Aligning Word Vectors on Low-Resource Languages with Wiktionary

Mike Izbicki
Claremont McKenna College
mike@izbicki.me

Abstract

Aligned word embeddings have become a popular technique for low-resource natural language processing. Most existing evaluation datasets are generated automatically from machine translations systems, so they have many errors and exist only for high-resource languages. We introduce the Wiktionary bilingual lexicon collection, which provides high-quality human annotated translations for words in 298 languages to English. We use these lexicons to train and evaluate the largest published collection of aligned word embeddings on 157 different languages. All of our code and data is publicly available at https://github.com/mikeizbicki/wiktionary_bli.

1 Introduction

A bilingual lexicon is a mapping of words from a source language into a target language. The *bilingual lexicon induction* (BLI) problem is the task of learning such a mapping from data. Most recent solutions to this problem follow a two step procedure: First, train word vectors on a large monolingual corpus for each language individually using a standard algorithm like word2vec (Mikolov et al., 2013a), GloVe (Pennington et al., 2014), or fastText (Bojanowski et al., 2017). Then, learn a transformation that aligns these two vector spaces into a common space (e.g. Mikolov et al., 2013b; Xing et al., 2015; Joulin et al., 2018; Artetxe et al., 2018a; Zhang et al., 2019; Glavaš et al., 2019; Vulić et al., 2019). The BLI problem is then solved by performing nearest neighbor queries in the common space. The focus of this work is the ground truth bilingual lexicon used to train and evaluate these models.

Recent previous work has relied on the MUSE lexicon collection (Conneau et al., 2017). This collection provides bilingual lexicons between 45 languages and English. This lexicon is generated from a machine translation system, and so suffers from a number of problems. First, many of the mappings in the lexicon do not contain real words in either the source or target language (see Figure 1 for examples from Thai). Second, the distribution of words is inconsistent between languages, with many languages containing only proper nouns in their training and test sets. Due to these problems, Kementchedjhieva et al. (2019) suggest that future research "avoids drawing conclusions from quantitative results on this BLI dataset." Other datasets (described in Section 2 below) have even worse limitations.

This paper introduces a new bilingual lexicon collection based on Wiktionary. Wiktionary contains more than 7 million words in 8166 languages and has been collaboratively edited by 3.9 million users.[1] Our specific contributions are:

1. We use Wiktionary to construct high-quality bilingual lexicons suitable for training and evaluating BLI models from 298 languages into English. Most of these languages are extremely low-resource, and many of them are extinct. We provide the first BLI datasets for 253 of these languages, and for the remaining 45 we improve the quality of existing datasets. Our lexicon collection is the first to allow meaningful cross-lingual performance comparisons on the BLI task.

2. We train the largest collection of BLI models to date. Grave et al. (2018) provide pretrained word vectors in 157 languages, and we train BLI models between each of these languages and English. 112 of these languages had not previously had BLI models trained on them because no training/evaluation data previously existed. Of these 112 previously unstudied languages, we identify 15 as having particularly good performance (Armenian, Austurian,

[1] https://en.wiktionary.org/wiki/Wiktionary:Statistics

Thai "Word"	English "Translation"
แคลอรี่	calories
โคมลอย	lanterns
univ	univ
bdfutbol	bdfutbol
efm	efm
พล็อต	plot
getparent	getparent
roca	roca
เป๊ะ	exactly
annie	annie

Figure 1: The last 10 data points for the Thai test files in the widely used MUSE dataset (Conneau et al., 2017). These translation pairs were machine generated without any human input, and this results in bad translation pairs. For example, Thai words should always written in the Thai script, but many words are written in Latin script. Words like `getparent` do not even correspond to words in any natural language and are an artifact of JavaScript code incorrectly included in the original source material. Our Wiktionary dataset contains only high-quality human verified translations and so does not have these problems.

Azerbaijani, Basque, Belarusian, Esperanto, Galician, Georgian, Malayalam, Norwegian Nynorsk, Serbian, Serbo-Croatian, and Welsh) and thus potentially suitable for downstream cross-lingual tasks.

The remainder of this paper is organized as follows. Section 2 discusses related work. Section 3 describes the lexicon construction procedures. Section 4 experimentally demonstrates that the resulting lexicons are of high quality, and trains the new models.

2 Related Work

Applications. Aligned word embeddings have many applications. They are an important component in many document-level translation systems of low-resource languages (Di Gangi and Federico, 2017; Neishi et al., 2017; Artetxe et al., 2017b, 2018b; Qi et al., 2018; Ding and Duh, 2018; Kim et al., 2018; Xia et al., 2019; Font and Costa-Jussa, 2019; Chen and Basirat, 2020). They are also used on non-translation tasks like cross-lingual morphological segmentation (Chimalamarri et al., 2020), dependency parsing (Ahmad et al., 2018), information retrieval (Vulić and Moens, 2015), and document classification (Klementiev et al., 2012; Mogadala and Rettinger, 2016). Our Wiktionary dataset allows better aligned word embeddings to be trained on more languages, allowing all of these "downstream tasks" to be extended into these other languages as well.

