages in cross-lingual embeddings. In an ideal cross-lingual embedding, word similarity should be independent of language—i.e., words within a language should not be more similar to each other than to words in another language. One test of this is *modularity*, a network measurement that measures the strength of clusters in a graph. When we apply this measure to a nearest neighbor graph, imperfect cross-lingual embeddings are sorted into modular, distinct regions. The correlation of this measurement with accuracy on two downstream tasks demonstrates that modularity can serve as an intrinsic metric of embedding quality.

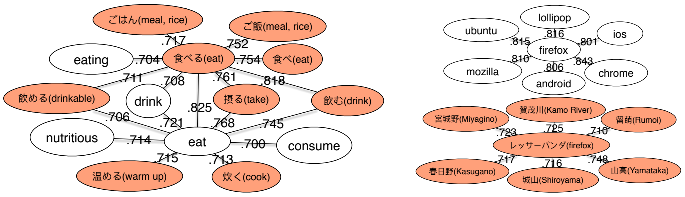
1 INTRODUCTION

The success of monolingual word embeddings in natural language processing (Mikolov et al., 2013b) has encouraged extensions to cross-lingual settings. Cross-lingual embeddings work well for classification (Klementiev et al., 2012; Ammar et al., 2016) and machine translation (Lample et al., 2018; Artetxe et al., 2018), even with few bilingual pairs (Artetxe et al., 2017).

But cross-lingual embeddings are not perfect; Section 2 discusses the ways they can fail to capture meaning across languages. The key underlying assumption for cross-lingual embeddings in many applications is that monolingual embeddings are consistent across language. However, this assumption does not always hold: embeddings can be bad monolingually, senses can be mismatched, or cross-lingual training could fail (Section 4).

We focus on the problems that arise when cross-lingual embeddings are *modular* by language: words in one language only appear next to words of the same language (Figure 1). We can diagnose this problem via graph representations of embeddings (Section 3). We connect vertices (words) based on their similarity; this representation allows us to apply concepts from network science to understand embeddings. Our hypothesis is that modularity can reveal whether an embedding is good or not; low-modularity embeddings should work better in cross-lingual tasks. We make the following contributions:

- We characterize what makes cross-lingual embeddings good or bad, using modularity to summarize the structure of the embedding space: cross-lingual embeddings with high modularity are hypothesized to perform poorly.
- We experimentally explore the relationship between modularity and downstream performance on two tasks: cross-lingual document classification in Italian, Japanese, Spanish, and Danish, and low-resource document retrieval in Hungarian and Amharic, finding strong correlations between modularity and performance ($-.704$ and $-.357$, respectively).
- We analyze the utility of modularity as a metric for evaluating cross-lingual embeddings. It captures complementary information that is more predictive of downstream performance than two existing evaluation metrics.



(a) low modularity

(b) high modularity

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

2 Diagnosing Cross-Lingual Embeddings

It is often helpful to understand the *intrinsic* characteristics of embeddings that make them useful. This section first describes cross-lingual embeddings and the ideal characteristics good embeddings should have. We then discuss three potential problems with cross-lingual embeddings, describing a possible method of identifying each problem intrinsically.

2.1 Background: Cross-Lingual Embeddings

Word embeddings assign a low-dimensional vector for each word given a monolingual corpus. *Cross-lingual* embeddings assign words from different languages into a vector in a shared Euclidean 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).

One approach to building cross-lingual embeddings is to learn a post-hoc mapping between independently constructed monolingual embeddings (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.

Ruder et al. (2017) describe two other approaches to training cross-lingual embeddings: (1) creating an artificial corpus with words from different languages using a bilingual lexical resource, and (2) jointly learning two embeddings for each language. However, we focus on mapping-based approaches because of their applicability to low-resource languages by not requiring large bilingual dictionaries or parallel corpora (Artetxe et al., 2017; Conneau et al., 2018).

2.2 When Do Embeddings Fail?

Good embeddings should be both monolingually coherent and cross-lingually consistent. We now describe three problems that can arise when learning cross-lingual embeddings.

