word2word: Toward Effective and Scalable Bilingual Lexicon Extraction

Yo Joong Choe[*]
Kakao
yj.c@kakaocorp.com

Kyubyong Park[*]
Kakao Brain
ryan.ai@kakaobrain.com

Dongwoo Kim[*]
Kakao Brain
art.hur@kakaobrain.com

Abstract

We present *word2word*, a novel model for finding cross-lingual word correspondences only using a sentence-level parallel corpus. Our model is based on the observation that not only source and target words but also source words themselves can be highly correlated. Inspired by this, we introduce a new scoring metric that captures the dependence between every source-target word pair. This approach significantly improves upon conventional co-occurrence-based methods across European and non-European languages. Evaluations on human labels yield up to 80% top-1 precision and 91% top-5 precision.

1 Introduction

Bilingual lexicon extraction (Fung, 1998; Déjean et al., 2002; Gaussier et al., 2004) refers to the classical natural language task of finding word-level correspondences from a (parallel) corpus. It can be a challenging yet significant component of important downstream tasks, including cross-lingual word embeddings (Ruder et al., 2017; Levy et al., 2017; Adams et al., 2017) and machine translation (Koehn et al., 2007; Sutskever et al., 2014; Vaswani et al., 2017; Östling and Tiedemann, 2017).

We present *word2word*, an effective and scalable scoring model that uses only sentence-level alignments to find word-level correspondences. Our contributions can be summarized as follows:

- We introduce a new bilingual lexicon extraction model that outperforms popular baselines such as co-occurrences and pointwise mutual information. Our model is computationally efficient and is effective across both European and non-European languages.
- We release all of our results, including our source code, a human-labelled validation set, and a large set of word correspondences across 60+ languages.

2 Related Work

One of the first statistical approaches to bilingual lexicon extraction can be found in (Fung, 1998), in which Fung introduces a statistical methodology for extracting relevant target lexicon for each given source word, and extends the task from using a parallel corpus to using a non-parallel corpus. The latter line of work extends this task to other non-parallel or comparable corpora in (Rapp, 1999; Ravi and Knight, 2011; Morin and Prochasson, 2011; Vulić and Moens, 2016; Zhang et al., 2017).

In the unsupervised setting, recent work based on neural embeddings (Mikolov et al., 2013a; Conneau et al., 2017) often led to competitive results by making use of the syntactic information from large monolingual corpora (e.g. Wikipedia). It is important to note that these unsupervised methods still perform poorly in settings between European and non-European languages (e.g. 32.5% top-1 precision accuracy from English to Chinese reported in Conneau et al. (2017)), perhaps because their distant linguistic structures.

3 Methods

Throughout this paper, we assume that we have a phrase- or sentence-level parallel corpus between a pair of languages (e.g. English-French). Publicly available datasets of this format include the Europarl dataset (Koehn, 2005) and the OpenSubtitles dataset (Lison and Tiedemann, 2016).

For each language pair, we designate one language as the *source* language (denoted by the letter X) and the other as the *target* language (denoted

[*] Equal contribution.

by the letter Y). We use $\mathcal{X}$ and $\mathcal{Y}$ to denote the source and target vocabulary sets, respectively.

Since bilingual word-to-word mappings are hardly one-to-one (Fung, 1998; Somers, 2001), we proceed with a scoring approach where we give a relevance score between every source-target word pair and find target words with the highest scores.

3.1 Co-occurrences

A simple baseline approach for our goal is to estimate the probability of seeing a target word $y \in \mathcal{Y}$, e.g. *pomme*, if the source sentence contains the source word $x \in \mathcal{X}$, e.g. *apple*. This conditional probability is defined as $p(y|x) = \frac{p(x,y)}{p(x)}$.[1]

To estimate the joint probability in the numerator, we can use the number of *(cross-lingual) co-occurrences* of x and y in each sentence pair. Similarly, we can estimate the marginal probability of each x by counting all occurrences of x in the entire data. Together, our estimate of the conditional probability $p(y|x)$ is given by

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

where $\hat{p}$ denotes the empirical probability of p, and $\#(\cdot)$ denotes the number of (co-)occurrences of the word or word pair in each sentence. Once we have this estimate for each y, we can choose the most likely correspondence(s) of x by choosing

$$\hat{y} = \underset{y \in \mathcal{Y}}{\operatorname{argmax}}\, \hat{p}(y \mid x) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}}\, \#(x,y) \tag{2}$$

3.2 Pointwise Mutual Information

The simple co-occurrence model does not take into account that some words, especially stop words, are much more frequent than others, regardless of the context. Pointwise mutual information further accounts for the frequency of any target word by simply normalizing the conditional probability $p(y|x)$ with the marginal probability of the target word itself, i.e. $p(y)$. By taking the logarithm, this normalized probability gives us the following formula and the corresponding sample estimate:

$$\mathsf{PMI}(x,y) = \log \frac{p(x,y)}{p(x)p(y)} \approx \log \frac{\#(x,y)}{\#(x)\#(y)} \tag{3}$$

[1] To be precise, we let x denote the binary random variable that encodes whether or not the word occurs in a sentence.

