Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3036–3045
Marseille, 11–16 May 2020

word2word: A Collection of Bilingual Lexicons for 3,564 Language Pairs

Yo Joong Choe*, **Kyubyong Park***, **Dongwoo Kim***
Kakao Brain
20, Pangyoyeok-ro 241, Bundang-gu, Seongnam-si, Gyeonggi-do, Korea
{yj.choe,kyubyong.park,dongwoo.kim}@kakaobrain.com

Abstract

We present *word2word*, a publicly available dataset and an open-source Python package for cross-lingual word translations extracted from sentence-level parallel corpora. Our dataset provides top-k word translations in 3,564 (directed) language pairs across 62 languages in OpenSubtitles2018 (Lison et al., 2018). To obtain this dataset, we use a count-based bilingual lexicon extraction model based on the observation that not only source and target words but also source words themselves can be highly correlated. We illustrate that the resulting bilingual lexicons have high coverage and attain competitive translation quality for several language pairs. We wrap our dataset and model in an easy-to-use Python library, which supports downloading and retrieving top-k word translations in any of the supported language pairs as well as computing top-k word translations for custom parallel corpora.
Keywords: bilingual lexicon, word translation, Python toolkit

1. Introduction

Bilingual lexicons (Fung, 1998) are valuable resources for cross-lingual tasks, including low-resource machine translation (Ramesh and Sankaranarayanan, 2018; Gū et al., 2019) and cross-lingual word embeddings (Ruder et al., 2017). However, it is often difficult to find a large enough set of bilingual lexicons that is freely and readily available across various language pairs (Levy et al., 2017). For example, standard bilingual dictionaries like Wiktionary[1] often do not explicitly provide word correspondences but refers or redirects to the query word's dictionary form:

- **Query:** *travaillé* (French for 'worked')
  Result: (verb) past principle of *travailler* 'work'
- **Query:** 먹었다 (Korean for 'ate')
  Result: *redirects to* 먹다 *'eat'*

Not only does this make it tedious to find word-level correspondences across many query words, this is particularly problematic when we try to find word correspondences for languages where some dictionary forms are rarely used in ordinary discourse, such as the case of 먹다 in the Korean language.

While the task of bilingual lexicon extraction (BLE) has been popular in both early and recent literature, spanning from count-based approaches (Fung, 1998; Vulić and Moens, 2013; Liu et al., 2013) to using cross-lingual word embeddings (Ruder et al., 2017; Mikolov et al., 2013a; Gouws et al., 2015; Conneau et al., 2017; Levy et al., 2017; Artetxe et al., 2018; Artetxe et al., 2019), few were focused on building high-coverage bilingual lexicons across many language pairs, possibly including non-Indo-European languages. In fact, many of the recent studies and their accompanying packages (Conneau et al., 2017; Artetxe et al., 2018; Glavaš et al., 2019) aim at evaluating cross-lingual word embeddings, so that they involve at most 10-100s of language pairs and 1-5K words for each pair.

# Languages	62
# Language Pairs	3,564
Avg. Lexicon Size	127,023
Avg. # Translations Per Word	8.8

Table 1: Overview of the *word2word* dataset.

Motivated by the lack of publicly available and high-coverage bilingual lexicons across diverse languages, we present *word2word*, a large collection of bilingual lexicons for 3,564 language pairs across 62 languages that is wrapped around an open-source and easy-to-use Python interface. We extract top-k bilingual word correspondences from all parallel corpora provided by OpenSubtitles2018[2] (Lison et al., 2018), using a count-based model that takes into account both monolingual and cross-lingual co-occurrences. The package also provides interface for obtaining bilingual lexicons for custom parallel corpora in any other language pairs and domains not covered by OpenSubtitles2018.

2. The *word2word* Dataset

2.1. Data Statistics

The *word2word* dataset spans across 3,564 directed language pairs between 62 languages in the OpenSubtitles2018 dataset, a collection of translated movie subtitles extracted from OpenSubtitles.org[3]. By design, our methodology covers 100% of words present in the source sentences, making the lexicon size much larger than existing bilingual dictionaries. The lexicon also contains up to top-10 word translations in the target language. We provide an overview of the entire dataset in Table 1.

In Table 2, we provide summary statistics for bilingual lexicons between English and some of the major languages (both European and non-European). For each pair, the lexicon size ranges from 76.2K (English-Russian) to

*Equal contribution.
[1] https://en.wiktionary.org
[2] http://opus.nlpl.eu/OpenSubtitles-v2018.php
[3] http://www.opensubtitles.org/

Language Pair	Lexicon Size	# Unique Translations	Avg. # Translations Per Word	# Sentences Used
Arabic-English	335.5K	86.0K	9.7	29.8M
English-Arabic	97.6K	191.6K	9.5	29.8M
S.Chinese-English	214.0K	87.0K	9.5	11.2M
English-S.Chinese	101.6K	139.1K	9.4	11.2M
T.Chinese-English	201.7K	72.5K	9.5	4.8M
English-T.Chinese	85.8K	119.7K	9.2	4.8M
French-English	92.1K	59.1K	9.8	41.8M
English-French	72.1K	71.4K	9.7	41.8M
Italian-English	111.5K	63.9K	9.7	35.2M
German-English	127.0K	64.8K	9.7	22.5M
English-German	73.6K	95.9K	9.6	22.5M
English-Italian	75.4K	83.9K	9.6	35.2M
Japanese-English	83.3K	75.2K	9.2	2.1M
English-Japanese	102.1K	63.8K	9.3	2.1M
Korean-English	87.2K	75.8K	9.3	1.4M
English-Korean	105.5K	69.8K	9.1	1.4M
Russian-English	213.4K	68.7K	9.7	25.9M
English-Russian	76.2K	155.8K	9.5	25.9M
Spanish-English	107.1K	60.8K	9.8	61.4M
English-Spanish	73.9K	82.5K	9.7	61.4M
Thai-English	155.6K	84.2K	9.4	3.3M
English-Thai	109.2K	99.2K	9.2	3.3M
Vietnamese-English	96.6K	76.6K	9.0	3.5M
English-Vietnamese	96.4K	75.3K	9.3	3.5M

Table 2: Summary statistics for the *word2word* dataset between selected languages and English. Lexicon size refers to the number of unique words in source language for which translations exist. S.Chinese and T.Chinese refer to simplified and traditional Chinese, respectively.

335.5K (Arabic-English), demonstrating the broad coverage of words in the dataset. For each of these words, the dataset includes an average of 9 or more highest-scored translations according to our extraction approach described in Section 3.1. Lexicon size for all language pairs can be found in Appendix B.

2.2. Examples

In Table 3, we present samples of top-5 word translations in the English↔French and English↔Korean bilingual lexicons. For each language pair, we randomly sample five words from the top-10,000 frequent words in the source lexicon and provide their top-5 word translations. This is to show translations for words that are relatively more likely used than others in typical discourse.

3. Methodology

3.1. Bilingual Lexicon Extraction

Bilingual lexicon extraction (BLE) is a classical natural language task where the goal is to find word-level correspondences from a (parallel) corpus. There are many different approaches to BLE, such as word alignment methods (Brown et al., 1993; Vogel et al., 1996; Koehn et al., 2007) and cross-lingual word representations (Ruder et al., 2017; Mikolov et al., 2013a; Liu et al., 2013; Gouws et al., 2015; Conneau et al., 2017).

Among them, we focus on simple approaches that can work well with various sizes of parallel corpora that are present in OpenSubtitles2018, which ranges from 129 sentence pairs in Armenian-Indonesian to 61M sentence pairs in English-Spanish. In particular, we avoid methods that require high-resource parallel corpora (e.g., neural machine translation) or external corpora (e.g., unsupervised or semi-supervised cross-lingual word embeddings). Also, since bilingual word-to-word mappings are hardly one-to-one (Fung, 1998; Somers, 2001; Levy et al., 2017), we consider methods that yield relevance scores between every source-target word pair, such that we can extract not just one but the top-k correspondences. For these reasons, we consider approaches based on (monolingual and cross-lingual) co-occurrence counts: co-occurrences, pointwise mutual information (PMI), and co-occurrences with controlled predictive effects (CPE).