2.2.1 Incoherent Neighbors

The most basic problem in any word embedding is if nearby words in the embedding space are not related in some way, i.e., the embeddings are incoherent. This can happen if the embeddings are suboptimally trained or trained from too little data.

How to diagnose: One widely used intrinsic measure used to evaluate the coherence of monolingual embeddings is QVEC (Tsvetkov et al., 2015). QVEC finds the optimal alignment of each dimension of given a vector derived from an annotated corpus (e.g., "supersenses") and each dimension of a word embedding, then calculates the score as the sum of correlations across all aligned dimensions. QVEC has been extended to use CCA (QVEC-CCA) to output a score in $[-1, 1]$ to make the scores comparable across embeddings with different dimensions (Ammar et al., 2016). However, both QVEC and QVEC-CCA 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), which is a prohibitive restriction for most languages.

2.2.2 Mismatched Senses across Languages

For embeddings that only have a single vector for each word, polysemy can lead to cross-lingual inconsistency. The monolingual embedding of a word in one language sometimes captures a different sense of the word than the monolingual embedding in another language. For example, the English embedding of "firefox" encodes the software sense, while "レッサーパンダ (firefox)" in the Japanese embedding encodes the animal sense (Figure 1).

How to diagnose: A simple method to detect a mismatch across languages is to calculate the cosine similarity between the embeddings of a pair of direct translations (Conneau et al., 2018). A low similarity indicates that they are not cross-lingually consistent. While cross-lingual inconsistency could be caused by many reasons, a common cause of low similarity, even negative similarity, is a mismatch in word sense (Section 5).

2.2.3 Clustering by Language

As a result of inconsistency across languages, even when the sense of translation pairs and its nearest neighbors are matched, intra-lingual words still sometimes cluster together more closely than their cross-lingual counterparts. This is apparent by a word having more intra-lingual nearest neighbors than cross-lingual nearest neighbors. For example, Figure 2 shows that the intra-lingual nearest neighbors of "slow", which are semantically similar or its morphological variants, have higher similarity than cross-lingual words in the embedding space.

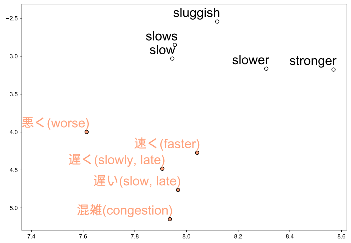


Figure 2: Local t-SNE (van der Maaten and Hinton, 2008) of an EN-JP cross-lingual embedding.

How to diagnose: Existing approaches do not reliably detect this problem. In the next section, we propose a graph-based metric to detect when embeddings exhibit clustering by languages.

3 Graph-Based Diagnostics

We posit that we can understand the quality of cross-lingual embeddings by analyzing characteristics of a lexical graph (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 can compute the similarity between two words by calculating their vector similarity. We encode this similarity in a weighted adjacency matrix A: $A_{ij} = \max(0, \text{cos\_sim}(v_i, v_j))$. However, nodes are only connected to their k-nearest neighbors (Section 5.3 examines the sensitivity to k); all other edges become zero. Finally, each node i has a label g_i indicating the word's language.

3.1 Modularity of Graphs

With a labeled graph, we can now ask whether the graph is *modular* (Newman, 2010, assortative). 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. For example, if two words across two languages have identical meanings in their respective languages, then they should have nearly identical vector representations in an ideal cross-lingual embedding. If the nearest neighbors of the words are instead within the same language, then the languages are not being mapped into the cross-lingual space consistently. 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, where the expected proportion is: given the nodes of the graph, if you randomly connect two nodes, how likely are they to be of the same language? If modularity is positive, it means that nodes within a language are connected more often than would be expected by chance, and therefore the structure is assortative by language.

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,

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

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 are actually connected to the same language:

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

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

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

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

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

The higher the modularity, the more words from the same language appear as nearest neighbors. Figure 1 shows the example of a lexical graph with low modularity (left, $Q_{norm} = 0.152$) 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 embeddings with lower modularity will be more successful at cross-lingual transfer in downstream tasks. If this hypothesis holds, then modularity could be a useful metric for cross-lingual evaluation.