Wiktionary. Wiktionary is a valuable resource and widely used by the NLP community. A google scholar search for "Wiktionary" produces 21 000 results on diverse tasks such as synonym detection (Navarro et al., 2009), idiom extraction (Muzny and Zettlemoyer, 2013), and word sense disambiguation (Ben Aouicha et al., 2018). The prior works most closely related to our own are general purpose information extractors (e.g. Acs, 2014; Sérasset, 2015; Nastase and Strapparava, 2015; Kirov et al., 2016; Sajous et al., 2020; Wu and Yarowsky, 2020). Although these extractors can be used to extract translation information, they have not been used explicitly for the purpose of generating datasets for machine translation problems like the BLI problem.

Alternative Datasets. Many datasets have been proposed for the training and evaluation of word vectors. Prior to the MUSE lexicons (Conneau et al., 2017), papers studying BLI all used their own ad-hoc datasets. For example: Mikolov et al. (2013b) introduce Spanish-English and Czech-English lexicons; Dinu and Baroni (2014) introduce an Italian-English lexicon; Artetxe et al. (2017a) introduce German-English and Finish-English lexicons; and Zhang et al. (2017) introduce Spanish-English and Chinese-English lexicons. Since the introduction of MUSE, Glavaš et al. (2019) followed a similar procedure to create an additional 28 bilingual lexicons for high-resource non-English language pairs. Using a machine translation system makes it impossible to create lexicons for low-resource languages without introducing serious mistakes as seen in Figure 1. Furthermore, inter-language comparisons should not be done because the topics covered by the languages' test sets vary considerably (Kementchedjhieva et al., 2019). Our Wiktionary dataset fixes all of these problems.

3 Dataset Overview

In this section, We first describe the data extraction process, then we describe how we split the data into training and test sets. Both steps use language-agnostic approaches. Our goal is to make the data for each language as similar as possible so that cross-lingual evaluations can be made in a fair and consistent manner.

Category	Small	Full
Adjective	50	350
Adverb	25	150
Conjunction	–	25
Determiner	–	25
Interjection	–	25
Noun	125	500
Number	–	50
Pronoun	–	25
Proper noun	–	50
Verb	50	300
Total	250	1500

Table 1: The number of source words of each part of speech for the small and full test sets.

3.1 Data Extraction

Users enter all their data into Wiktionary using the MediaWiki Markdown language. This language is designed primarily for human editors, but contains sufficient semantic annotations to enable machine parsing of entries. We extract all words, also storing the associated language, part of speech, and English-language definition.

Table 2 summarizes the total number of words extracted for selected languages, including a breakdown by part of speech. Many of the languages for which we provide BLI data are now extinct. For example, Ancient Greek has 11381 data points, Old English has 7362, and Tocharian B has 1807.

3.2 Train/Test Splits

In order to train BLI models, we need to split the data extracted above into training and testing sets for each language. We follow the precedent of the MUSE dataset and have the "full test set" contain 1500 words. To facilitate comparison between low resource languages for which it will be difficult to find 1500 meaningful words, we also create a "small test set", which is a subset of the full test set containing 250 words.

We take particular care to construct these test sets so that fair comparisons can be made between languages. In particular, we use the part of speech information extracted from Wiktionary to ensure that each test set has the same number of words in each part of speech. Table 1 shows the number of words. The small test set includes only the "semantic" parts of speech (Adjective, Adverb, Noun, Verb) and not the "syntactic" parts of speech because many low-resource languages lack entries for the syntactic parts of speech and we believe the semantic parts of speech to be more intuitively meaningful.

To populate the small test set, we select the most frequent words from each category. A sampling strategy could result in a harder test set for languages with more words to choose from because they might select less-frequently used words. The remaining words in the large test set are sampled uniformly from the 10 000 most common words for each category. In practice, this allows ranked as high as 20 000 to be included in the test set. This choice makes the Wiktionary test sets significantly harder than the MUSE test sets, which use the 5000-6500 most frequent words regardless of their part of speech.

Finally, we note that not all languages will be able to fully construct test sets according to the procedures above. For example, the Finish lexicon is large (76 375 words), but it only has 17 determiners, and so the final full test set cannot contain 1500 words. This is not due to a defect of the Wiktionary dataset in this language, but just due to the fact that Finish naturally has fewer determiners than other languages. We do not resolve this conflict by adding more words of a different part of speech to the test set, as this would distort the proportions of each part of speech, making the results less comparable. Instead, we simply use a smaller test set. Most languages have a truncated test set due to this effect. Table 3 shows the number of languages with different size test sets. We suggest that meaningful inter-lingual comparisons can be made with models evaluated on 80% of a complete small test set, and so there are 298 languages that can be evaluated using our Wiktionary dataset. Of course many of these languages will have essentially no training data available, and so these languages represent an extreme test-case for unsupervised vector alignment algorithms.

4 Experiments

We perform three experiments. The first experiment measures the importance of the size of the BLI training dataset on model performance. The second experiment compares the quality of the MUSE and Wiktionary lexicons. The final experiment trains BLI models on 112 new, previously unstudied languages.

