Bilingual Lexicon Induction via Unsupervised Bitext Construction and Word Alignment

Haoyue Shi *
TTI-Chicago
freda@ttic.edu

Luke Zettlemoyer
University of Washington
Facebook AI Research
lsz@fb.com

Sida I. Wang
Facebook AI Research
sida@fb.com

Abstract

Bilingual lexicons map words in one language to their translations in another, and are typically induced by learning linear projections to align monolingual word embedding spaces. In this paper, we show it is possible to produce much higher quality lexicons with methods that combine (1) unsupervised bitext mining and (2) unsupervised word alignment. Directly applying a pipeline that uses recent algorithms for both subproblems significantly improves induced lexicon quality and further gains are possible by learning to filter the resulting lexical entries, with both unsupervised and semi-supervised schemes. Our final model outperforms the state of the art on the BUCC 2020 shared task by 14 F_1 points averaged over 12 language pairs, while also providing a more interpretable approach that allows for rich reasoning of word meaning in context. Further analysis of our output and the standard reference lexicons suggests they are of comparable quality, and new benchmarks may be needed to measure further progress on this task.[1]

1 Introduction

Bilingual lexicons map words in one language to their translations in another, and can be automatically induced by learning linear projections to align monolingual word embedding spaces (Artetxe et al., 2016; Smith et al., 2017; Lample et al., 2018, *inter alia*). Although very successful in practice, the linear nature of these methods encodes unrealistic simplifying assumptions (e.g. all translations of a word have similar embeddings). In this paper, we show it is possible to produce much higher quality lexicons without these restrictions by introducing new methods that combine (1) unsupervised bitext mining and (2) unsupervised word alignment.

We show that simply pipelining recent algorithms for unsupervised bitext mining (Tran et al., 2020) and unsupervised word alignment (Sabet et al., 2020) significantly improves bilingual lexicon induction (BLI) quality, and that further gains are possible by learning to filter the resulting lexical entries. Improving on a recent method for doing BLI via unsupervised machine translation (Artetxe et al., 2019), we show that unsupervised mining produces better bitext for lexicon induction than translation, especially for less frequent words.

These core contributions are established by systematic experiments in the class of bitext construction and alignment methods (Figure 1). Our full induction algorithm filters the lexicon found via the initial unsupervised pipeline. The filtering can be either fully unsupervised or weakly-supervised: for the former, we filter using simple heuristics and global statistics; for the latter, we train a multi-layer perceptron (MLP) to predict the probability of a word pair being in the lexicon, where the features are global statistics of word alignments.

In addition to BLI, our method can also be directly adapted to improve word alignment and reach competitive or better alignment accuracy than the state of the art on all investigated language pairs. We find that improved alignment in sentence representations (Tran et al., 2020) leads to better contextual word alignments using local similarity (Sabet et al., 2020).

Our final BLI approach outperforms the previous state of the art on the BUCC 2020 shared task (Rapp et al., 2020) by 14 F_1 points averaged over 12 language pairs. Manual analysis shows that most of our false positives are due to the incompleteness of the reference and that our lexicon is comparable to the reference lexicon and the output of a supervised system. Because both of our key building blocks make use of the pretrainined contextual representations from mBART (Liu et al.,

*Work done during internship at Facebook AI Research.
[1]Code is publicly available at https://github.com/facebookresearch/bitext-lexind.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 813–826
August 1–6, 2021.