3.1.1. Co-occurrences

The simplest baseline for our goal is to compute the co-occurrences between each source word x and target word y. For each source word x, we can score any target word y based on the conditional probability $p(y|x) \propto p(x, y)$:

$$p(y|x) = \frac{p(x,y)}{p(x)} \approx \frac{\#(x,y)}{\#(x)} \propto \#(x,y) \qquad (1)$$

where $\#(\cdot)$ denotes the number of (co-)occurrence counts of the word or word pair across the parallel corpus. The top-k translations of source word x can be computed as the top-k target words with respect to their co-occurrence counts with x.

Word	Top-5 Translations				
English	French				
exceptional	exceptionnel	exceptionnelle	exceptionnels	exceptionnelles	exception
whether	plaise	décider	importe	question	savoir
committee	comité	éthique	accueil	commission	central
clown	clown	clowns	bouffon	guignol	cirque
spread	dispersez-vous	propagation	répandre	propager	répandu
French	English				
hobbs	hobbs	abigail	garret	jacob	garrett
mêlé	mixed	involved	middle	part	murder
établir	establish	establishing	set	able	connection
taule	slammer	joint	locked	jail	prison
chaussettes	socks	sock	stockings	pairs	underwear
English	Korean				
slaughtered	학살	도륙	도살	당했	살육
shadow	그림자	그늘	알맞	어둠	존재
Charles	찰스	제프리	Charles	조프리	램퍼트
concerns	걱정	우려	염려	관한	판단력
reverse	역	뒤집	후진	거꾸로	되돌리
Korean	English				
아유	arm	Thank	thrilled	killing	NamWon
상어	shark	Shark	sharks	Tank	Tiger
쥐	rat	rats	mouse	mice	squeeze
기꺼이	willing	happy	pleasure	gladly	willingly
어떤	Some	kind	which	any	anything

Table 3: Randomly sampled words and their top-5 translations in the English↔French and English↔Korean *word2word* bilingual lexicons. Top-5 translations are listed in the descending order of scores.

3.1.2. Pointwise Mutual Information

Another simple baseline is pointwise mutual information (PMI), which further accounts for the monolingual frequency of a candidate target word y:

$$\begin{aligned} \mathsf{PMI}(x,y) &= \log \frac{p(x,y)}{p(x)p(y)} \\ &\approx \log \frac{\#(x,y)}{\#(x)\#(y)} \propto \log \#(x,y) - \log \#(y) \end{aligned} \tag{2}$$

Compared to the co-occurrence model in (1), PMI can help prevent stop words from obtaining high scores.

The use of PMI has been connected to the skip-gram with negative sampling (SGNS) (Levy and Goldberg, 2014) model of *word2vec* (Mikolov et al., 2013b). PMI can also be interpreted as a conditional version of TF-IDF (Fung, 1998).

3.1.3. Controlled Predictive Effects

While conditional probability and PMI are proportional to cross-lingual co-occurrence counts, they can fail to distinguish exactly which source word in the sentence is the most predictive of the corresponding target word in the translated sentence. For example, given a English-French pair (*the apple juice*, *la jus de pomme*), these baseline methods cannot isolate the effect of *apple*, as opposed to *the* and *juice*, on *pomme*.

To deal with this issue, we add a correction term that averages the probability of seeing y given a confounder x' in the source language, i.e. $p(y|x')$. This probability is then weighted by the probability of actually seeing that confounder, i.e. $p(x'|x)$. This correction can be explained intuitively by the dashed arrows in the schematic graphical model in Figure 1– it reflects the conditional independence

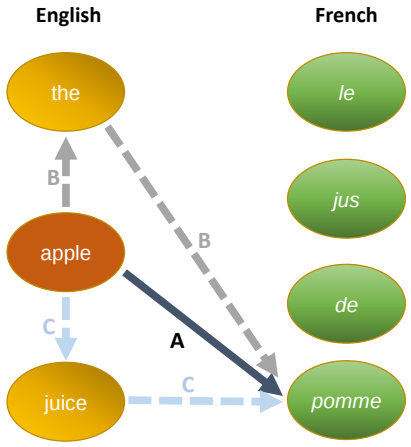

Figure 1: A schematic graphical model of English and French words. Co-occurrence and PMI models focus on the relationship from *apple* to *pomme* (**A**). CPE further controls for the confounding effect of other collocates like *the* (B) and *juice* (C).

Metric (%)	Method	en-es	es-en	en-fr	fr-en	en-de	de-en	en-ru	ru-en	en-zh	zh-en	en-it	it-en
# Sentence Pairs		61.4M		41.8M		22.5M		25.9M		4.8M		35.2M	
P@1	Co-occurrence	22.3	25.5	18.7	21.9	10.5	23.5	3.3	11.4	5.4	3.8	24.9	24.1
	PMI	72.7	72.3	73.9	72.1	62.1	71.9	32.8	55.0	24.8	33.1	68.1	69.5
	MUSE	81.7	**83.3**	82.3	**82.4**	74.0	72.4	51.7	63.7	42.7	37.5	66.2	58.7
	CPE	**82.4**	79.5	**83.6**	80.7	**82.4**	**81.1**	**66.7**	**68.9**	**56.0**	**58.7**	**80.9**	**82.1**
P@5	Co-occurrence	67.8	71.4	63.1	66.3	63.7	65.5	52.3	51.8	46.0	36.3	61.9	68.5
	PMI	**92.3**	**90.4**	**92.5**	**90.1**	90.5	**88.1**	74.1	79.5	58.7	66.1	**90.3**	**91.1**
	MUSE	-	-	-	-	-	-	-	-	-	-	80.4	76.5
	CPE	90.1	88.4	91.7	89.3	**90.7**	87.7	**79.5**	**80.0**	**73.5**	**72.8**	89.8	89.9

Table 4: Precision (%) on 1,500 word translations (test split from MUSE) for language pairs evaluated in the MUSE paper. P@1 and P@5 denote the precision of top-1 and top-5 predictions, respectively. The ISO 639-1 language codes are used (en: English, es: Spanish, fr: French, de: German, ru: Russian, zh: traditional Chinese, it: Italian).

relationships between words that the baseline models do not. We call the resulting approach as the method of *controlled predictive effects (CPE)*.

Formally, we define the corrected CPE score as follows:

$$\begin{aligned}\mathsf{CPE}(y \mid x) &= p(y \mid x) - \sum_{x' \in \mathcal{X}} p(y \mid x')p(x' \mid x) \\ &= \sum_{x' \in \mathcal{X}} \mathsf{CPE}_{y|x}(x')p(x' \mid x) \qquad (3)\end{aligned}$$

where $\mathcal{X}$ is the source vocabulary and $\mathsf{CPE}_{y|x}(x')$ denotes the CPE term of any other source word x' when predicting y from x. Formally, this term is defined as

$$\mathsf{CPE}_{y|x}(x') = p(y \mid x, x') - p(y \mid x') \qquad (4)$$

This CPE term measures the effect of *additionally* seeing x (*apple*) when predicting y (*pomme*), after controlling for the effect of any other x' (*the*), which the model views as a confounder. If $\mathsf{CPE}_{y|x}(x') = 0$, then $x \perp\!\!\!\perp y \mid x'$, meaning that after observing a confounder x', x is no longer related to y. The CPE term for each confounder x' is then marginalized over all possible confounders to give a final score, weighted by the probability of seeing that confounder in a sentence with x. Note that $\mathsf{CPE}_{y|x}(x) = 0$, meaning that, after seeing x when predicting y, there is no additional effect by seeing x (again).