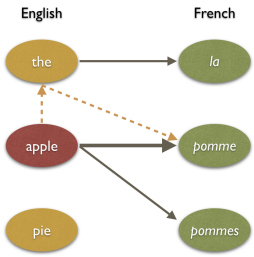

Figure 1: A schematic graphical model of English and French words. The co-occurrence and PMI models only consider the relationship from *apple* to *pomme* (thick gray arrow). *word2word* further controls for the confounding effect of other English words like *the* (dotted yellow arrows).

PMI measures the amount of information about y that is provided by the presence of x, given that x and y co-occurred. PMI has commonly been used to find collocates and associated word pairs in other parts of natural language processing (see, e.g., (Fung, 1998)).

The use of pointwise mutual information can also be viewed as implicitly using a skip-gram with negative sampling (SGNS) model (Levy and Goldberg, 2014), most notably including the skip-gram version of *word2vec* (Mikolov et al., 2013b). PMI can also be interpreted as a conditional version of TF-IDF (Fung, 1998).

3.3 *word2word*

While both the conditional probability and PMI are proportional to cross-lingual co-occurrence counts, they can fail to distinguish exactly which source word x is the most predictive of the candidate target word y. For example, given a sentence pair (*the apple is mine*, *la pomme est á moi*), the baseline methods cannot effectively isolate the effect of *apple* (as opposed to *the*, *is* and *mine*) on *pomme*.

In order to deal with this issue, *word2word* adds 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 yellow dashed arrows in the schematic graphical model in Figure 1– it reflects the conditional independence relationships between words that the baseline models do not.

Formally, we define the corrected *word2word*

Metric (%)	Model	en-es	es-en	en-fr	fr-en	en-de	de-en	en-fi	fi-en	en-tr	tr-en	en-zh	zh-en	en-vi	vi-en
# Sentence Pairs		42M		34M		14M		19M		37M		9.3M		2.2M	
# Validation Words		940	828	933	894	960	934	447	619	811	516	953	839	887	890
P@1	Co-occurrence	19.1	22.1	12.6	8.4	51.8	45.0	17.7	14.2	8.8	5.8	13.3	27.5	46.7	18.9
	PMI	63.5	57.2	51.0	53.4	49.2	54.3	27.5	48.0	33.5	46.5	39.3	39.7	40.2	29.7
	word2word	**80.1**	**79.1**	**71.3**	**67.4**	**76.5**	**72.7**	**54.4**	**67.4**	**70.4**	**59.9**	**68.7**	**56.4**	**62.1**	**55.1**
P@5	Co-occurrence	51.4	67.0	40.3	38.9	69.9	72.5	45.4	53.3	52.5	37.6	59.6	54.7	65.7	47.5
	PMI	89.4	85.9	86.2	81.2	83.2	80.5	55.7	79.2	75.8	78.5	80.1	70.1	77.9	66.3
	word2word	**90.9**	**87.8**	**88.0**	**82.9**	**86.2**	**82.9**	**66.2**	**79.8**	**85.5**	**81.2**	**86.3**	**72.7**	**80.0**	**70.7**

Table 1: Precision (%) on 1,000 sampled words with human labels. 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, fi: Finnish, tr: Turkish, zh: Chinese, vi: Vietnamese).

score as follows:

$$\begin{aligned}\mathsf{w2w}(y \mid x) &= p(y \mid x) - \sum_{x' \in \mathcal{X}} p(y \mid x')p(x' \mid x) \\ &= \sum_{x' \in \mathcal{X}} \mathsf{CPE}_{y|x}(x')p(x' \mid x) \qquad (4)\end{aligned}$$

where $\mathsf{CPE}_{y|x}(x')$ denotes the **controlled predictive effect (CPE)** of any other source word x' when predicting y from x. Formally, the CPE is defined as

$$\mathsf{CPE}_{y|x}(x') = p(y \mid x, x') - p(y \mid x') \qquad (5)$$

The CPE 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 for each confounder x' is then marginalized over all possible confounders to give a final score.