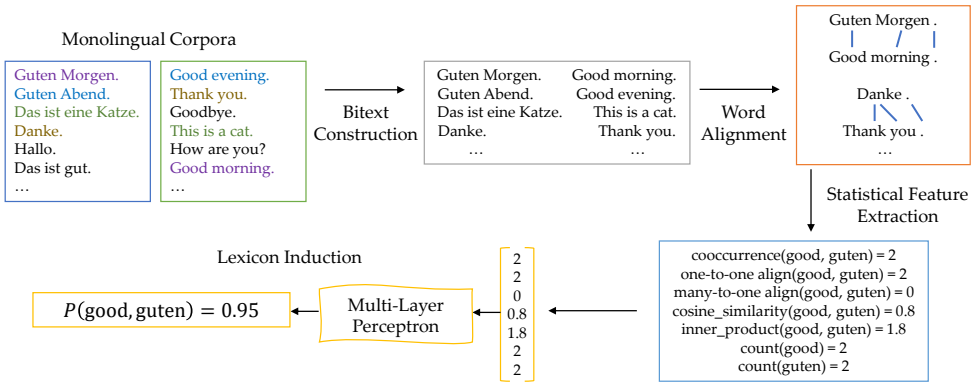


Figure 1: Overview of the proposed retrieval–based supervised BLI framework. Best viewed in color.

2020) and CRISS (Tran et al., 2020), we can also interpret these results as clear evidence that lexicon induction benefits from contextualized reasoning at the token level, in strong contrast to nearly all existing methods that learn linear projections on word types.

2 Related Work

Bilingual lexicon induction (BLI). The task of BLI aims to induce a bilingual lexicon (i.e., word translation) from comparable monolingual corpora (e.g., Wikipedia in different languages). Following Mikolov et al. (2013), most methods train a linear projection to align two monolingual embedding spaces. For supervised BLI, a seed lexicon is used to learn the projection matrix (Artetxe et al., 2016; Smith et al., 2017; Joulin et al., 2018). For unsupervised BLI, the projection matrix is typically found by an iterative procedure such as adversarial learning (Lample et al., 2018; Zhang et al., 2017), or iterative refinement initialized by a statistical heuristics (Hoshen and Wolf, 2018; Artetxe et al., 2018). Artetxe et al. (2019) show strong gains over previous works by word aligning bitext generated with unsupervised machine translation. We show that retrieval-based bitext mining and contextual word alignment achieves even better performance.

Word alignment. Word alignment is a fundamental problem in statistical machine translation, of which the goal is to align words that are translations of each in within parallel sentences (Brown et al., 1993). Most methods assume parallel sentences for training data (Och and Ney, 2003; Dyer et al., 2013; Peter et al., 2017, *inter alia*). In contrast, Sabet et al. (2020) propose `SimAlign`, which does not train on parallel sentences but instead aligns words that have the most similar pre-trained multilingual representations (Devlin et al., 2019; Conneau et al., 2019). `SimAlign` achieves competitive or superior performance than conventional alignment methods despite not using parallel sentences, and provides one of the baseline components for our work. We also present a simple yet effective method to improve performance over `SimAlign` (Section 5).

Bitext mining/parallel corpus mining. Bitext mining has been a long studied task (Resnik, 1999; Shi et al., 2006; Abdul-Rauf and Schwenk, 2009, *inter alia*). Most methods train neural multilingual encoders on bitext, which are then used with efficent nearest neighbor search to expand the training set (Espana-Bonet et al., 2017; Schwenk, 2018; Guo et al., 2018; Artetxe and Schwenk, 2019a, *inter alia*). Recent work has also shown that unsupervised mining is possible (Tran et al., 2020; Keung et al., 2020). We use CRISS (Tran et al., 2020)[2] as one of our component models.

3 Baseline Components

We build on unsupervised methods for word alignment and bitext construction, as reviewed below.

3.1 Unsupervised Word Alignment

`SimAlign` (Sabet et al., 2020) is an unsupervised word aligner based on the similarity of contextualized token embeddings. Given a pair of parallel sentences, `SimAlign` computes embeddings using pretrained multilingual language models such as mBERT and XLM-R, and forms a matrix whose entries are the cosine similarities between every source token vector and every target token vector.

[2] https://github.com/pytorch/fairseq/tree/master/examples/criss

Based on the similarity matrix, the *argmax* algorithm aligns the positions that are the simultaneous column-wise and row-wise maxima. To increase recall, Sabet et al. (2020) also propose *itermax*, which applies *argmax* iteratively while excluding previously aligned positions.

3.2 Unsupervised Bitext Construction

We consider two methods for bitext construction: unsupervised machine translation (generation; Artetxe et al., 2019, Section 3.2) and bitext retrieval (retrieval; Tran et al., 2020, Section 3.2).