In practice, summing the CPE scores over all words in the source vocabulary can be inefficient. Because many of the (unrelated) words in the vocabulary do not play a role in the confounding, we select the top-m source words with the highest co-occurrence counts and correct for their effects only. We used $m = 5,000$ in our experiments and found that using a larger m did not make a meaningful difference on the quality of top-1 and top-5 correspondences.

3.1.4. Evaluation on MUSE Bilingual Dictionaries

We first evaluate the methods on the same ground-truth bilingual dictionaries as MUSE[4], a cross-lingual neural embedding model. Each dictionary contains 1,500 words and their translations obtained using an internal translation tool from the authors. Although we consider MUSE's performance as a reference, we do note that it is difficult to make a fair comparison against MUSE: the count-based methods use parallel corpora from OpenSubtitles2018, while MUSE embeddings are instead learned from monolingual Wikipedia data (for its unsupervised version) and an additional 5,000-word bilingual lexicon (for its supervised version).

In Table 4, we report the top-1 and top-5 precision scores (P@1 and P@5, respectively) of the count-based methods and MUSE embeddings across twelve[5] directed language pairs that were used to evaluate MUSE in its paper (Conneau et al., 2017): English-Spanish, English-French, German-English, English-Russian, English-Chinese (traditional), and English-Italian, all in both directions. For MUSE, we report its best reported performance (only top-1 precision is reported, except for en-it and it-en) among its supervised and unsupervised variants.

Our main finding is that the CPE method consistently and significantly outperforms the co-occurrence and PMI baselines at top-1 precision score. We also find that CPE outperforms MUSE on most of the reported language pairs, especially when the number of sentence pairs is comparatively small (e.g., 13-21% improvement between English and Chinese, for which there are about 6% as many sentence pairs as those between English and Spanish). In terms of the top-5 precision score, the CPE method performs comparatively well with the PMI method, which performs better on some of the selected language pairs. Compared to the CPE method, we suspect that the PMI method overly favors rare words because it directly penalizes word counts, so that the most likely correspondence (which isn't necessarily the least common) is pushed back to later ranks. More examples can be found in Appendix A

3.1.5. Evaluation on Non-European Languages

Next, we compare the performance of co-occurrence, PMI, and CPE methods on language pairs between English and some of the major non-European languages: Arabic, simplified Chinese, Japanese, Korean, Thai, and Vietnamese. As we detail in Section 3.2., these languages commonly require special word segmentation techniques. Also, they

[4] https://github.com/facebookresearch/MUSE

[5] The MUSE paper also presents the results on English-Esperanto and Esperanto-English, but the ground-truth dictionary is no longer available online. See https://github.com/facebookresearch/MUSE/issues/34.

Metric (%)	Method	en-ar	ar-en	en-zh	zh-en	en-ja	ja-en	en-ko	ko-en	en-th	th-en	en-vi	vi-en
# Sentence Pairs		29.8M		11.2M		2.1M		1.4M		3.3M		3.5M	
P@1	Co-occurrence	23.3	1.1	2.1	0.4	5.0	0.3	22.9	0.4	0.6	0.5	4.0	2.1
	PMI	13.3	20.7	8.5	20.6	33.5	16.7	14.0	14.9	18.3	13.4	20.5	16.5
	CPE	**30.3**	**27.9**	**48.3**	**34.3**	**49.3**	**40.4**	**39.1**	**38.1**	**48.1**	**31.0**	**30.0**	**37.7**
P@5	Co-occurrence	46.9	35.2	50.5	27.1	30.7	29.1	36.6	26.9	55.6	24.4	39.3	28.3
	PMI	57.0	**61.6**	78.7	**65.3**	64.0	60.5	48.8	57.7	64.5	52.7	**50.1**	60.4
	CPE	**58.1**	50.5	**80.9**	60.1	**66.8**	**66.4**	**54.9**	**60.0**	**69.3**	**53.1**	48.9	**62.2**

Table 5: Precision (%) on 2,000 word translations between six *non-European* languages and English (source words randomly sampled from OpenSubtitles2018; gold labels taken from Google Translate). P@1 and P@5 denote the precision of top-1 and top-5 predictions, respectively. The ISO 639-1 language codes are used (ar: Arabic, zh: simplified Chinese, ja: Japanese, ko: Korean, th: Thai, vi: Vietnamese).

Language	Python Tokenizer Module	Reference
Arabic	`pyarabic.araby`	(Zerrouki, 2010)
Chinese (Simplified)	`Mykytea`	(Neubig et al., 2011)
Chinese (Traditional)	`jieba`	n/a
Japanese	`Mykytea`	(Neubig et al., 2011)
Korean	`konlpy.tag.Mecab`	(Park and Cho, 2014)
Thai	`pythainlp`	n/a
Vietnamese	`pyvi`	n/a
Others	`nltk.tokenize.TokTokTokenizer`	(Bird et al., 2009; Dehdari, 2014)

Table 6: List of Python tokenizer modules used for each language.

typically have relatively smaller amounts of sentences paired with English, making it more challenging for the models to achieve high precision.

Unfortunately, we learned in our early experiments that the MUSE test set translations are far from being perfect for these non-European languages. For example, in English-Vietnamese, we found that 80% of the 1,500 word pairs in the test set had the same word twice as a pair (e.g. crimson-crimson, Suzuki-Suzuki, Randall-Randall). Thus, for the non-European languages, we instead evaluate on translations using Google Translate[6], a proprietary[7] web software for machine translation. To construct this test set, we first sample 2,000 words from the monolingual word distribution of that language pair's OpenSubtitles2018 parallel corpus. We use temperature-based smoothing ($T = 1.25$) for the distribution to include more low-frequency words in the test set and also filter out words that include characters not from its alphabet (e.g., *Charles* in Korean). Then, for each of the 2,000 sampled words, we retrieve "common" and "uncommon" translations[8] from Google Translate and treat them as ground truth labels.

The results are summarized in Table 5. Here, we see more evidence that the CPE method performs significantly better than both the co-occurrence and the PMI methods in top-1 precision as well as top-5 precision. The performance gap tends to be larger both when the language's words are not whitespace-separated (e.g., Chinese and Japanese) and when there are a relatively small number of paired sentences (e.g., Korean and Thai). Based on the results from Tables 4 and 5, we employ the CPE method to produce the *word2word* dataset.

3.2. Word Segmentation

Since many of the 62 languages we consider are sensitive to word segmentation, we use language-specific tokenization tools when necessary. Specifically, we use publicly available tokenization packages for morphologically complex languages, i.e., Arabic (Attia, 2007) and Korean, and languages in which words are not separated by spaces, i.e., Chinese, Japanese, Thai, and Vietnamese[9]. For all other languages, we use the tok-tok tokenizer (Dehdari, 2014) implemented in NLTK (Bird et al., 2009). Table 6 summarizes the tokenization packages we used in the *word2word* dataset and their references.

4. The *word2word* Python Interface

As part of releasing the dataset and making it easily accessible and reproducible, we also introduce the *word2word* Python package. The open-source package provides an easy-to-use interface for both downloading and accessing bilingual lexicons for any of the 3,564 language pairs and building a custom bilingual lexicon on other language pairs for which there is a parallel corpus. Our source code is available on PyPi as `https://pypi.org/project/word2word/`.

[6] `https://translate.google.com/`

[7] We note that, because Google Translate is proprietary and not open-source, its results may change depending on the time of access. Our evaluations use Google Translate results accessed on July 19, 2019.

[8] For word translations, Google Translate categorizes its translations to three categories: common, uncommon, and rare translations.

[9] Spaces in Vietnamese delimit syllables.