4 EXPERIMENTS: MODULARITY AND DOWNSTREAM SUCCESS

We now investigate whether modularity can predict the effectiveness of cross-lingual embeddings on two downstream tasks: cross-lingual document classification and document retrieval in low-resource languages. If modularity has a strong correlation with task performance, it can effectively characterize the quality of embeddings. To speed up the extraction of k-nearest neighbors, we use random projection trees (Dasgupta and Freund, 2008).[1] We tune k on the German RCV2 dataset, and set $k = 3$. We further discuss the sensitivity of k in Section 5.3.

4.1 EXPERIMENT SETUP: WORD EMBEDDINGS

To investigate the relationship between embedding effectiveness and modularity, we explore four different cross-lingual embedding methods for six different language pairs using a mapping approach with a bilingual dictionary, i.e., the cross-lingual embeddings are trained by learning a mapping between independently trained monolingual embeddings.

Table 1: Dataset statistics.

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

All monolingual embeddings (Table 1) are trained using a skip-gram model with negative sampling (Mikolov et al., 2013b). The dimension size is set to 100 and 200. News articles except for Amharic are from the 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 lexicons from Rolston and Kirchhoff (2016) induce all cross-lingual embeddings for all languages except for Danish, which uses Wiktionary.

We used the following four methods for learning cross-lingual mappings:

[1] `https://github.com/spotify/annoy`

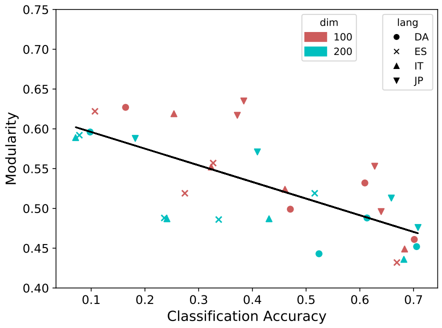


Figure 3: Classification accuracy and modularity of cross-lingual embeddings ($\rho = -.704$).

Table 2: Average classification accuracy on (EN $\rightarrow$ DA, ES, IT, JP) along with the average modularity of the corresponding cross-lingual embeddings trained with different methods. This table contains the same data as Figure 3 to the left, but grouped by training method. MSE+Orth has the highest accuracy, which is captured by its low modularity.

Method	**Acc.**	**Modularity**
Unsupervised	0.166	0.606
MSE	0.399	0.533
CCA	0.502	0.513
MSE+Orth	0.628	0.461

Mean-squared error (MSE) Mikolov et al. (2013a) learn a projection matrix between two embeddings by minimizing the mean-squared error of a bilingual entry in a dictionary. We use the implementation by Artetxe et al. (2016).

MSE with orthogonal constraints (MSE+Orth) The downside of learning a projection matrix without any constraints is that it can change cosine similarity of words in the original monolingual embedding space. To preserve cosine similarities in the original monolingual embedding space, Xing et al. (2015) extend this approach by adding length normalization step to preserve the cosine similarities in the original monolingual embeddings. Artetxe et al. (2016) apply further preprocessing by mean centering the monolingual embeddings before learning the projection matrix.

Canonical Correlation Analysis (CCA) Faruqui and Dyer (2014) use CCA to map two separate monolingual embeddings into a shared space by maximizing the correlation between translation pairs in a dictionary. We use the implementation by Faruqui and Dyer (2014).

Unsupervised Cross-Lingual Embedding (Unsupervised) Unlike the first three methods which use an external bilingual lexicon to train the cross-lingual embeddings, Conneau et al. (2018) use an adversarial approach to align two embedding spaces without using an external bilingual lexicon. We use the implementation by Conneau et al. (2018).