Generation Artetxe et al. (2019) train an unsupervised machine translation model with monolingual corpora, generate bitext with the obtained model, and further use the generated bitext to induce bilingual lexicons. We replace their statistical unsupervised translation model with CRISS, a recent high quality unsupervised machine translation model which is expected to produce much higher quality bitext (i.e., translations). For each sentence in the two monolingual corpora, we generate a translation to the other language using beam search or nucleus sampling (Holtzman et al., 2020).

Retrieval Tran et al. (2020) show that the CRISS encoder module provides as a high-quality sentence encoder for cross-lingual retrieval: they take the average across the contextualized embeddings of tokens as sentence representation, perform nearest neighbor search with FAISS (Johnson et al., 2019),[3] and mine bitext using the margin-based max-score method (Artetxe and Schwenk, 2019a).[4]

The score between sentence representations $\mathbf{s}$ and $\mathbf{t}$ is defined by

$$score(\mathbf{s}, \mathbf{t}) \quad (1)$$
$$= \frac{\cos(\mathbf{s}, \mathbf{t})}{\sum_{\mathbf{t}' \in NN_k(\mathbf{t})} \frac{\cos(\mathbf{s}, \mathbf{t}')}{2k} + \sum_{\mathbf{s}' \in NN_k(\mathbf{s})} \frac{\cos(\mathbf{s}', \mathbf{t})}{2k}},$$

where $NN_k(\cdot)$ denotes the set of k nearest neighbors of a vector in the corresponding space. In this work, we keep the top 20% of the sentence pairs with scores larger than 1 as the constructed bitext.

4 Proposed Framework for BLI

Our framework for bilingual lexicon induction takes separate monolingual corpora and the pretrained CRISS model as input, and outputs a list of bilingual word pairs as the induced lexicon. The framework consists of two parts: (i) an unsupervised bitext construction module which generates or retrieves bitext from separate monolingual corpora without explicit supervision (Section 3.2), and (ii) a lexicon induction module which induces bilingual lexicon from the constructed bitext based on the statistics of cross-lingual word alignment. For the lexicon induction module, we compare two approaches: fully unsupervised induction (Section 4.1) which does not use any extra supervision, and weakly supervised induction (Section 4.2) that uses a seed lexicon as input.

4.1 Fully Unsupervised Induction

We align the constructed bitext with CRISS-based `SimAlign`, and propose to use smoothed matched ratio for a pair of bilingual word type $\langle s, t \rangle$

$$\rho(s, t) = \frac{\mathrm{mat}(s, t)}{\mathrm{coc}(s, t) + \lambda}$$

as the metric to induce lexicon, where $\mathrm{mat}(s, t)$ and $\mathrm{coc}(s, t)$ denote the one-to-one matching count (e.g., guten-good; Figure 1) and co-occurrence count of $\langle s, t \rangle$ appearing in a sentence pair respectively, and λ is a non-negative smoothing term.[5]

During inference, we predict the target word t with the highest $\rho(s, t)$ for each source word s. Like most previous work (Artetxe et al., 2016; Smith et al., 2017; Lample et al., 2018, *inter alia*), this method translates each source word to exactly one target word.

4.2 Weakly Supervised Induction

We also propose a weakly supervised method, which assumes access to a seed lexicon. This lexicon is used to train a classifier to further filter the potential lexical entries.

For a pair of word type $\langle s, t \rangle$, our classifier uses the following global features:

- Count of alignment: we consider both one-to-one alignment (Section 4.1) and many-to-one alignment (e.g., danke-you and danke-thank; Figure 1) of s and t separately as two features, since the task of lexicon induction is arguably biased toward one-to-one alignment.
- Count of co-occurrence used in Section 4.1.

[3] https://github.com/facebookresearch/faiss

[4] We used max-score (Artetxe and Schwenk, 2019a) as it strongly outperforms the other methods they proposed.

[5] We use $\lambda = 20$. This reduces the effect of noisy alignment: the most extreme case is that both $\mathrm{mat}(s, t)$ and $\mathrm{coc}(s, t)$ are 1, but it is probably not desirable despite the high matched ratio of 1.