In practice, summing the CPE scores over all words in the source vocabulary is inefficient, because most words in the vocabulary do not play a role in the confounding. Thus, for each candidate target word, we select the top-k source words with the highest co-occurrence counts. We used $k = 100$ in our experiments and found that using a larger k did not impact our top-1 and top-5 predictions in any meaningful way.

4 Experiments

We now compare our *word2word* model (Section 3.3) against the two aforementioned baseline models (Sections 3.1 and 3.2) in real data settings. Further experiment details and analyses can be found in the supplementary material.

4.1 The OpenSubtitles Dataset

We compute all scores using the OpenSubtitles[2] dataset (Lison and Tiedemann, 2016), which contains publicly available parallel corpora of crowd-sourced movie subtitles across 65 different languages. Among many of the publicly available parallel corpora[3], we use this dataset because it covers a large number of both European and non-European languages. As they are movie subtitles, the parallel translations in the OpenSubtitles dataset are naturally made at the phrase- or sentence-level. The number of translations between each pair of languages varies from 3.5M (English-Vietnamese) to 61M (English-Spanish).

4.2 Validation Using Human Labels

In order to measure the performance of our model accurately, we used gold standard labels taken from bilingual human labellers, each of whom hand-labelled the predictions from the three models on 1,000 words in their respective language pairs. For example, a human labeller with native fluency in English and Vietnamese looked at 1,000 English (Vietnamese) words and labelled correct or incorrect the top-5 Vietnamese (English) predictions from both baselines and our model. These gold standard labels were only used during final evaluation and *not* during training.

In Table 1, we report the top-1 and top-5 precision scores of the baseline models and our model across 7 different language pairs. The *word2word* model consistently and significantly outperforms the co-occurrence and PMI models at top-1 precision accuracy, while performing better than or as well as the PMI model in top-5 precision accuracy.

[2] https://www.opensubtitles.org/

[3] e.g., as found in http://opus.nlpl.eu/.

English	Model	Top-5 Predictions in Spanish					Top-5 Predictions in Chinese				
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
	word2word	**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**
	word2word	**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
	word2word	**biblioteca**	la	librería	pública	tarjetas	**图书馆**	书房	里	**圖書館**	去

Table 2: Selected *word2word* translations of English words into Spanish and Chinese. Top-5 predictions are listed in decreasing order of the model's scores. Boldfaced target words indicate correct translations.

4.3 Sample Results

In Table 2, we show examples of word-level correspondences found by the baseline models as well as our model.

4.3.1 Co-occurrences

The baseline co-occurrence model performs poorly in most languages and in both categories (P@1 and P@5).[4] As exemplified in Table 2, 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.

4.3.2 Comparing *word2word* against PMI

The comparison between PMI and *word2word* is more interesting, because the two models rely on different hypotheses: PMI accounts for term frequencies in the *target* language, while *word2word* controls for the effects of other words in the *source* language.

Overall, we find in the error cases 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*).

[4]The co-occurrence model performs surprisingly well in P@1 in English to/from German tasks. While examining this result is an interesting future work on its own, for now we hypothesize that this is due to the similarity between the two languages in terms of conjugation and grammar structures.

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.

5 Conclusion and Future Work

In this paper, we introduce a scalable model that can effectively find word-level correspondences across many language pairs, using one of the various publicly available sentence-level parallel corpora. *word2word* significantly outperforms co-occurrence-based baselines including PMI and is well-founded on linguistic intuition, including the many-to-many relationship between words across languages (see Figure 1). Based on our results, we also open-source our code and results for easy use in downstream tasks such as machine translation and cross-lingual word embeddings.

There are several ways to build upon our work. One interesting direction is to give a measure of confidence based on the *word2word* score. This will allow us to not only tell how trustworthy the correspondence is, but also determine a varying number of correspondences for each word, as opposed to looking at all of the top-k predictions. Another future work is to apply the learned bilingual lexicon to machine translation tasks involving non-European and/or relatively low-resource languages. Because our model is computationally efficient and also effective with small amounts of parallel data (compared to those required by deep learning models), we expect to see both performance gains and interpretability by combining the bilingual lexicon with word alignment models.

References

Oliver Adams, Adam Makarucha, Graham Neubig, Steven Bird, and Trevor Cohn. 2017. Cross-lingual word embeddings for low-resource language modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 937–947.