For all experiments, we align the common crawl vectors provided by Grave et al. (2018) to the English-language vectors trained on the common crawl provided by Mikolov et al. (2018). We use

			Parts of Speech									
Rank	Language	Total	Adj	Adv	Conj	Det	Interj	Noun	Num	Pron	PN	Verb
1	Italian	82 948	22 045	3 799	91	49	123	45 264	108	118	2 809	8 542
2	Finnish	76 375	11 832	3 843	48	17	298	46 631	145	123	1 381	12 057
3	Chinese	75 750	7 142	1 813	192	18	199	43 472	111	387	9 892	12 524
4	Spanish	69 086	17 827	2 605	37	54	201	39 353	53	93	2 488	6 375
5	French	60 692	15 613	2 857	26	25	183	33 444	94	100	2 492	5 858
6	Romanian	54 068	11 873	545	25	38	118	29 537	44	89	7 310	4 489
7	Japanese	47 965	3 052	1 029	94	0	231	32 936	67	225	3 330	7 001
8	German	47 128	11 385	1 071	60	37	141	25 004	242	96	3 116	5 976
9	Serbo-Croatian	47 040	8 793	3 606	92	3	106	24 579	84	231	1 524	8 022
10	Portuguese	41 621	9 428	1 458	33	6	213	22 579	55	75	3 592	4 182
11	Polish	40 096	7 427	1 855	81	0	181	20 939	123	97	1 106	8 287
12	Russian	38 799	7 258	1 467	45	13	215	18 876	52	93	1 680	9 100
13	Dutch	34 716	4 952	792	49	59	161	16 415	105	110	7 539	4 534
14	Macedonian	30 149	7 356	2 681	30	26	91	13 382	53	69	578	5 883
15	Czech	26 958	6 557	702	65	0	133	14 972	45	83	849	3 552
16	Latin	23 155	6 545	1 112	58	19	47	10 074	97	56	2 004	3 143
17	Korean	22 796	790	511	2	97	89	17 814	161	89	1 276	1 967
18	Catalan	22 024	4 528	965	14	19	48	12 266	104	81	1 033	2 966
19	Hungarian	21 660	4 735	967	69	27	143	11 605	215	138	677	3 084
20	Swedish	18 933	3 543	1 002	45	13	88	10 461	145	111	781	2 744
⋮												
101	Zulu	2 208	24	35	15	1	9	1 346	0	42	3	733
102	Volapük	2 194	198	72	20	18	8	1 454	42	48	119	215
103	Basque	2 168	210	59	14	9	18	1 487	31	36	114	190
104	Yoruba	2 165	62	33	10	13	11	1 503	73	38	170	252
105	Westrobothnian	2 107	410	111	15	5	9	879	10	26	5	637
106	Northern Kurdish	2 079	255	42	8	0	5	1 536	21	17	58	137
107	Cimbrian	2 020	199	106	24	13	9	1 095	49	67	23	435
108	Interlingua	2 017	430	60	8	19	7	1 072	24	29	67	301
109	Old Irish	2 013	322	33	33	18	3	1 033	22	42	57	450
110	Egyptian	2 001	67	34	1	25	12	991	21	74	110	666
⋮												
201	Laz	688	20	4	0	0	3	641	7	1	9	3
202	Chechen	686	74	14	3	0	0	498	25	6	17	49
203	Karelian	683	69	13	0	9	0	484	16	29	14	49
204	Tuvan	683	109	27	8	6	4	363	20	27	7	112
205	Low German	683	82	20	8	1	2	487	22	12	10	39
206	Romagnol	683	102	13	3	2	1	415	12	6	12	117
207	Piedmontese	674	118	2	0	0	1	389	28	11	9	116
208	Kavalan	669	63	7	1	0	1	568	11	14	0	4
209	Maquiritari	654	0	72	0	0	3	347	7	27	6	192
210	Zazaki	648	51	15	6	0	3	463	27	21	19	43
⋮												
501	Khaling	127	2	9	1	0	0	76	0	19	1	19
502	Muong	127	17	2	0	0	0	77	11	4	0	16
503	Western Lawa	126	11	1	0	0	1	77	2	1	0	33
504	Picard	126	6	3	0	1	0	77	0	5	3	31
505	Old Marathi	125	17	5	0	0	0	88	0	0	2	13
506	Pohnpeian	125	17	1	1	4	3	70	1	1	1	26
507	Saaroa	123	1	0	0	0	0	121	1	0	0	0
508	Jingpho	121	6	1	0	0	0	81	9	2	0	22
509	Sierra Miwok	120	6	8	1	3	0	88	0	0	0	14
510	Khorezmian Turkic	120	1	1	0	0	0	1	0	0	0	117
⋮												

Table 2: Number of words in the Wiktionary Dataset broken down by their part of speech. Nouns form the bulk every language's vocabulary. The column abbreviations are Adj: Adjective, Adv: Adverb, Conj: Conjunction, Det: Determiner, Interj: Interjection, Num: Number, Pron: Pronoun, PN: Proper Noun.

Percent	Small	Full
100	164	8
90	236	81
80	298	104
70	356	124
60	412	153
50	478	185

Table 3: The "Small" and "Large" columns indicate the number of languages whose completed test set is "Percent" the size that it is supposed to be. For example, only 8 languages can construct a proper full test set with 1500 source words, but 104 languages can construct a full test set with $(80\%)(1500) = 1120$ words.