- The count of s in the source language and t in the target language.[6]
- Non-contextualized word similarity: we feed the word type itself into CRISS, use the average pooling of the output subword embeddings, and consider both cosine similarity and dot-product similarity as features.

For a counting feature c, we take $\log(c + \theta_c)$, where θ consists of learnable parameters. There are 7 features in total, which is denoted by $\mathbf{x}_{\langle s,t\rangle} \in \mathbb{R}^7$.

We compute the probability of a pair of words $\langle s, t\rangle$ being in the induced lexicon $P_\Theta(s,t)$[7] by a ReLU activated multi-layer perceptron (MLP):

$$\hat{\boldsymbol{h}}_{\langle s,t\rangle} = \mathrm{ReLU}\left(\mathbf{W}_1 \mathbf{x}_{\langle s,t\rangle} + \mathbf{b}_1\right)$$
$$P_\Theta(s,t) = \sigma\left(\mathbf{w}_2 \cdot \hat{\boldsymbol{h}}_{\langle s,t\rangle} + b_2\right),$$

where $\sigma(\cdot)$ denotes the sigmoid function, and $\Theta = \{\mathbf{W}_1, \mathbf{b}_1, \mathbf{w_2}, b_2\}$ denotes the learnable parameters of the model.

Recall that we are able to access a seed lexicon, which consists of pairs of word translations. In the training stage, we seek to maximize the log likelihood:

$$\Theta^* = \arg\max_\Theta \sum_{\langle s,t\rangle \in \mathcal{D}_+} \log P_\Theta(s,t) + \sum_{\langle s',t'\rangle \in \mathcal{D}_-} \log\left(1 - P_\Theta(s',t')\right),$$

where $\mathcal{D}_+$ and $\mathcal{D}_-$ denotes the positive training set (i.e., the seed lexicon) and the negative training set respectively. We construct the negative training set by extracting all bilingual word pairs that co-occurred but are not in the seed word pairs.

We tune two hyperparameters δ and n to maximize the F_1 score on the seed lexicon and use them for inference, where δ denotes the prediction threshold and n denotes the maximum number of translations for each source word, following Laville et al. (2020) who estimate these hyperparameters based on heuristics. The inference algorithm is summarized in Algorithm 1.

Algorithm 1: Inference algorithm for weakly-supervised lexicon induction.

Input: Thresholds δ, n, Model parameters Θ, source words S
Output: Induced lexicon $\mathcal{L}$
$\mathcal{L} \leftarrow \emptyset$
for $s \in S$ **do**
  $(\langle s,t_1\rangle, \ldots, \langle s,t_k\rangle) \leftarrow$ bilingual word pairs sorted by the descending order of $P_\Theta(s,t_i)$
  $k' = \max\{j \mid P_\Theta(s,t_j) \geq \delta, j \in [k]\}$
  $m = \min(n, k')$
  $\mathcal{L} \leftarrow \mathcal{L} \cup \{\langle s,t_1\rangle, \ldots, \langle s,t_m\rangle\}$
end

5 Extension to Word Alignment

The idea of using an MLP to induce lexicon with weak supervision (Section 4.2) can be directly extended to word alignment. Let $\mathcal{B} = \{\langle \mathcal{S}_i, \mathcal{T}_i\rangle\}_{i=1}^N$ denote the constructed bitext in Section 3.2, where N denotes the number of sentence pairs, and $\mathcal{S}_i$ and $\mathcal{T}_i$ denote a pair of sentences in the source and target language respectively. In a pair of bitext $\langle \mathcal{S}, \mathcal{T}\rangle$, $\mathcal{S} = \langle s_1, \ldots, s_{\ell_s}\rangle$ and $\mathcal{T} = \langle t_1, \ldots, t_{\ell_s}\rangle$ denote sentences consist of word tokens s_i or t_i.