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

Hervé Déjean, Éric Gaussier, and Fatia Sadat. 2002. An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.

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

Eric Gaussier, J-M Renders, Irina Matveeva, Cyril Goutte, and Hervé Déjean. 2004. A geometric view on bilingual lexicon extraction from comparable corpora. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 526. Association for Computational Linguistics.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, 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.

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

Omer Levy, Anders Søgaard, and Yoav Goldberg. 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.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Emmanuel Morin and Emmanuel Prochasson. 2011. Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora. In *Proceedings of the 4th workshop on building and using comparable corpora: comparable corpora and the web*, pages 27–34. Association for Computational linguistics.

Robert Östling and Jörg Tiedemann. 2017. Neural machine translation for low-resource languages. *arXiv preprint arXiv:1708.05729*.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 519–526. Association for Computational Linguistics.

Sujith Ravi and Kevin Knight. 2011. Deciphering foreign language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 12–21. Association for Computational Linguistics.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2017. A survey of cross-lingual embedding models. *arXiv preprint arXiv:1706.04902*.

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

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Llion Jones, Jakob Uszkoreit, Aidan N Gomez, and Ł ukasz Kaiser. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5851–5861.

Ivan Vulić and Marie-Francine Moens. 2016. Bilingual distributed word representations from document-aligned comparable data. *Journal of Artificial Intelligence Research*, 55:953–994.

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)*, volume 1, pages 1959–1970.

A Supplementary Material

A.1 The OpenSubtitles Dataset

We used the 2016 version of the OpenSubtitles dataset[5] for all our models. There are 65 languages included in this dataset. Upon public release, we will include the learned bilingual lexicon on as many of the 65 language pairs. We expect that our model will perform in a reasonably consistent manner across most languages, including the non-European ones and those with smaller amounts of data.

A.2 Language Pairs

For our experiments, we chose seven languages (in addition to English) that were at least 20MB in size from the OpenSubtitles dataset and were frequently featured in common shared tasks such as WMT. We only considered the translations from/to these languages to/from English, in order to be able to qualitatively evaluate our results. The seven languages and the number of sentence pairs with English are as follows:

- Spanish: 42M sentence pairs (3.4GB)
- French: 34M sentence pairs (2.4GB)
- German: 14M sentence pairs (972MB)
- Finnish: 19M sentence pairs (1.3GB)
- Turkish: 37M sentence pairs (2.5GB)
- Chinese: 9.3M sentence pairs (609MB)
- Vietnamese: 2.2M sentence pairs (166MB)

A.3 Tokenization Softwares

The choice of tokenization softwares across different languages is a crucial design choice for our models, and the results may differ significantly in languages with non-whitespace-based tokenization strategies (e.g. the CJK family). We used the following tokenization softwares, all of which are packages in Python:

- Chinese: `jieba` v0.39[6]
- Vietnamese: `pyvi` v0.0.8.0[7]
- German: split by whitespaces, cases kept

[5] http://opus.nlpl.eu/OpenSubtitles2016.php

[6] `list(jieba.cut(sent, cut_all=False))`

[7] `pyvi.pyvi.ViTokenizer.tokenize(sent).split()`

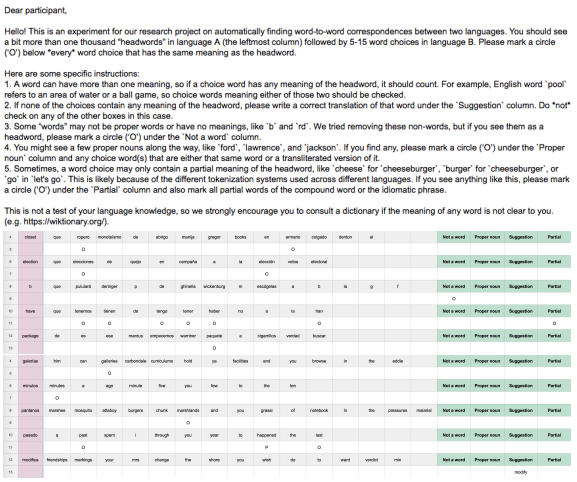

Figure 2: Screenshots of surveys that were sent to bilingual human labellers.