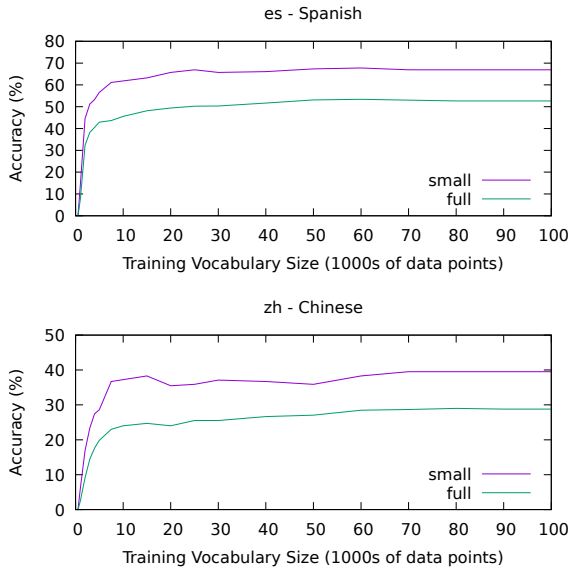

Figure 2: BLI accuracy as a function of dataset size.

the iterative normalization preprocessing procedure (Zhang et al., 2019) to transform both the source and target language vectors before learning. This is different than the most common evaluation setup in the literature, which aligns vectors trained on wikipedia provided by Bojanowski et al. (2017). We use this non-standard setting because our preliminary tests (not shown) found it to give significantly better results for the low-resource languages that we study in Section 4.3 and equivalent results for the high-resource languages.

4.1 Training Dictionary Size

The goal of this experiment is to measure the effect of training dataset size on the performance of supervised BLI models. There are two reasons for performing this experiment. The first is computational. The runtime and memory usage of most BLI training algorithms is proportional to the input training set size. So for languages with large training sets, we want to learn at what size should we truncate the dataset in order to speed up training without sacrificing performance. The second reason is statistical. The size of the training sets we extracted from Wiktionary follow a power law distribution, with a small number of high-resource languages having many translations, but most languages having few translations. We want to understand how having a small training set will effect the BLI performance of these datasets.

To perform the experiment, we construct modified training sets by taking the first n samples from the Wiktionary training set, where n ranges from 0 to 100 000. For each truncated training set, we train the supervised VecMap model (Artetxe et al., 2018a), and evaluate on both the small and full test sets. Results are shown in Figure 2 for the Spanish-English and Chinese-English language pairs. In both cases, BLI accuracy rapidly increases as the number of training samples reaches 5k, and then tapers off. After 20k training points, there is minimal improvement and the performance occasionally decreases due to statistical randomness.[2] This is consistent with previous findings on the effect of training dataset size using the MUSE dataset (Vulić and Korhonen, 2016; Qiu et al., 2018; Glavaš et al., 2019).

In the experiments below, we will train many models. For computational reasons, we truncate the training set size to 20k and expect not to lose any accuracy. We also know that if we observe extremely poor BLI performance in an experiment with at least (about) 5k entries in the BLI training set, then the poor performance is likely not explained by the size of the BLI training set but by some other cause.

4.2 MUSE Corpus vs Wiktionary Corpus

Our next experiment attempts to measure the quality of the MUSE and Wiktionary datasets for the 45 language pairs supported by both datasets. The first three columns of Table 4 show summary statistics of both datasets (details are provided in the table caption). The fourth column is the most interesting, and is the focus of our explanation here.

We train the VecMap (Artetxe et al., 2018a) BLI

[2]Other language pairs are not shown for space reasons, but all had similar results.

model on each language pair, once on the MUSE dataset and once on the Wiktionary dataset. Then for both models, we evaluate on the Wiktionary dataset. The results are shown in the rightmost column of Table 4. Surprisingly, the MUSE training set outperforms the Wiktionary training set for 22/45 of the languages despite coming from a seemingly different distribution. This suggests that despite the high quality nature of the Wiktionary test set data, it is not complete, and more data from more data sources could still be used to improve the alignment of vector spaces.

We hypothesize two reasons to explain this effect. First, the effect only happens when the MUSE training set is much larger than the Wiktionary training set. For example, in the case of Slovak, the MUSE training set has 36 891 data points and the Wiktionary set has only 5 396. The experiments in Section 4.1 above suggest that our Wiktionary dataset's size of about 5k words is large enough to get meaningful results, but that a larger dictionary would still improve performance. Second, Wiktionary is naturally biased towards containing the "dictionary" (i.e. unconjugated) forms of words. Slovak is a fusional language with many inflected forms for each word, and this helps explain the smaller size of the wiktionary dataset.

4.3 The Grave et al. (2018) Languages

Grave et al. (2018) released word vectors in 157 languages trained on the common crawl corpus (a multi-petabyte collection of webpages). All 45 of the languages in the MUSE corpus studied above appear in the Grave et al. (2018) corpus; so in this section we focus on the 112 languages that do not. As far as we are aware, no one has previously attempted to align these embeddings, and there are no previously published datasets of bilingual lexicons suitable for training or evaluation. The Wiktionary corpus is therefore the first publicly available dataset for training and testing alignment models in these languages. The size of each language's dataset and the accuracy for each model on the small and full test set are shown in Table 5.