4.1. Implementation

The *word2word* package is built entirely using Python 3. The package includes scripts for downloading and preprocessing parallel corpora from OpenSubtitles2018, including word segmentation, and for computing the CPE scores for all available word tokens within each parallel corpus. After processing, the package stores the bilingual lexicon as a Python `pickle` file, typically sized a few megabytes per language pair. The `pickle` file contains a Python dictionary that maps each source word to a list of top-10 word correspondences in $O(1)$ time. This allows bilingual lexicons to be portable and accessible.

4.2. Usage

The Python interface provides a simple API to download and access the *word2word* dataset. As demonstrated in Figure 2, word translations for any query word can be retrieved as a list with a few lines of Python code.

```
from word2word import Word2word

en2fr = Word2word('en', 'fr')
print(en2fr('apple'))
# ['pomme', 'pommes', 'pommier',
#  'tartes', 'fleurs']
```

Figure 2: The *word2word* Python interface for retrieving word translations.

4.3. Building a Custom Bilingual Lexicon

The *word2word* package also allows training a custom bilingual lexicon using a different parallel corpus. This can be useful in cases where there are larger and/or higher-quality parallel corpora available for the language pair of interest or when utilizing word translations for a particular domain (e.g., government, law, and medical). This process can also be done using a few lines of Python code, as demonstrated in Figure 3. For an OpenSubtitles2018 corpus of a million parallel sentences, building a bilingual lexicon takes approximately 10 minutes using 8 CPUs.

```
from word2word import Word2word

my_en2fr = Word2word.make(
    'en', 'fr', 'data/pubmed.en-fr'
)
# ...building...done!
print(my_en2fr('mitochondrial'))
# ['mitochondriale', 'mitochondriales',
#  'mitochondrial', 'cytopathies',
#  'mitochondriaux']
```

Figure 3: The *word2word* Python interface for building a custom bilingual lexicon. Once built, the lexicon can be accessed in the same way as done in Figure 2.

5. Conclusion

In this paper, we present the *word2word* dataset, a publicly available collection of bilingual lexicons for 3,564 language pairs that are extracted from OpenSubtitles2018. The bilingual lexicons have high coverage (up to hundreds of thousands words) for many language pairs and provide word translations of similar or better quality compared to those from a state-of-the-art embedding model. We also release the *word2word* Python package, with which the user can easily access the dataset or build a custom lexicon for different parallel corpora. We hope that the dataset and its Python interface can facilitate research on improving cross-lingual models, including machine translation models (Ramesh and Sankaranarayanan, 2018; Gū et al., 2019) and cross-lingual word embeddings (Conneau et al., 2017; Ruder et al., 2017).

6. Bibliographical References

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

Artetxe, M., Labaka, G., and Agirre, E. (2019). Bilingual lexicon induction through unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5002–5007, Florence, Italy, July. Association for Computational Linguistics.

Attia, M. A. (2007). Arabic tokenization system. In *Proceedings of the 2007 workshop on computational approaches to semitic languages: Common issues and resources*, pages 65–72. Association for Computational Linguistics.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2017). Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Dehdari, J. (2014). *A Neurophysiologically-Inspired Statistical Language Model*. Ph.D. thesis, The Ohio State University.

Fung, P. (1998). A statistical view on bilingual lexicon extraction: from parallel corpora to non-parallel corpora. *Machine Translation and the Information Soup*, pages 1–17.

Glavaš, G., Litschko, R., Ruder, S., and Vulić, I. (2019). How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 710–721, Florence, Italy, July. Association for Computational Linguistics.

Gouws, S., Bengio, Y., and Corrado, G. (2015). Bilbowa: Fast bilingual distributed representations without word alignments. In Francis Bach et al., editors, *Proceedings*

of the 32nd International Conference on Machine Learning, volume 37 of *Proceedings of Machine Learning Research*, pages 748–756, Lille, France, 07–09 Jul. PMLR.

Gū, J., Shavarani, H. S., and Sarkar, A. (2019). Pointer-based fusion of bilingual lexicons into neural machine translation. *arXiv preprint arXiv:1909.07907*.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

Levy, O. and Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, pages 2177–2185.

Levy, O., Søgaard, A., and Goldberg, Y. (2017). A strong baseline for learning cross-lingual word embeddings from sentence alignments. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 765–774.

Lison, P., Tiedemann, J., and Kouylekov, M. (2018). Opensubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Liu, X., Duh, K., and Matsumoto, Y. (2013). Topic models + word alignment = a flexible framework for extracting bilingual dictionary from comparable corpus. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 212–221, Sofia, Bulgaria, August. Association for Computational Linguistics.

Mikolov, T., Le, Q. V., and Sutskever, I. (2013a). Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Neubig, G., Nakata, Y., and Mori, S. (2011). Pointwise prediction for robust, adaptable japanese morphological analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 529–533. Association for Computational Linguistics.

Park, E. L. and Cho, S. (2014). Konlpy: Korean natural language processing in python. In *Proceedings of the 26th Annual Conference on Human & Cognitive Language Technology*, Chuncheon, Korea, October.

Ramesh, S. H. and Sankaranarayanan, K. P. (2018). Neural machine translation for low resource languages using bilingual lexicon induced from comparable corpora. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 112–119, New Orleans, Louisiana, USA, June. Association for Computational Linguistics.

Ruder, S., Vulić, I., and Søgaard, A. (2017). A survey of cross-lingual embedding models. *arXiv preprint arXiv:1706.04902*.

Somers, H. (2001). Bilingual parallel corpora and language engineering. In *Proc. Anglo-Indian Workshop" Language Engineering for South-Asian languages*.

Vogel, S., Ney, H., and Tillmann, C. (1996). Hmm-based word alignment in statistical translation. In *Proceedings of the 16th conference on Computational linguistics-Volume 2*, pages 836–841. Association for Computational Linguistics.

Vulić, I. and Moens, M.-F. (2013). Cross-lingual semantic similarity of words as the similarity of their semantic word responses. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 106–116, Atlanta, Georgia, June. Association for Computational Linguistics.

Zerrouki, T. (2010). pyarabic, an arabic language library for python.

A Sample Translations from Different Extraction Methods

In Table 7, we compare the BLE methods described in Section 3.1. from illustrative examples of their extracted bilingual lexicons for English to Spanish and English to Simplified Chinese. These examples show that the CPE approach provides the correct correspondence as its top-1 translation in both languages, while the PMI approach seems to excessively favor rarer words among the co-occurrences. As illustrated in the English-Chinese example, this can be particularly problematic with languages such as Chinese, where word segmentation is highly nontrivial. The co-occurrence method prefers stop words that are frequent over the entire document, rather than the corresponding words.

A1. Co-occurrences

The baseline co-occurrence model performs poorly in both experiments (Tables 4 and 5). As exemplified in Table 7, we find that the top-5 predictions in many cases are primarily stop words, such as *la* (the), *de* (of), and *que* (that) in Spanish and 的 (of), 你 (you), and 我 (I, me) in Chinese, because they frequently occur in any sentence, regardless of context.

A2. Comparing PMI and CPE

Comparing translations using PMI and CPE, we find in Table 7 that PMI favors less frequent words excessively. This results in two kinds of error cases: (a) when PMI overemphasizes rare words in the target vocabulary, e.g. *solarización* for *library* in en-es, and (b) when PMI misses correct words in the target language that are relatively frequently used, e.g. *bien* for *good* in en-es. Another consequence is that PMI prefers less common variants of the same word, in particular conjugations and past/future tenses as well as typos, when two forms of the same word have comparable counts (e.g. *obligados* preferred over *obligado* in Spanish for the English *obliged*).