For a pair of bitext, `SimAlign` with a specified inference algorithm produces word alignment $\mathcal{A} = \{\langle a_i, b_i\rangle\}_i$, denoting that the word tokens s_{a_i} and t_{b_i} are aligned. Sabet et al. (2020) has proposed different algorithms to induce alignment from the same similarity matrix, and the best method varies across language pairs. In this work, we consider the relatively conservative (i.e., having higher precision) *argmax* and the higher recall *itermax* algorithm (Sabet et al., 2020), and denote the alignments by $\mathcal{A}_{argmax}$ and $\mathcal{A}_{itermax}$ respectively.

We substitute the non-contextualized word similarity feature (Section 4.2) with contextualized word similarity where the corresponding word embedding is computed by averaging the final-layer contextualized subword embeddings of CRISS. The cosine similarities and dot-products of these embeddings are included as features.

Instead of the binary classification in Section 4.2, we do ternary classification for word alignments. For a pair of word tokens $\langle s_i, t_j\rangle$, the gold label $y_{\langle s_i,t_j\rangle}$ is defined as

$$\mathbb{1}[\langle i,j\rangle \in \mathcal{A}_{argmax}] + \mathbb{1}[\langle i,j\rangle \in \mathcal{A}_{itermax}].$$

Intuitively, the labels 0 and 2 represents confident alignment or non-alignment by both methods, while the label 1 models the potential alignment.

The MLP takes the features $\mathbf{x}_{\langle s_i,t_j\rangle} \in \mathbb{R}^7$ of the word token pair, and compute the probability of

[6] `SimAlign` sometimes mistakenly align rare words to punctuation, and such features can help exclude such pairs.

[7] Not to be confused with joint probability.

each label y by

$$\hat{\boldsymbol{h}} = \mathrm{ReLU}\left(\mathbf{W}_1 \mathbf{x}_{\langle s_i, t_j \rangle} + \mathbf{b}_1\right)$$
$$\boldsymbol{g} = \mathbf{W}_2 \cdot \hat{\boldsymbol{h}} + \mathbf{b}_2$$
$$P_\Phi(y \mid s_i, t_j, \mathcal{S}, \mathcal{T}) = \frac{\exp\left(g_y\right)}{\sum_{y'} \exp\left(g_{y'}\right)},$$

where $\Phi = \{\mathbf{W}_1 \mathbf{W}_2, \mathbf{b}_1, \mathbf{b}_2\}$. On the training stage, we maximize the log-likelihood of ground-truth labels:

$$\Phi^* = \arg\max_{\Phi} \sum_{\langle \mathcal{S}, \mathcal{T} \rangle \in \mathcal{B}} \sum_{s_i \in \mathcal{S}} \sum_{t_j \in \mathcal{T}} \log P_\Phi(y_{\langle s_i, t_j \rangle} \mid s_i, t_j, \mathcal{S}, \mathcal{T}).$$

On the inference stage, we keep all word token pairs $\langle s_i, t_j \rangle$ that have

$$\mathbb{E}_P[y] := \sum_y y \cdot P(y \mid s_i, t_j, \mathcal{S}, \mathcal{T}) > 1$$

as the prediction.

6 Experimental Setup and Baselines

Throughout our experiments, we use a two-layer perceptron with the hidden size of 8 for both lexicon induction and word alignment. We optimize all of our models using Adam (Kingma and Ba, 2015) with the initial learning rate 5×10^{-4}. For our bitext construction methods, we retrieve the best matching sentence or translate the sentences in the source language Wikipedia; for baseline models, we use their default settings.

For evaluation, we use the BUCC 2020 BLI shared task dataset (Rapp et al., 2020) and metric (F_1). Like most recent work, this evaluation is based on MUSE (Lample et al., 2018).[8] We primarily report the BUCC evaluation because it considers recall in addition to precision. However, because most recent work only evaluates on precision, we include those evaluations in Appendix D.

We compare the following baselines:

BUCC. Best results from the BUCC 2020 (Rapp et al., 2020) for each language pairs, we take the maximum F_1 score between the best closed-track results (Severini et al., 2020; Laville et al., 2020) and open-track ones (Severini et al., 2020). Our method would be considered open track since the pretrained models used a much larger data set (Common Crawl 25) than the BUCC 2020 closed-track (Wikipedia or Wacky; Baroni et al., 2009).