We train 3 alignment models on each language: the Procrustes (Xing et al., 2015) and Bootstrap Procrustes (Glavaš et al., 2019; Vulić et al., 2019) as implemented by the MUSE project, and VecMap (Artetxe et al., 2018a). There are many other supervised methods and unsupervised methods that would be interesting to train on these datasets, but we did not have the computational resources to do so. Thirteen languages achieve an accuracy on the full test set greater than 30: Esperanto (50.00), Galician (46.62), Armenian (39.15), Azerbaijani (37.38), Georgian (37.30), Austurian (36.92), Basque (36.32), Belarusian (35.75), Welsh (34.84), Malayalam (33.62), Serbo-Croatian (33.17), Norwegian Nynorsk (32.35), and Serbian (30.76). An additional 2 languages achieve an accuracy on the small test set greater than 30: Urdu (37.08) and Mongolian (31.38). We call out the 30% threshold in particular because these languages achieve competitive performance with the languages from the widely used MUSE test set (Table 4), and therefore are good candidates for downstream applications. Because of the careful construction of the test set, as described in Section 3.2 above, it is reasonable to compare the absolute performance between languages. Such comparisons were not recommended for the MUSE dataset (Kementchedjhieva et al., 2019) due to the high variability in quality and content between languages.

We observe that the higher-resource languages (top of table) tend to have better BLI performance than the lower resource languages (bottom of table). We suggest that this difference is not due to a lower quality of the Wiktionary lexicons, but to the lower quality of the Grave et al. (2018) word vectors trained on smaller datasets. We note that in our dictionary size experiment from Section 4.1 above, training lexicons as small as 5k examples give strong performance when the monolingual word vectors are high quality. In Table 5, however, we see performance drop off long before this 5k mark. This is particularly notable in the Latin and Sanskrit languages. Both languages have a large Wiktionary dataset (41 278 and 11 363), but poor BLI performance (13.03 and 2.98 on the full test set). We attribute this to the fact that these languages are of particular interest to the Wiktionary community for their historical importance, and thus have a lot of entries; but their historical nature also means there are few webpages written in these languages, and so the word vectors trained on the common crawl corpus will be of low quality. Word vectors trained on small corpuses are known to be less stable (Pierrejean and Tanguy, 2018; Wendlandt et al., 2018; Leszczynski et al., 2020; Burdick et al., 2021) and therefore difficult to align even with large BLI training data (Vulić et al., 2020).

Source Language		Full Vocab Size		Fraction Distinct		Distinct Vocab Size		BLI Accuracy	
		MUSE	Wikt	MUSE	Wikt	MUSE	Wikt	MUSE	Wikt
af	Afrikaans	**37 421**	4 848	0.30	**0.95**	**11 226**	4 605	**42.13**	35.08
ar	Arabic	**31 355**	26 361	**1.00**	**1.00**	**31 355**	26 361	**31.94**	30.35
bg	Bulgarian	**55 170**	13 827	**1.00**	**1.00**	**55 170**	13 827	48.91	**52.84**
bn	Bengali	**23 829**	5 712	**1.00**	**1.00**	**23 829**	5 712	**28.34**	26.68
bs	Bosnian	43 318	**73 449**	0.38	**0.99**	16 460	**72 714**	**35.95**	29.49
ca	Catalan	78 081	**116 348**	0.30	**0.99**	23 424	**115 184**	**49.79**	49.53
cs	Czech	**64 211**	35 879	0.55	**0.98**	**35 316**	35 161	47.78	**49.67**
da	Danish	**81 959**	16 680	0.46	**0.94**	**37 701**	15 679	49.79	**53.56**
de	German	**101 997**	68 029	0.52	**0.94**	53 038	**63 947**	47.46	**48.88**
el	Greek	**45 515**	32 519	**1.00**	**1.00**	**45 515**	32 519	53.02	**55.45**
es	Spanish	**112 583**	91 066	0.45	**0.95**	50 662	**86 512**	54.40	**54.67**
et	Estonian	**32 776**	6 901	0.64	**0.98**	**20 976**	6 762	**50.04**	48.07
fa	Persian	**41 321**	14 238	**1.00**	**1.00**	**41 321**	14 238	37.39	**39.40**
fi	Finnish	43 102	**105 030**	0.62	**0.99**	26 723	**103 979**	**43.90**	43.11
fr	French	**113 324**	78 837	0.35	**0.90**	39 663	**70 953**	**53.92**	53.57
he	Hebrew	**45 679**	12 234	**1.00**	**1.00**	**45 679**	12 234	33.47	**35.32**
hi	Hindi	**31 046**	21 887	**1.00**	**1.00**	**31 046**	21 887	33.99	**38.28**
hr	Croatian	56 424	**73 449**	0.49	**0.99**	27 647	**72 714**	**47.57**	45.21
hu	Hungarian	**42 823**	34 569	0.62	**0.99**	26 550	**34 223**	45.48	**49.29**
id	Indonesian	**96 518**	12 269	0.30	**0.97**	**28 955**	11 900	35.20	**40.15**
it	Italian	103 613	**119 697**	0.40	**0.98**	41 445	**117 303**	**46.43**	45.47
ja	Japanese	25 969	**73 669**	**1.00**	**1.00**	25 969	**73 669**	**24.96**	24.96
ko	Korean	20 549	**34 739**	**1.00**	**1.00**	20 549	**34 739**	23.84	**31.64**
lt	Lithuanian	**33 435**	6 270	0.55	**1.00**	**18 389**	6 270	**51.22**	49.86
lv	Latvian	**46 385**	14 428	0.72	**1.00**	**33 397**	14 428	50.11	**52.12**
mk	Macedonian	**43 935**	41 054	**1.00**	**1.00**	**43 935**	41 054	37.97	**40.23**
ms	Malay	**73 092**	5 821	0.23	**0.97**	**16 811**	5 646	27.60	**28.56**
nl	Dutch	**93 853**	67 309	0.38	**0.97**	35 664	**65 289**	**39.78**	36.57
no	Norwegian Bokmål	**75 171**	21 386	0.37	**0.95**	**27 813**	20 316	**54.24**	43.96
pl	Polish	**73 901**	66 225	0.48	**0.98**	35 472	**64 900**	**44.18**	41.11
pt	Portuguese	**108 686**	55 927	0.42	**0.95**	45 648	**53 130**	58.42	**58.68**
ro	Romanian	**80 821**	65 122	0.39	**0.93**	31 520	**60 563**	**48.96**	48.58
ru	Russian	48 714	**70 740**	**1.00**	**1.00**	48 714	**70 740**	**46.99**	39.86
sk	Slovak	**65 878**	5 681	0.56	**0.95**	**36 891**	5 396	**54.29**	53.27
sl	Slovene	**62 890**	4 401	0.53	**0.99**	**33 331**	4 356	**49.40**	40.86
sq	Albanian	**52 090**	8 628	0.53	**1.00**	**27 607**	8 628	**36.97**	33.47
sv	Swedish	**82 348**	27 724	0.42	**0.95**	**34 586**	26 337	49.12	**52.82**
ta	Tamil	**21 230**	8 376	**1.00**	**1.00**	**21 230**	8 376	**29.11**	21.20
th	Thai	**25 332**	19 988	0.38	**1.00**	9 626	**19 988**	19.70	**23.33**
tl	Tagalog	**34 984**	17 817	0.28	**0.98**	9 795	**17 460**	28.24	**30.14**
tr	Turkish	**68 611**	15 271	0.42	**0.98**	**28 816**	14 965	34.51	**40.49**
uk	Ukrainian	**40 723**	16 910	**1.00**	**1.00**	**40 723**	16 910	59.10	**59.18**
vi	Vietnamese	**76 364**	9 708	0.08	**1.00**	6 109	**9 708**	11.34	**12.34**
zh	Chinese	21 597	**119 459**	**1.00**	**1.00**	21 597	**119 459**	24.66	**27.78**
Total Best		**35**	10	14	**45**	**24**	21	22	**23**