Because of the second reason, we also find that *word2word* tends to be more robust to tokenization issues, which are common in non-whitespace-separated languages like Chinese. For example, since the tokenizer failed to separate 张开嘴 (open mouth), which in general occurs far less frequently than 嘴 (mouth), PMI favors 张开嘴 over the more frequent 嘴 as its first choice.

B Full Dataset Statistics

In Table 8, we list the sizes of all 3,564 bilingual lexicons in the *word2word* dataset. By size, we refer to the number of source words for which translations exist. For each source word, we extract up to 10 (9+ on average) most likely translations according to the CPE method described in 3.1.3.

English	Methods	Top-5 Translations in Spanish					Top-5 Translations in Simplified Chinese				
its	Co-occurrence	de	la	que	el	y	的	**它**	了	是	我
	PMI	propia	**sus**	**su**	tierra	poder	**它**	政府	国家	失去	由
	CPE	**su**	**sus**	propia	tierra	cada	**它**	将	自己	国家	中
good	Co-occurrence	que	de	no	**bien**	es	**好**	的	你	我	很
	PMI	**buenas**	noches	**buenos**	**buena**	**buen**	祝你好运	晚安	好消息	早上好	早安
	CPE	**bien**	**buena**	**buenas**	**buen**	**bueno**	**好**	很	**不错**	晚安	早上好
mouth	Co-occurrence	la	**boca**	de	que	no	的	你	我	**嘴**	了
	PMI	**boca**	cerrada	pico	mantén	abre	张开嘴	嘴里	大嘴巴	张嘴	**嘴巴**
	CPE	**boca**	cerrada	abre	palabras	labios	**嘴**	嘴里	**嘴巴**	闭上	闭嘴
library	Co-occurrence	la	**biblioteca**	de	en	que	**图书馆**	的	我	在	你
	PMI	solarización	**biblioteca**	soltándola	library	librería	英	**图书馆**	**圖書館**	藏书室	书房
	CPE	**biblioteca**	la	librería	pública	tarjetas	**图书馆**	书房	里	**圖書館**	去

Table 7: Selected *word2word* translations of English words into Spanish and simplified Chinese. Top-5 predictions are listed in the decreasing order of the model's scores. Boldfaced target words indicate correct translations.