- English, Spanish, French, Finnish, Turkish: split by whitespaces, cases ignored

A.4 Labeling Methodology

For each of the seven languages we chose, we asked a human labeller with native fluency in both the chosen language and English to fill out a survey in which the labeller is asked to check if any of the candidate words is a correct translation of the query word on the far left column. We selected 1,000 words in both languages (see Section A.5 for how we chose these 1,000 words) and generated at most 15 candidates, which were the union of the top-5 predictions from the three models we tested – co-occurrences, PMI, and *word2word*.

For each query word, we also gave four slots for special cases: not a word, proper noun, partial, suggestion. Labelers were asked to mark one of the first three cases when appropriate; we explain the partial meaning category in detail below. We excluded non-words (e.g. typos such as *b* and *rd*) from our evaluation, because they were not properly defined vocabulary words, while we included proper nouns, because in most cases they have direct or transliterated correspondences (e.g. John-John in en-es, Bob-鲍勃 in en-zh). When none of the candidate answers were correct correspondences, labellers were also asked to write down a correct translation in the suggestion box. For evaluation purposes these suggestions were not utilized.

A.4.1 Partial Meanings

Depending on tokenization or the differences in grammatical structures between languages, there

can be more than one target word that corresponds to a source word, and vice versa. This problem also occurs commonly when tokenizers between different language split compound words with different granularity.

For example, the English word *cheeseburger* can be decomposed into *cheese* and *burger*, in which case there can be no exact single correspondence in the target language. Another example is the Spanish word *tenemos*, which translates to *we have*.

In these cases, the evaluators were asked to check the "Partial" box and mark all parts or supersets of the query word – unless there was another correct answer that contained the exact meaning. In our main results, we removed all query words marked as "Partial".

A.5 Choosing a Representative Validation Set for Human Labelling

The 1,000 words in each language are sampled from a smoothed word frequency distribution across all sentences in that language, where the smoothing is made on a similar manner to label smoothing in knowledge distillation (Hinton et al., 2015). Specifically, for each language, we denote the empirical distribution with probability $p(x) = \#(x)/N$ for each $x \in \mathcal{X}$, with $\#(x)$ being the total number of occurrences of the word $x \in \mathcal{X}$ and N being the number of total word occurrences in that language. Let $z(x) = \log p(x) = \log n(x) - \log N$ be the corresponding log-counts, which in this context are interpreted as logits. Then, the smoothed distribution is defined by the following probabilities:

$$q(x) = \frac{\exp(z(x)/T)}{\sum_{x \in \mathcal{X}} \exp(z(x)/T)}$$

with some global temperature parameter $T > 0$. Note that $q(x) = p(x)$ when $T = 1.0$. The use of this temperature parameter encourages the less frequent words to enter the gold standard set rather than just stop words, and we want our database to include good lexical correspondences not only for the most frequent words but also for the less common words. We found *before conducting human evaluation* that $T = 2.0$ allows for a sufficiently diverse (in terms of frequencies) selection of words.

Empirically, this approach gives more diverse (again, in terms of frequencies) set of words than other methods such as adding pseudocounts (corresponding to the Dirichlet-multinomial posterior mean) or using the *maximum a posteriori (MAP)* estimate of the Dirichlet-multinomial model.

A.6 Challenges for a Systematic Validation using MUSE

Multilingual Unsupervised and Supervised Embeddings (MUSE)[8] is a neural word embeddings library based on (Conneau et al., 2017), that also contains 110 bilingual dictionaries labelled using an internal translation tool.

While we also attempted to validate our method against MUSE, we found that direct comparison with MUSE on the dataset is difficult for two reasons. First, our coverage of the query words in the MUSE validation set is low, in some cases due to the lack of variety of words in the OpenSubtitles dataset and in other cases due to the prevalence of proper nouns in the MUSE validation set. For example, in the English-Vietnamese validation set (`en-vi.5000-6500.txt`), we found that 1,199 (80%) out of the 1,500 English words have the same English word as its only Vietnamese translation: *crimson* is listed as the only Vietnamese translation for the English word *crimson*, *precious* for *precious*, and many other cases are simply proper nouns like *Leningrad*, *Suzuki*, and *Randall*.

Second, we also found that the labels are not as reliable as that of humans' or Google Translate's. We believe that a large part of this is due to the use of an internal translation tool rather than human validation or crowdsourcing. We note below that, in certain language pairs, the labels more than often did not agree with any of Google Translate's "common translation."[9]

- English to Chinese: 36.2%
- English to Vietnamese: 41.4%

In other cases we tested, in particular within European languages, the agreement ranged from 75% to 95%.

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

[9] Google Translate rates each of its word-level translations as common, uncommon, or rare.