Table 4: A comparison of the MUSE and Wiktionary datasets. The "Full Vocab Size" measures the total number of source/target word pairs in each dataset. Recall, however, that the MUSE dataset is machine translated, and has many artifacts from this process. One such artifact is the presence of many duplicate entries where the source and target words are the same, and frequently not valid words in either language (See Figure 1 for examples in Thai). The "Fraction Distinct" column measures the fraction of source/target word pairs where the source value does not equal the target. This number is extremely low for many of the MUSE lexicons (e.g. 0.38 for Thai and 0.08 for Vietnamese) due to the machine translation generation process. This number is high for all of the Wiktionary lexicons because they are sourced from high quality human generated translations. All of the duplicate entries are the result of true cognate words between the source language and English. The "Distinct Vocab Size" column computes the total number of distinct source/target pairs in each lexicon, and is equal to the Full Vocab Size column times the Fraction Distinct column. We see that many of the MUSE lexicons are still larger than the Wiktionary lexicons because they allow conjugates of words to appear in a lexicon multiple times, but this does not happen in the Wiktionary lexicon. Finally the "BLI Accuracy" column presents the result of a MUSE-trained model and a Wiktionary trained model on the Wiktionary test set. See Section 4.2 for details.

Rank	Source	Language	Vocab Size	Small Test Set			Full Test Set		
				Proc	Proc-B	VecMap	Proc	Proc-B	VecMap
1	sr	Serbian	73 449	28.51	**47.79**	42.57	19.43	**30.76**	29.27
2	sh	Serbo-Croatian	73 449	42.34	**52.42**	46.37	28.64	**33.17**	31.78
3	la	Latin	41 278	14.11	14.11	**21.37**	9.65	9.65	**13.03**
4	ga	Irish	26 579	24.79	24.79	**29.75**	16.20	16.20	**18.81**
5	hy	Armenian	22 748	50.00	**52.02**	50.40	37.91	38.10	**39.15**
6	nn	Norwegian Nynorsk	19 881	26.53	26.53	**28.98**	30.17	30.17	**33.25**
7	is	Icelandic	19 570	32.52	36.59	**39.43**	29.82	29.91	**34.59**
8	gl	Galician	19 155	47.77	**52.63**	51.82	45.06	45.32	**46.62**
9	ka	Georgian	18 898	36.71	36.71	**42.19**	34.14	34.14	**37.30**
10	eo	Esperanto	18 534	49.36	49.36	**54.94**	47.14	47.14	**50.00**
11	te	Telugu	13 289	14.11	14.11	**15.77**	13.81	13.81	**16.19**
12	gd	Scottish Gaelic	12 443	9.80	9.80	**11.84**	8.32	8.32	**8.53**
13	km	Khmer	11 378	21.54	**26.42**	22.76	16.28	17.56	**18.50**
14	sa	Sanskrit	11 363	**4.08**	4.08	1.22	**2.98**	2.98	1.56
15	kk	Kazakh	11 323	**26.67**	26.67	23.33	27.11	27.11	**27.20**
16	ceb	Cebuano	10 853	**5.88**	5.88	5.88	**8.22**	8.22	3.88
17	az	Azerbaijani	10 713	37.65	**39.27**	38.46	36.09	**37.38**	37.12
18	azb	Southern Azerbaijani	10 713	**2.10**	2.10	1.40	**2.91**	2.91	1.82
19	cy	Welsh	10 459	40.57	**46.31**	44.26	30.98	34.34	**34.84**
20	io	Ido	8 127	**14.00**	14.00	14.00	12.30	12.30	**12.75**
21	gv	Manx	8 105	**5.93**	5.93	3.39	**6.39**	6.39	3.27
22	mt	Maltese	8 089	**0.00**	0.00	0.00	**0.00**	0.00	0.00
23	ml	Malayalam	7 465	27.78	28.63	**30.34**	29.60	32.47	**33.62**