	af	ar	bg	bn	br	bs	ca	cs	da	de	el	en	eo	es	et	eu	fa	fi	fr	gl	he	hi	hr	hu	hy	id	is	it	ja	ka	kk	ko	lt	lv	mk	ml	ms	nl	no	pl	pt	pt_br	ro	ru	si	sk	sl	sq	sr	sv	ta	te	th	tl	tr	uk	ur	vi	ze_en	ze_zh	zh_cn	zh_tw
af	0	8K	11K	2K	0	4K	0	11K	9K	11K	15K	19K	1K	16K	5K	0	4K	8K	11K	0	9K	1K	9K	11K	0	4K	0	10K	2K	0	0	0	972	1K	4K	3K	1K	12K	5K	10K	12K	16K	16K	10K	1K	5K	8K	1K	10K	6K	1K	0	2K	0	13K	3K	0	3K	1K	0	6K	3K
ar	18K	0	347K	158K	17K	337K	182K	342K	341K	342K	347K	335K	27K	343K	330K	236K	307K	353K	347K	70K	347K	58K	343K	350K	6K	319K	331K	346K	336K	102K	2K	336K	328K	197K	316K	159K	321K	364K	355K	353K	346K	347K	343K	332K	196K	348K	340K	330K	344K	351K	26K	21K	327K	12K	335K	241K	26K	309K	231K	302K	334K	326K
bg	20K	211K	0	98K	15K	206K	117K	204K	209K	201K	196K	188K	24K	196K	208K	162K	220K	211K	205K	45K	213K	36K	199K	202K	2K	211K	200K	205K	205K	73K	4K	211K	206K	133K	213K	98K	194K	210K	216K	209K	199K	196K	194K	207K	125K	209K	207K	196K	195K	215K	18K	16K	201K	10K	193K	161K	20K	206K	134K	188K	220K	223K
bn	2K	89K	86K	0	0	57K	12K	85K	79K	79K	88K	97K	0	89K	75K	24K	72K	82K	80K	8K	85K	10K	85K	86K	0	85K	23K	77K	52K	8K	0	36K	41K	23K	57K	30K	62K	82K	72K	87K	84K	90K	88K	77K	25K	48K	81K	44K	83K	82K	3K	0	48K	3K	87K	21K	6K	69K	11K	10K	78K	63K
br	0	9K	9K	0	0	4K	5K	10K	8K	9K	10K	10K	2K	10K	9K	1K	5K	9K	10K	1K	9K	0	9K	10K	0	7K	5K	8K	0	0	0	0	0	0	1K	2K	0	9K	8K	10K	10K	10K	10K	8K	0	5K	9K	5K	9K	7K	0	0	0	0	10K	2K	0	0	0	0	6K	0
bs	8K	276K	264K	84K	5K	0	93K	260K	279K	269K	259K	253K	13K	262K	266K	123K	275K	276K	267K	38K	272K	26K	261K	271K	200	265K	216K	276K	241K	65K	2K	155K	215K	122K	272K	85K	215K	279K	278K	277K	272K	262K	260K	274K	108K	273K	268K	261K	254K	280K	13K	13K	280K	5K	259K	122K	13K	274K	128K	131K	278K	283K
ca	0	82K	84K	11K	5K	52K	0	89K	68K	73K	87K	95K	0	96K	66K	19K	49K	77K	86K	12K	83K	5K	82K	88K	0	60K	18K	86K	40K	6K	0	21K	23K	10K	42K	7K	22K	86K	57K	89K	85K	95K	88K	74K	8K	41K	74K	24K	83K	69K	0	0	40K	0	90K	23K	0	46K	22K	21K	51K	33K
cs	21K	264K	262K	123K	18K	259K	149K	0	266K	258K	252K	245K	33K	250K	255K	208K	267K	268K	262K	65K	266K	44K	260K	258K	7K	256K	263K	270K	251K	86K	2K	256K	249K	167K	262K	124K	236K	266K	268K	260K	261K	256K	252K	262K	151K	255K	265K	247K	252K	265K	19K	18K	254K	11K	251K	215K	23K	252K	211K	253K	265K	272K
da	10K	164K	159K	68K	7K	167K	65K	156K	0	160K	157K	151K	9K	159K	161K	92K	155K	160K	162K	23K	166K	24K	159K	159K	0	157K	158K	163K	149K	44K	3K	126K	153K	90K	157K	59K	146K	162K	160K	161K	159K	161K	158K	166K	77K	162K	161K	150K	157K	160K	12K	12K	151K	3K	155K	86K	12K	151K	115K	118K	171K	166K
de	15K	155K	157K	72K	10K	163K	78K	153K	160K	0	143K	142K	18K	147K	162K	112K	172K	155K	152K	36K	162K	28K	151K	156K	4K	165K	163K	157K	166K	45K	4K	145K	161K	94K	168K	69K	157K	156K	161K	155K	157K	148K	143K	164K	80K	171K	158K	166K	148K	159K	12K	10K	171K	11K	145K	103K	14K	163K	105K	92K	167K	174K
el	25K	241K	236K	108K	14K	241K	119K	222K	248K	229K	0	197K	30K	211K	231K	178K	246K	255K	218K	66K	251K	38K	214K	222K	6K	237K	220K	228K	221K	71K	3K	220K	223K	143K	232K	113K	213K	237K	241K	226K	226K	216K	214K	246K	131K	234K	245K	215K	213K	261K	17K	15K	218K	12K	211K	171K	19K	225K	136K	184K	252K	251K
en	18K	98K	96K	60K	9K	101K	67K	95K	98K	94K	87K	0	22K	97K	103K	89K	107K	93K	91K	51K	104K	26K	89K	94K	4K	101K	98K	95K	102K	41K	3K	106K	98K	67K	104K	59K	93K	91K	95K	93K	89K	99K	89K	96K	73K	106K	102K	98K	86K	98K	14K	12K	109K	10K	86K	94K	14K	96K	112K	121K	102K	106K
eo	2K	20K	21K	0	2K	10K	0	27K	12K	22K	26K	38K	0	34K	18K	4K	8K	18K	23K	544	19K	2K	20K	24K	2K	13K	2K	23K	7K	0	1K	2K	3K	2K	2K	2K	4K	19K	12K	26K	22K	31K	27K	24K	3K	7K	20K	4K	26K	17K	0	0	8K	3K	28K	5K	0	4K	1K	4K	15K	9K
es	22K	144K	142K	77K	13K	143K	97K	137K	145K	138K	132K	131K	30K	0	149K	130K	151K	140K	134K	74K	150K	33K	133K	136K	6K	147K	142K	138K	148K	57K	3K	145K	141K	98K	145K	79K	140K	142K	144K	137K	136K	136K	132K	141K	96K	151K	147K	140K	130K	144K	18K	16K	151K	12K	134K	125K	17K	141K	141K	153K	149K	153K
et	8K	251K	257K	102K	15K	254K	102K	245K	260K	256K	259K	256K	21K	264K	0	143K	255K	258K	262K	38K	257K	32K	256K	257K	488	251K	249K	256K	250K	61K	2K	164K	251K	143K	247K	88K	238K	263K	265K	260K	258K	259K	257K	262K	122K	251K	253K	243K	254K	261K	13K	14K	243K	4K	256K	137K	19K	244K	142K	154K	266K	261K
eu	0	167K	175K	29K	2K	103K	24K	185K	137K	157K	185K	200K	5K	202K	129K	0	95K	172K	181K	14K	167K	12K	167K	182K	0	107K	35K	167K	66K	11K	0	43K	55K	29K	62K	29K	59K	176K	119K	183K	178K	198K	189K	141K	25K	90K	154K	37K	174K	160K	5K	1K	47K	3K	191K	31K	3K	81K	19K	19K	127K	110K
fa	5K	215K	222K	86K	7K	211K	62K	220K	216K	217K	226K	213K	9K	224K	210K	95K	0	221K	223K	25K	218K	27K	216K	217K	0	203K	99K	212K	213K	45K	2K	152K	156K	86K	207K	67K	207K	224K	214K	231K	216K	222K	220K	210K	93K	205K	220K	187K	220K	223K	12K	9K	194K	9K	219K	85K	17K	210K	52K	76K	208K	216K
fi	14K	338K	332K	136K	16K	339K	158K	329K	333K	331K	331K	319K	25K	328K	330K	236K	340K	0	339K	51K	340K	45K	331K	332K	3K	326K	329K	328K	334K	94K	5K	307K	323K	197K	327K	126K	313K	339K	331K	332K	334K	331K	328K	344K	172K	336K	335K	323K	329K	338K	21K	18K	334K	10K	327K	225K	22K	327K	256K	273K	342K	343K
fr	13K	120K	120K	62K	11K	121K	75K	114K	122K	115K	109K	108K	18K	112K	124K	99K	127K	117K	0	44K	127K	26K	110K	116K	2K	125K	120K	116K	124K	43K	2K	122K	123K	78K	122K	62K	120K	117K	121K	116K	115K	112K	109K	124K	72K	125K	124K	121K	108K	120K	12K	9K	122K	11K	110K	94K	14K	120K	101K	96K	126K	131K
gl	0	45K	41K	8K	2K	28K	14K	50K	30K	43K	54K	83K	648	87K	32K	13K	27K	37K	63K	0	34K	1K	43K	45K	0	23K	6K	58K	18K	3K	0	11K	14K	7K	21K	6K	12K	41K	25K	50K	47K	75K	50K	41K	6K	19K	37K	12K	52K	32K	0	0	19K	0	62K	8K	2K	21K	5K	4K	34K	22K