4.2 Task 1: Cross-Lingual Document Classification

We classify documents from the Reuters RCV1 and RCV2 corpora (Lewis et al., 2004). The documents are labeled with one of four categories (Corporate/Industrial, Economics, Government/Social, Markets). We follow Klementiev et al. (2012), but we use all English documents as training data and use all of the documents in each target language as held-out data. After removing out-of-vocabulary words, the documents in each language are split into 10% as tuning data and 90% as test data. The test data in each language contains 10,067 documents for Danish, 25,566 for Italian, 58,950 for Japanese, and 16,790 for Spanish. We exclude HU and AM because Reuters lacks those languages. We use a deep averaging network Iyyer et al. (2015) with three layers, 100 hidden states, and 15 epochs to train a classifier. In preliminary experiments, we found that a deep averaging network resulted in better accuracy compared to an averaged perceptron (Collins, 2002) following Klementiev et al. (2012).

Results Figure 3 shows the relationship between classification accuracy using each embedding and the modularity of the corresponding lexical graphs. The Spearman's correlation between modularity and classification accuracy on all languages is $\rho = -0.704$. Upon computing the correlations within each language pair, we find that modularity has a very strong correlation within EN-JP embeddings ($\rho = -0.881$), a strong correlation within EN-IT ($\rho = -0.731$), and a moderate correlation within EN-ES embeddings ($\rho = -0.707$) and EN-DA embeddings ($\rho = -0.690$). The best classification accuracy was achieved with embeddings trained by MSE+Orth (Table 2), which is reflected by the low modularity of these embeddings.

Error Analysis One example of an error in the EN → JP classification task is a document predicted as "Corporate/Industrial", but labeled as "Markets". One of the keywords in this document "終値 (closing price)" has intra-lingual nearest neighbors (Table 3). This issue is causing failure in the transfer of information across languages.

Table 3: Nearest neighbors in an EN-JP embedding.

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)

4.3 Task 2: Low-Resource Document Retrieval

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

We induce keywords in a target language by taking the nearest neighbors of the English seed words in an cross-lingual embedding. Using the extracted terms, we retrieve disaster-related documents from the annotated LORELEI corpora (Strassel and Tracey, 2016) by keyword matching and assess the coverage and relevance of terms extracted. Specifically, we extract the n nearest neighbors of each seed word, then report the area under the precision-recall curve (AUC) with varying n.

Seed words We select sixteen disaster-related English seed words (see Appendix A), manually selected from the Wikipedia articles, "*Natural hazard*" and "*Anthropogenic hazard*". Examples of seed terms include "earthquake" and "flood".

Labeled corpora As positively labeled documents, we use documents from the LORELEI project tagged as disaster-related, 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 languages we investigate (Christodouloupoulos and Steedman, 2015). Since disasters are discussed in some chapters of the Bible, we took only the chapters of the gospels (89 documents), which do not discuss disasters, and treated these as non-disaster-related documents.

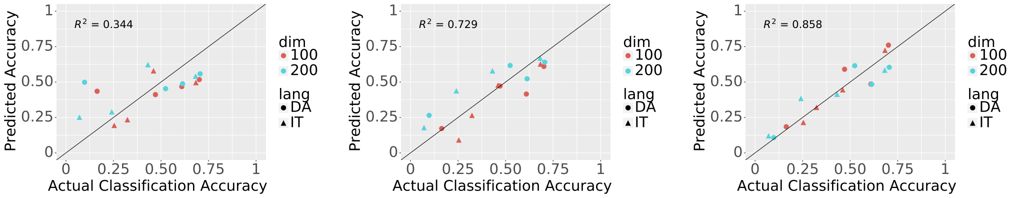
Table 4: Modularity (Mod) and the area under the precision-recall curve (AUC) on document retrieval (EN → AM, HU) using different numbers of cross-lingual nearest neighbors.

Lang.	Method	AUC	Mod
AM	Unsupervised	0.236	0.579
	MSE	0.578	0.628
	CCA	0.345	0.501
	MSE+Orth	0.606	0.480
HU	Unsupervised	0.424	0.620
	MSE	0.561	0.598
	CCA	0.675	0.506
	MSE+Orth	0.612	0.447
Spearman Correlation ρ		-0.357	

Results Modularity has a moderate Spearman's correlation with AUC (Table 4). While modularity focuses on the assortativity of cross-lingual word embeddings over the entire vocabulary, this task is more focused on small, specific subset of words, which may explain why the correlations are lower than for classification.