24	lb	Luxembourgish	7 438	4.88	**8.54**	5.69	10.89	**11.99**	6.13
25	sw	Swahili	7 324	12.45	12.86	**15.35**	15.94	**18.01**	17.49
26	ur	Urdu	7 013	26.25	**37.08**	26.25	23.64	**29.73**	24.81
27	yi	Yiddish	6 869	4.86	4.86	**5.26**	7.79	7.79	**9.79**
28	my	Burmese	5 902	13.52	**16.80**	13.11	13.77	**15.90**	15.26
29	ast	Asturian	5 645	27.98	30.86	**33.33**	30.12	33.65	**36.92**
30	bcl	Bikol Central	5 069	**6.22**	6.22	4.98	**5.16**	5.16	3.53
31	be	Belarusian	4 598	27.92	**30.00**	27.92	32.81	**35.75**	33.92
32	mn	Mongolian	4 470	21.34	**31.38**	19.67	21.10	**27.97**	19.14
33	as	Assamese	4 341	1.28	1.28	**1.70**	4.49	4.49	**4.72**
34	oc	Occitan	4 317	16.96	18.70	**20.87**	20.26	**22.23**	22.23
35	gu	Gujarati	4 068	10.78	**18.53**	10.34	10.85	**16.81**	12.51
36	ba	Bashkir	4 053	7.00	7.00	**9.47**	9.03	9.03	**9.54**
37	sco	Scots	3 734	**9.09**	9.09	3.54	**10.82**	10.82	8.19
38	mg	Malagasy	3 634	**5.26**	5.26	4.78	**4.05**	4.05	3.64
39	vec	Venetian	3 570	**3.24**	3.24	2.43	**3.94**	3.94	3.48
40	yo	Yoruba	3 536	**1.34**	1.34	0.00	**0.87**	0.87	0.00
41	bo	Tibetan	3 438	**1.02**	1.02	0.00	**1.48**	1.48	0.00
42	sah	Yakut	3 265	**2.87**	2.87	0.82	**4.00**	4.00	2.83
43	qu	Quechua	3 190	**4.13**	4.13	0.00	**3.60**	3.60	0.00
44	eu	Basque	3 117	20.89	**38.67**	23.56	25.42	**36.32**	23.48
45	mr	Marathi	3 016	11.02	**22.88**	13.98	13.73	**19.82**	12.85
46	pnb	Western Punjabi	2 692	0.00	**0.52**	0.00	0.00	**0.37**	0.00
47	pa	Punjabi	2 692	**4.66**	4.66	2.54	**5.35**	5.35	3.36
48	vo	Volapük	2 673	**2.98**	2.98	0.85	**3.92**	3.92	2.04
49	ku	Northern Kurdish	2 658	**2.87**	1.64	2.87	5.52	**5.79**	3.77
50	rm	Romansch	2 479	**2.03**	2.03	1.22	**5.64**	5.64	4.05
51	ia	Interlingua	2 397	7.29	**10.12**	4.45	8.25	**9.68**	5.22
52	ne	Nepali	2 185	7.88	**10.37**	2.07	10.11	**10.51**	4.65
53	fy	West Frisian	2 137	7.50	**10.83**	4.17	10.48	**14.06**	5.41
54	scn	Sicilian	2 050	**2.86**	2.86	1.63	**5.41**	5.41	2.85
55	ug	Uyghur	1 919	**1.72**	1.72	0.00	**3.78**	3.78	0.15
56	als	Alemannic German	1 888	**0.50**	0.50	0.00	**2.64**	2.64	0.00
57	uz	Uzbek	1 845	**8.06**	8.06	7.11	**12.64**	12.64	8.58
58	kn	Kannada	1 837	6.19	**20.00**	8.10	9.12	**21.40**	7.19
59	tg	Tajik	1 725	11.79	**18.34**	14.69	**16.25**	0.00	0.00
60	jv	Javanese	1 641	**0.00**	0.00	0.00	**0.00**	0.00	0.00

Table 5: Results of the experiment described in Section 4.3. Displayed are results on the languages with the 60 largest lexicons from the Grave et al. (2018) corpus that are not also included in the MUSE corpus.

References

Judit Acs. 2014. Pivot-based multilingual dictionary building using wiktionary. *LREC*.

Wasi Uddin Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2018. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. *arXiv preprint arXiv:1811.00570*.

Maria Antoniak and David Mimno. 2018. Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*, 6:107–119.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017a. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. Unsupervised statistical machine translation. *arXiv preprint arXiv:1809.01272*.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2017b. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*.