he	16K	250K	246K	109K	15K	236K	123K	240K	250K	248K	249K	239K	23K	251K	230K	166K	235K	253K	258K	40K	0	40K	248K	248K	3K	227K	230K	252K	223K	73K	4K	215K	219K	142K	242K	105K	215K	256K	258K	247K	245K	246K	244K	240K	127K	228K	245K	216K	247K	263K	22K	16K	222K	11K	241K	163K	20K	218K	135K	179K	254K	245K
hi	1K	24K	22K	8K	0	14K	4K	21K	19K	21K	21K	27K	1K	24K	18K	7K	18K	21K	21K	987	21K	0	21K	21K	0	17K	4K	19K	15K	1K	0	11K	12K	7K	10K	4K	13K	21K	16K	23K	18K	25K	24K	17K	8K	14K	18K	12K	22K	20K	12K	9K	13K	790	22K	4K	2K	14K	2K	805	19K	16K
hr	17K	252K	245K	119K	14K	249K	138K	242K	250K	237K	230K	226K	23K	230K	250K	179K	269K	251K	245K	57K	253K	45K	0	246K	5K	243K	246K	245K	282K	84K	3K	272K	243K	154K	246K	109K	242K	242K	249K	245K	232K	230K	236K	248K	153K	253K	247K	238K	233K	255K	21K	20K	261K	11K	235K	192K	22K	245K	199K	229K	260K	261K
hu	21K	365K	363K	157K	19K	360K	195K	356K	369K	363K	342K	331K	35K	348K	360K	274K	360K	379K	360K	71K	368K	50K	360K	0	4K	352K	358K	364K	350K	108K	4K	349K	343K	213K	351K	149K	327K	379K	374K	362K	365K	350K	347K	358K	200K	357K	366K	341K	343K	375K	21K	19K	345K	15K	339K	274K	25K	348K	291K	325K	373K	369K
hy	0	6K	3K	0	0	216	0	8K	0	6K	8K	8K	3K	8K	526	0	0	3K	3K	0	3K	0	5K	3K	0	455	0	826	0	0	0	0	0	0	520	154	0	3K	0	7K	6K	8K	8K	3K	0	202	585	425	8K	3K	0	0	0	0	8K	0	0	0	0	0	4K	3K
id	4K	132K	133K	60K	7K	134K	50K	130K	132K	131K	132K	126K	11K	133K	128K	63K	134K	132K	134K	17K	132K	20K	129K	131K	394	0	92K	130K	120K	29K	1K	88K	113K	65K	129K	48K	117K	134K	0	134K	132K	132K	130K	129K	65K	130K	130K	125K	129K	133K	10K	9K	129K	6K	130K	56K	13K	125K	68K	68K	134K	134K
is	0	251K	249K	30K	6K	197K	22K	245K	249K	247K	248K	236K	2K	245K	242K	35K	116K	252K	249K	5K	248K	6K	246K	243K	0	175K	0	247K	61K	16K	2K	33K	111K	60K	139K	24K	96K	252K	228K	250K	245K	247K	244K	212K	30K	142K	243K	121K	246K	255K	3K	0	75K	2K	243K	28K	2K	125K	25K	16K	177K	131K
it	14K	145K	149K	67K	11K	151K	84K	146K	147K	144K	139K	135K	20K	139K	149K	110K	156K	144K	142K	49K	148K	26K	142K	146K	787	151K	151K	0	138K	42K	3K	135K	146K	89K	156K	64K	138K	149K	150K	148K	147K	140K	140K	151K	78K	158K	145K	153K	138K	146K	10K	9K	139K	12K	138K	110K	14K	150K	111K	100K	153K	159K
ja	2K	77K	78K	28K	0	65K	24K	79K	75K	76K	78K	83K	7K	80K	75K	30K	72K	77K	79K	10K	76K	12K	77K	77K	0	78K	24K	79K	0	15K	4K	42K	39K	25K	58K	18K	52K	78K	76K	79K	76K	79K	78K	77K	28K	62K	77K	41K	78K	76K	7K	4K	60K	6K	78K	26K	6K	63K	18K	22K	77K	63K
ka	0	82K	87K	10K	0	61K	10K	83K	74K	66K	86K	92K	0	89K	65K	14K	59K	80K	77K	4K	81K	2K	81K	82K	0	54K	16K	61K	31K	0	0	20K	28K	10K	49K	8K	25K	84K	58K	85K	81K	87K	88K	63K	9K	28K	78K	33K	83K	73K	0	0	37K	2K	86K	13K	4K	42K	2K	7K	55K	35K
kk	0	2K	5K	0	0	3K	0	2K	5K	5K	4K	5K	2K	4K	2K	0	2K	5K	2K	0	4K	0	4K	4K	0	2K	2K	4K	2K	0	0	0	2K	2K	0	0	2K	3K	4K	6K	5K	4K	4K	4K	0	2K	4K	0	5K	5K	0	0	2K	0	4K	2K	0	2K	0	0	2K	0
ko	0	82K	82K	23K	0	56K	19K	81K	70K	73K	81K	87K	1K	83K	60K	22K	60K	79K	83K	10K	84K	10K	83K	82K	0	61K	18K	84K	45K	11K	0	0	31K	20K	40K	14K	41K	82K	60K	83K	83K	85K	84K	79K	20K	47K	73K	26K	81K	71K	6K	1K	40K	1K	83K	20K	6K	49K	14K	16K	58K	49K
lt	1K	312K	312K	69K	0	252K	37K	304K	311K	296K	309K	296K	3K	305K	306K	76K	226K	313K	311K	18K	307K	23K	306K	303K	0	269K	148K	303K	140K	31K	2K	93K	0	124K	203K	55K	165K	314K	270K	312K	304K	307K	303K	293K	70K	214K	308K	173K	307K	316K	15K	13K	136K	3K	301K	63K	7K	195K	31K	43K	269K	213K
lv	2K	165K	168K	33K	0	124K	14K	167K	157K	150K	166K	170K	2K	172K	156K	34K	104K	168K	163K	8K	167K	12K	166K	167K	0	137K	72K	155K	65K	12K	2K	43K	108K	0	93K	27K	81K	167K	131K	170K	169K	169K	171K	158K	30K	106K	157K	85K	165K	147K	5K	5K	67K	0	170K	31K	2K	100K	17K	19K	125K	106K
mk	6K	227K	226K	71K	2K	220K	54K	221K	225K	214K	228K	214K	2K	220K	218K	63K	216K	226K	222K	22K	222K	16K	217K	218K	514	210K	128K	213K	144K	42K	0	85K	141K	73K	0	62K	148K	231K	216K	230K	222K	225K	219K	214K	74K	206K	221K	169K	216K	228K	7K	3K	162K	6K	217K	69K	13K	195K	41K	57K	226K	199K
ml	5K	244K	223K	66K	3K	142K	10K	233K	177K	197K	241K	268K	3K	251K	162K	51K	148K	202K	222K	9K	210K	9K	210K	222K	129	162K	41K	185K	75K	12K	0	42K	73K	34K	120K	0	89K	202K	146K	224K	218K	254K	242K	180K	39K	98K	190K	77K	227K	199K	6K	0	77K	7K	246K	21K	2K	118K	9K	17K	185K	132K
ms	1K	117K	118K	40K	0	93K	16K	115K	113K	111K	116K	116K	3K	116K	110K	30K	110K	115K	114K	9K	113K	14K	114K	115K	0	112K	50K	111K	70K	15K	1K	49K	63K	37K	80K	26K	0	117K	101K	118K	114K	117K	117K	107K	43K	86K	113K	65K	115K	114K	6K	4K	69K	5K	115K	28K	10K	94K	28K	24K	114K	91K
nl	14K	130K	128K	60K	9K	135K	73K	125K	132K	128K	117K	116K	14K	119K	134K	102K	136K	130K	123K	28K	135K	24K	123K	128K	2K	134K	132K	131K	133K	43K	2K	133K	130K	82K	133K	56K	128K	0	132K	129K	125K	118K	117K	139K	77K	134K	132K	127K	120K	132K	12K	10K	131K	7K	122K	100K	13K	131K	129K	132K	139K	140K
no	5K	179K	174K	65K	8K	179K	58K	171K	171K	172K	166K	160K	10K	167K	179K	86K	173K	171K	171K	19K	180K	21K	167K	174K	0	0	154K	177K	161K	37K	3K	115K	148K	81K	171K	50K	155K	173K	0	176K	169K	168K	165K	177K	70K	174K	180K	153K	166K	173K	8K	9K	161K	3K	166K	74K	12K	171K	78K	78K	186K	180K
pl	20K	262K	258K	125K	18K	265K	157K	253K	266K	260K	247K	240K	34K	249K	257K	206K	267K	263K	260K	70K	266K	44K	259K	258K	7K	251K	264K	268K	262K	89K	5K	264K	260K	168K	257K	122K	249K	264K	268K	0	257K	249K	247K	261K	158K	265K	260K	256K	248K	266K	20K	18K	262K	12K	245K	212K	23K	253K	260K	278K	270K	269K
pt	18K	157K	154K	76K	12K	156K	88K	154K	156K	158K	150K	144K	19K	149K	152K	123K	155K	157K	154K	43K	156K	27K	148K	157K	4K	154K	151K	158K	146K	54K	4K	147K	147K	101K	153K	75K	142K	156K	157K	157K	0	147K	148K	158K	92K	156K	154K	143K	149K	161K	14K	14K	145K	10K	146K	113K	17K	148K	129K	142K	160K	162K
pt_br	20K	137K	135K	69K	11K	137K	84K	131K	139K	130K	126K	125K	25K	128K	136K	119K	137K	135K	129K	58K	143K	30K	124K	130K	6K	133K	131K	131K	128K	50K	3K	128K	127K	88K	138K	73K	125K	132K	137K	130K	129K	0	125K	133K	85K	137K	137K	126K	126K	140K	15K	12K	128K	11K	127K	112K	15K	130K	126K	142K	140K	140K