5 Analysis: Understanding Modularity as an Evaluation Metric

The previous section shows that modularity captures whether an embedding is useful, which suggests that modularity could be used as an intrinsic evaluation metric. Here, we investigate whether modularity can capture *distinct* information compared to existing evaluation measures: QVEC-CCA and cosine similarity between translation pairs (Section 5.1). We also investigate why simpler metrics do not work well (Section 5.2) and analyze the effect of the number of nearest neighbors k on these results (Section 5.3).



(a) Omitting Modularity (b) Omitting QVEC-CCA (c) Omitting average cos_sim

Figure 4: We predict the cross-lingual document classification results for DA and IT from Figure 3 using two out of three embedding evaluation techniques. Ablating modularity causes by far the largest decrease ($R^2 = 0.878$ when using all three features) in R^2, showing that it captures information complementary to the other evaluation metrics.

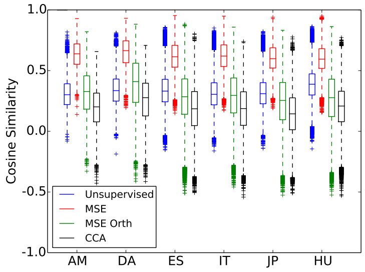


Figure 5: Cosine similarities of translation pairs of different embeddings, where each language is paired with English.

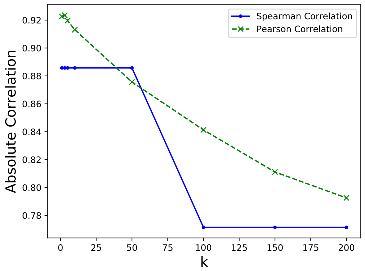


Figure 6: Correlation between modularity and classification performance (EN→DE) with different numbers of neighbors k.

5.1 Ablation Study Using Linear Regression

We fit a linear regression model (see Appendix B) to predict the classification accuracy given three intrinsic measures: QVEC-CCA, average cosine similarity of translations, and modularity. We ablate each of the three measures, fitting linear regression for IT and DA cross-lingual document classification (Figure 3). We limit to IT and DA because aligned supersense annotations to English ones (Miller et al., 1993), required to compute 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).

Omitting modularity hurts the ability to predict the accuracy on cross-lingual document classification substantially, while omitting the other two measures has only a small effect (Figure 4). Thus, modularity is complementary to the other two measures and is strongly predictive of classification accuracy compared to these existing measures.

5.2 Sense-Mismatched Translation Pairs

To gain a better understanding of why measures other than modularity do not predict accuracy well, we examine the cosine similarities of embeddings of direct translations, which one would expect to be high in a good embedding. Surprisingly, some of the translation pairs used in the seed lexicons to create the cross-lingual embeddings have *negative* cosine similarities (Figure 5). Furthermore, Figure 5 and Table 2 both indicate that the cross-lingual embeddings trained by MSE give lower classification accuracy than the ones trained by MSE+Orth, yet the overall cosine similarity between translation pairs is higher in the former. Often, high cosine similarity between translation pairs does not indicate better classification accuracy.

Upon inspection, the most common cause of negative similarities seems to be mismatches in the sense of polysemous words. For example, the pair “eddy” in EN and “āzurīti” (whirlpool) in AM has negative similarity (-0.329) in the EN-AM embedding space. This is because the EN representation of “eddy” has the name sense, and so its nearest cross-lingual neighbor is “michaels” instead of “whirlpool”. Similarly, “firefox” and “レッサーパンダ” (firefox) has low similarity (-0.526) and also has the sense mismatch between EN and JP embeddings.

