1. Introduction

Bilingual lexicons (Fung, 1998) are valuable resources for cross-lingual tasks, including low-resource machine translation (Ramesh and Sankaranarayanan, 2018; Gú 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 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.

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 (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.org2 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

---

1 Equal contribution.

2 http://opus.nlpl.eu/OpenSubtitles-v2018.php
   http://www.opensubtitles.org/

---

Keywords: bilingual lexicon, word translation, Python toolkit

Table 1: Overview of the word2word dataset.

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

---

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

---

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.
| 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 2 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 \(p(y|x) \propto p(x, y)\):

\[
p(y|x) = \frac{p(x, y)}{p(x)} \approx \frac{\#(x, y)}{\#(x)} \propto \#(x, y)
\]

(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       |
| whether    | plaise             |
| committee  | comité             |
| clown      | clown              |
| spread     | dispersez-vous     |
|            |                    |
| French     | English            |
| hobbs      | hobbs              |
| mêlé        | mixed              |
| établir     | establish          |
| taulé       | slammer            |
| chaussettes | socks              |
|            |                    |
| Korean     | English            |
| slaughtered | 학살               |
| shadow     | 그림자              |
| Charles    | 찰스               |
| concerns   | 걱정               |
| reverse    | 역                 |
|            |                    |

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 $y$:

$$PMI(x, y) = \log \frac{p(x, y)}{p(x)p(y)}$$

$$\approx \log \frac{\#(x, y)}{\#(x)\#(y)} \propto \log #(x, y) - \log #(y)$$

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 an 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...
Formally, we define the corrected CPE score as follows:

\[
CPE(y \mid x) = p(y \mid x) - \sum_{x' \in X} p(y \mid x')p(x' \mid x)
\]

\[
= \sum_{x' \in X} CPE_{y|x}(x')p(x' \mid x)
\] (3)

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

\[
CPE_{y|x}(x') = p(y \mid x, x') - p(y \mid x')
\] (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 \(CPE_{y|x}(x') = 0\), then \(x \perp y \mid x'\), 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 \(CPE_{y|x}(x) = 0\), 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\(^5\), 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\(^6\) 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

---

\(^{5}\) 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

\(^{6}\) 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                   | pythonlml               | 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.

...
4.1. Implementation

The word2word package is built entirely using Python 3. The package includes scripts for downloading and pre-processing 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.

```python
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.

```python
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; Gú 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
A Sample Translations from Different Extraction Methods

In Table 7, we compare the BLE methods described in Section 3.1.4 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 3.1.4.
| English | Methods | Top-5 Translations in Spanish | Top-5 Translations in Simplified Chinese |
|---------|---------|-----------------------------|----------------------------------------|
| its     | Co-occurrence PMI CPE | de la que el y propio sus su tierra poder | 的 了 是 我 |
|         |          | sus propio tierra cada     | 政府 国家 失去 由 |
|         |          |                           | 它 将 自己 国家 中 |
| good    | Co-occurrence PMI CPE | que de no bien es buenas noches buenos buena buen buen | 好 的 你 我 很 |
|         |          | bien buenas buen buenas buen bueno | 祝你好运 晚安 好消息 早上好 早安 |
|         |          |                           | 好 很 很不错 早晚好 早上好 |
| mouth   | Co-occurrence PMI CPE | la boca de que no boca cerrada pico mantén abre boca cerrada abre palabras labios | 的 你 我 嘴 了 |
|         |          |                           | 张开嘴 嘴里 大嘴巴 张嘴 嘴巴 |
|         |          |                           | 嘴 嘴里 嘴巴 闭上 闭嘴 |
| library | Co-occurrence PMI CPE | la biblioteca de en que solarización biblioteca soltándola library librería biblioteca la librería pública tarjetas | 图书馆 的 我 在 你 |
|         |          |                           | 英 图书馆 圖書館 藏书室 书房 |
|         |          |                           | 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.
Table 8: Bilingual lexicon counts for the entire word2word dataset. Each (row, column) entry is the number of words in the (row $\rightarrow$ column) bilingual lexicon. See [http://opus.nlpl.eu/OpenSubtitles2018.php](http://opus.nlpl.eu/OpenSubtitles2018.php) for language codes and original data sizes.