ro	24K	184K	177K	90K	12K	178K	107K	169K	182K	169K	168K	164K	28K	170K	180K	150K	180K	179K	174K	48K	185K	36K	166K	169K	7K	176K	170K	175K	169K	65K	3K	169K	164K	113K	179K	92K	166K	171K	178K	172K	168K	172K	0	186K	112K	177K	180K	157K	169K	185K	19K	17K	167K	11K	165K	145K	18K	171K	168K	179K	185K	183K
ru	21K	296K	288K	133K	15K	292K	133K	286K	310K	292K	282K	249K	32K	260K	297K	186K	300K	308K	286K	56K	290K	35K	287K	283K	3K	293K	249K	291K	289K	74K	4K	268K	274K	166K	297K	111K	270K	308K	307K	298K	291K	270K	282K	0	148K	297K	292K	288K	286K	306K	13K	13K	285K	15K	285K	214K	21K	290K	137K	171K	309K	325K
si	2K	160K	160K	32K	0	100K	10K	155K	136K	125K	159K	182K	3K	165K	128K	25K	111K	149K	144K	7K	151K	12K	149K	157K	0	129K	32K	134K	76K	9K	0	38K	57K	26K	85K	25K	87K	152K	107K	163K	153K	166K	165K	131K	0	68K	142K	57K	157K	145K	7K	3K	59K	4K	162K	18K	6K	91K	12K	15K	115K	88K
sk	9K	278K	279K	68K	7K	276K	65K	268K	285K	286K	286K	270K	9K	277K	275K	105K	265K	283K	290K	24K	285K	27K	275K	276K	195	265K	165K	288K	192K	33K	2K	130K	186K	107K	262K	63K	188K	288K	286K	289K	279K	280K	273K	289K	75K	0	280K	200K	277K	284K	11K	10K	249K	2K	279K	101K	13K	258K	98K	97K	289K	274K
sl	14K	223K	220K	102K	14K	224K	106K	225K	226K	223K	223K	220K	23K	225K	223K	148K	227K	229K	229K	42K	227K	30K	222K	226K	541	220K	216K	224K	223K	68K	3K	186K	214K	130K	220K	90K	202K	232K	228K	230K	221K	227K	221K	222K	117K	228K	0	212K	221K	228K	15K	14K	218K	10K	222K	146K	19K	215K	152K	171K	232K	228K
sq	2K	218K	214K	50K	7K	204K	26K	210K	210K	205K	215K	204K	5K	211K	210K	33K	175K	218K	212K	12K	214K	20K	210K	209K	399	196K	108K	210K	94K	30K	0	46K	114K	64K	167K	39K	112K	216K	186K	216K	209K	214K	209K	204K	48K	154K	209K	0	210K	213K	11K	10K	115K	7K	208K	44K	9K	149K	38K	43K	207K	145K
sr	20K	302K	289K	153K	14K	279K	171K	281K	302K	274K	266K	251K	36K	261K	306K	231K	320K	298K	264K	79K	319K	56K	268K	269K	7K	295K	289K	272K	344K	108K	6K	331K	308K	192K	311K	148K	308K	288K	300K	282K	279K	261K	271K	293K	200K	288K	298K	295K	0	306K	24K	22K	310K	11K	270K	237K	26K	299K	220K	263K	309K	319K
sv	7K	168K	167K	73K	7K	175K	68K	163K	168K	167K	163K	158K	13K	161K	172K	113K	170K	170K	165K	24K	176K	26K	165K	168K	2K	171K	171K	169K	167K	46K	3K	144K	167K	89K	173K	67K	159K	172K	171K	167K	167K	163K	163K	177K	89K	175K	168K	165K	164K	0	11K	10K	166K	6K	164K	95K	14K	166K	122K	122K	181K	178K
ta	2K	20K	17K	4K	0	11K	0	15K	14K	17K	15K	21K	0	17K	12K	5K	12K	19K	18K	0	20K	19K	16K	16K	0	13K	3K	11K	15K	0	0	11K	12K	5K	7K	5K	9K	19K	11K	16K	15K	20K	20K	12K	7K	10K	13K	11K	16K	13K	0	11K	12K	0	16K	0	0	11K	3K	887	19K	17K
te	0	13K	13K	0	0	10K	0	12K	13K	10K	13K	15K	0	14K	11K	1K	8K	13K	10K	0	13K	12K	13K	12K	0	10K	0	9K	6K	0	0	1K	9K	4K	3K	0	5K	13K	10K	13K	13K	13K	14K	9K	3K	8K	12K	9K	13K	11K	12K	0	2K	0	13K	0	0	8K	1K	0	12K	9K
th	3K	154K	162K	40K	0	168K	41K	159K	163K	165K	161K	156K	7K	157K	159K	34K	159K	164K	164K	14K	162K	17K	162K	158K	0	161K	54K	161K	119K	24K	1K	62K	76K	40K	131K	30K	88K	163K	164K	165K	159K	161K	158K	162K	39K	158K	161K	92K	161K	168K	9K	2K	0	7K	156K	48K	8K	131K	35K	43K	165K	123K
tl	0	8K	8K	2K	0	3K	0	8K	3K	11K	10K	13K	3K	11K	3K	3K	8K	7K	11K	0	8K	846	8K	10K	0	7K	2K	11K	10K	2K	0	2K	2K	0	5K	4K	7K	9K	4K	8K	10K	11K	9K	11K	4K	2K	7K	6K	7K	7K	0	0	8K	0	12K	2K	0	7K	0	0	8K	2K
tr	28K	303K	300K	147K	20K	298K	182K	296K	307K	296K	290K	280K	41K	292K	296K	252K	304K	311K	301K	98K	308K	50K	292K	293K	8K	297K	294K	299K	299K	105K	4K	297K	287K	192K	296K	158K	280K	306K	304K	298K	294K	295K	290K	306K	190K	308K	302K	287K	291K	315K	22K	20K	295K	19K	0	249K	27K	294K	279K	306K	310K	305K
uk	5K	271K	286K	36K	3K	159K	45K	301K	212K	216K	288K	316K	7K	310K	200K	42K	148K	264K	268K	12K	263K	8K	267K	285K	0	158K	37K	270K	103K	15K	2K	51K	67K	36K	114K	18K	76K	277K	161K	295K	266K	306K	297K	263K	25K	133K	239K	78K	275K	214K	0	0	98K	3K	291K	0	3K	119K	30K	28K	169K	114K
ur	0	14K	13K	6K	0	7K	0	13K	12K	12K	13K	14K	0	14K	13K	3K	11K	13K	13K	2K	13K	2K	13K	13K	0	13K	2K	12K	8K	3K	0	6K	4K	2K	7K	1K	11K	14K	12K	14K	13K	14K	14K	11K	4K	8K	13K	7K	14K	13K	0	0	6K	0	14K	3K	0	9K	0	0	11K	9K
vi	3K	104K	102K	41K	0	105K	31K	100K	105K	103K	101K	97K	3K	100K	105K	38K	105K	104K	101K	12K	102K	13K	100K	100K	0	102K	52K	102K	81K	19K	1K	56K	66K	40K	95K	29K	90K	103K	103K	103K	101K	100K	99K	103K	39K	100K	103K	76K	101K	105K	7K	6K	82K	6K	97K	36K	7K	0	37K	36K	107K	104K
ze_en	2K	90K	78K	8K	0	57K	16K	101K	92K	76K	78K	136K	1K	120K	68K	11K	37K	95K	101K	3K	70K	2K	92K	103K	0	66K	15K	91K	25K	2K	0	15K	14K	9K	23K	7K	27K	115K	58K	132K	107K	123K	118K	60K	8K	43K	86K	23K	87K	90K	2K	982	31K	0	111K	12K	0	45K	0	116K	67K	43K
ze_zh	0	961K	913K	16K	0	329K	47K	1M	694K	361K	794K	1M	4K	1M	443K	26K	257K	768K	677K	4K	716K	870	789K	881K	0	366K	20K	543K	129K	9K	0	58K	50K	20K	120K	13K	91K	1M	308K	1M	973K	1M	1M	527K	19K	196K	679K	92K	813K	635K	872	0	161K	0	1M	31K	0	179K	855K	0	1M	1M
zh_cn	8K	206K	220K	95K	7K	215K	67K	221K	215K	218K	221K	214K	13K	225K	208K	115K	211K	216K	221K	32K	210K	34K	217K	218K	3K	213K	153K	218K	230K	42K	1K	152K	195K	96K	225K	76K	217K	225K	227K	224K	215K	216K	217K	218K	95K	216K	208K	216K	218K	219K	15K	13K	223K	7K	216K	105K	17K	217K	107K	201K	0	176K
zh_tw	4K	196K	205K	70K	0	209K	41K	201K	203K	207K	206K	203K	10K	205K	198K	96K	189K	204K	209K	24K	201K	25K	203K	201K	3K	192K	106K	202K	152K	26K	0	108K	137K	78K	174K	60K	159K	204K	202K	206K	202K	205K	200K	207K	69K	196K	198K	133K	203K	205K	15K	7K	151K	2K	202K	59K	13K	198K	63K	197K	191K	0

Table 8: Bilingual lexicon counts for the entire *word2word* dataset. Each (row, column) entry is the number of words in the (row → column) bilingual lexicon. See `http://opus.nlpl.eu/OpenSubtitles2018.php` for language codes and original data sizes.