5.3 Hyperparameter Sensitivity

While modularity itself does not have any adjustable hyperparameters, our method has two hyperparameters, which are the number of nearest neighbors (k) and the number of trees (t) for computing the approximate k-nearest neighbors using random projection trees (Dasgupta and Freund, 2008) when constructing the lexical graph. 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 the hyperparameters. We observed that k had 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 used $k = 3$ since it gives the highest Pearson's and Spearman's correlation on the tuning dataset (Figure 6). Surprisingly, the absolute correlation between the downstream task decreases when setting $k > 3$, indicating nearest neighbors beyond $k = 3$ are only contributing noise.

6 Related Work

One major line of work on evaluating cross-lingual embeddings is comparing their similarity with a fixed set of cross-lingual word pairs rated by humans. In SemEval 2017 Task 2 (Camacho-Collados et al., 2017), correlations between word similarity and human ratings for a fixed set of language pairs evaluate cross-lingual embeddings. Another common way of evaluating a cross-lingual embedding is word translation accuracy using a bilingual lexicon (Upadhyay et al., 2016; Artetxe et al., 2016; 2017; Conneau et al., 2018; Søgaard et al., 2018). Translations are usually retrieved using the nearest cross-lingual neighbor measured by cosine similarity. Since retrieving closest inter-lingual nearest neighbors ignores intra-lingual neighbors, cross-lingual embeddings with words being clustered to its language could still be useful for the word translation task if the closest cross-lingual neighbor is the correct translation. However, for tasks like cross-lingual document classification, we show in Section 4.2 that having more intra-lingual nearest neighbors degrades accuracy.

Less work has focused on intrinsic measures that correlate with downstream tasks. Our work is closest to the work by Søgaard et al. (2018). They compute eigenvalue similarity between two monolingual lexical subgraphs built by subsampled words from two separate *monolingual* embeddings. The resulting eigenvalue similarity has a high correlation with the bilingual lexical induction task on unsupervised cross-lingual embeddings obtained using the method by Conneau et al. (2018). In contrast, our cross-lingual lexical graph is directly derived from *cross-lingual* embeddings. Furthermore, we do not build subgraphs for samples of words, but rather consider the *entire* lexical graph. Finally, we explore the correlation with a classification task, which is a task that requires assumptions about the consistency across languages.

Lastly, a related line of work is the automated evaluation of probabilistic topic models, which are another low-dimensional representation of words and documents. 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.

7 Conclusion

Cross-lingual embeddings are often assortative by language, meaning that words have higher intra-lingual similarity than cross-lingual similarity. Our intrinsic evaluation metric for cross-lingual embeddings based on graph modularity strongly correlates with downstream cross-lingual extrinsic evaluations. While modularity is not a direct measurement of the quality of cross-lingual embeddings, it captures a characteristic of embeddings that is both important for downstream tasks and not captured by other existing intrinsic measures such as QVEC-CCA or cosine similarity on translation pairs. Modularity has an additional advantage over the other two measures in that it does not require external resources, relying only on the structure of the embeddings themselves. We therefore suggest that practitioners should consider modularity when diagnosing and evaluating cross-lingual embeddings in addition to other approaches.

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.

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

Jordan Boyd-Graber and David M. Blei. 2009. Multilingual Topic Models for Unaligned Text. In *Proceedings of Uncertainty in Artificial Intelligence*.

Jose Camacho-Collados, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. 2017. Semeval-2017 task 2: Multilingual and cross-lingual semantic word similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, pages 15–26.

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*.

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*.

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. *CoRR*, abs/1309.4168.

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).

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. *CoRR*, abs/1706.04902.

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 *Conference of the North American Chapter of the Association for Computational Linguistics*.

A Seed Words

Table 5 shows the seed words we use to retrieve disaster-related documents in languages other than English in Section 4.3.

Table 5: Seed Words

criminality	sinkholes
terrorism	blizzard
war	drought
fire	hailstorm
avalanche	tornado
earthquake	flood
lahar	wildfire
landslide	disease

B Linear Regression Fitting

In Section 5.1, we model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$, where y is the cross-lingual classification accuracy on the Reuters corpus, β_i are model parameters, x_1 is the modularity, x_2 is the QVEC-CCA score, x_3 is the average cosine similarity of translation pairs, and ϵ is the error term. All input features are standardized (zero mean, unit variance) before fitting.