Mohamed Ben Aouicha, Mohamed Ali Hadj Taieb, and Hania Ibn Marai. 2018. Wordnet and wiktionary-based approach for word sense disambiguation. In *Transactions on Computational Collective Intelligence XXIX*, pages 123–143. Springer.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Laura Burdick, Jonathan K Kummerfeld, and Rada Mihalcea. 2021. Analyzing the surprising variability in word embedding stability across languages. *EMNLP*.

Shifei Chen and Ali Basirat. 2020. Cross-lingual word embeddings beyond zero-shot machine translation. *Swedish Language Technology Conference (SLTC-2020)*.

Santwana Chimalamarri, Dinkar Sitaram, and Ashritha Jain. 2020. Morphological segmentation to improve crosslingual word embeddings for low resource languages. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(5):1–15.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Mattia Antonino Di Gangi and Marcello Federico. 2017. Monolingual embeddings for low resourced neural machine translation. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 97–104.

Shuoyang Ding and Kevin Duh. 2018. How do source-side monolingual word embeddings impact neural machine translation? *arXiv preprint arXiv:1806.01515*.

Georgiana Dinu and Marco Baroni. 2014. How to make words with vectors: Phrase generation in distributional semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 624–633.

Joel Escudé Font and Marta R Costa-Jussa. 2019. Equalizing gender biases in neural machine translation with word embeddings techniques. *arXiv preprint arXiv:1901.03116*.

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

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Yova Kementchedjhieva, Mareike Hartmann, and Anders Søgaard. 2019. Lost in evaluation: Misleading benchmarks for bilingual dictionary induction. *EMNLP/IJCNLP*.

Yunsu Kim, Jiahui Geng, and Hermann Ney. 2018. Improving unsupervised word-by-word translation with language model and denoising autoencoder. *Empirical Methods in Natural Language Processing (EMNLP)*.

Christo Kirov, John Sylak-Glassman, Roger Que, and David Yarowsky. 2016. Very-large scale parsing and normalization of wiktionary morphological paradigms. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3121–3126.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*, pages 1459–1474.

Megan Leszczynski, Avner May, Jian Zhang, Sen Wu, Christopher Aberger, and Christopher Ré. 2020. Understanding the downstream instability of word embeddings. *Proceedings of Machine Learning and Systems*, 2:262–290.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

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

Grace Muzny and Luke Zettlemoyer. 2013. Automatic idiom identification in wiktionary. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1417–1421.

Vivi Nastase and Carlo Strapparava. 2015. knowitiary: A machine readable incarnation of wiktionary. *Int. J. Comput. Linguistics Appl.*, 6(2):61–82.

Emmanuel Navarro, Franck Sajous, Bruno Gaume, Laurent Prévot, Hsieh ShuKai, Kuo Tzu-Yi, Pierre Magistry, and Huang Chu-Ren. 2009. Wiktionary and nlp: Improving synonymy networks. In *ACL Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, pages 19–27.

Masato Neishi, Jin Sakuma, Satoshi Tohda, Shonosuke Ishiwatari, Naoki Yoshinaga, and Masashi Toyoda. 2017. A bag of useful tricks for practical neural machine translation: Embedding layer initialization and large batch size. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 99–109.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Bénédicte Pierrejean and Ludovic Tanguy. 2018. Towards qualitative word embeddings evaluation: Measuring neighbors variation. In *Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 32–39.

Ye Qi, Devendra Singh Sachan, Matthieu Felix, Sarguna Janani Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? *arXiv preprint arXiv:1804.06323*.

Yuanyuan Qiu, Hongzheng Li, Shen Li, Yingdi Jiang, Renfen Hu, and Lijiao Yang. 2018. Revisiting correlations between intrinsic and extrinsic evaluations of word embeddings. In *Chinese computational linguistics and natural language processing based on naturally annotated big data*, pages 209–221. Springer.

Franck Sajous, Basilio Calderone, and Nabil Hathout. 2020. Englawi: From human-to machine-readable wiktionary. In *12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 3016–3026.

Gilles Sérasset. 2015. Dbnary: Wiktionary as a lemon-based multilingual lexical resource in rdf. *Semantic Web*, 6(4):355–361.

Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Do we really need fully unsupervised cross-lingual embeddings? *EMNLP-IJCNLP*.

Ivan Vulić and Anna-Leena Korhonen. 2016. On the role of seed lexicons in learning bilingual word embeddings. *Proceedings of ACL*.

Ivan Vulić and Marie-Francine Moens. 2015. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 363–372.

Ivan Vulić, Sebastian Ruder, and Anders Søgaard. 2020. Are all good word vector spaces isomorphic? *arXiv preprint arXiv:2004.04070*.

Laura Wendlandt, Jonathan K Kummerfeld, and Rada Mihalcea. 2018. Factors influencing the surprising instability of word embeddings. *NAACL HLT*.

Winston Wu and David Yarowsky. 2020. Wiktionary normalization of translations and morphological information. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4683–4692.

Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. 2019. Generalized data augmentation for low-resource translation. *arXiv preprint arXiv:1906.03785*.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1006–1011.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1970.

Mozhi Zhang, Keyulu Xu, Ken-ichi Kawarabayashi, Stefanie Jegelka, and Jordan Boyd-Graber. 2019. Are girls neko or sh\= ojo? cross-lingual alignment of non-isomorphic embeddings with iterative normalization. *arXiv preprint arXiv:1906.01622*.