VecMap. Popular and robust method for aligning monolingual word embeddings via a linear projection and extracting lexicons. Here, we use the standard implementation[9] with FastText vectors (Bojanowski et al., 2017)[10] trained on the union of Wikipedia and Common Crawl corpus for each language.[11] We include both supervised and unsupervised versions.

WM. WikiMatrix (Schwenk et al., 2019)[12] is a dataset of mined bitext. The mining method LASER (Artetxe and Schwenk, 2019b) is trained on real bitext and then used to mine more bitext from the Wikipedia corpora to get the WikiMatrix dataset. We test our lexicon induction method with WikiMatrix bitext as the input and compare to our methods that do not use bitext supervision.

7 BLI Results and Analysis

7.1 Main Results

We evaluate bidirectional translations from beam search (GEN; Section 3.2), bidirectional translations from nucleus sampling (GEN-N; Holtzman et al., 2020),[13] and retrieval (RTV; Section 3.2). In addition, it is natural to concatenate the global statistical features (Section 4.2) from both GEN and RTV and we refer to this approach by GEN-RTV.

Our main results are presented in Table 1. All of our models (GEN, GEN-N, RTV, GEN-RTV) outperform the previous state of the art (BUCC) by a significant margin on all language pairs. Surprisingly, RTV and GEN-RTV even outperform WikiMatrix by average F_1 score, indicating that we do not need bitext supervision to obtain high-quality lexicons.

7.2 Automatic Analysis

Bitext quality. Since RTV achieves surprisingly high performance, we are interested in how much the quality of bitext affects the lexicon induction performance. We divide all retrieved bitexts with score (Eq. 1) larger than 1 equally into five sections with respect to the score, and compare the lexicon

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

[9] https://github.com/artetxem/VecMap

[10] https://github.com/facebookresearch/fastText

[11] https://github.com/facebookresearch/fastText/blob/master/docs/crawl-vectors.md; that is, our VecMap baselines have the same data availability with our main results.

[12] https://github.com/facebookresearch/LASER/tree/master/tasks/WikiMatrix

[13] We sample from the smallest word set whose cumulative probability mass exceeds 0.5 for next words.

Language Pair	*Weakly-Supervised*							*Unsupervised*		
	BUCC	VECMAP	WM	GEN	GEN-N	RTV	GEN-RTV	VECMAP	GEN	RTV
de-en	61.5	37.1	71.6	70.2	67.7	73.0	**74.2**	22.1	62.6	66.8
de-fr	76.8	43.2	79.8	79.1	79.2	78.9	**83.2**	27.1	79.4	80.3
en-de	54.5	33.2	62.1	62.7	59.3	64.4	**66.0**	33.7	51.0	56.2
en-es	62.6	45.3	71.8	73.7	69.6	**77.0**	75.3	44.1	60.2	65.6
en-fr	65.1	45.4	74.4	73.1	69.9	73.4	**76.3**	44.8	61.9	66.3
en-ru	41.4	29.2	**54.4**	43.5	37.9	53.1	53.1	24.6	28.4	45.4
en-zh	49.5	31.0	67.7	64.3	56.8	**69.9**	68.3	12.8	51.5	51.7
es-en	71.1	55.5	82.3	80.3	75.8	**82.8**	82.6	52.4	71.4	76.4
fr-de	71.0	46.2	**82.1**	80.0	78.7	80.9	81.7	46.0	76.4	77.3
fr-en	53.7	51.5	80.3	79.7	76.1	80.0	**83.2**	50.4	72.7	75.9
ru-en	57.1	44.8	72.7	61.1	59.2	72.7	**72.9**	42.1	51.8	68.0
zh-en	36.9	36.1	**64.1**	52.6	50.6	62.5	62.5	34.4	34.3	48.1
average	58.4	41.5	72.0	68.4	65.1	72.4	**73.3**	36.2	58.5	64.8

Table 1: F_1 scores ($\times 100$) on the BUCC 2020 test set (Rapp et al., 2020). The best number in each row is **bolded**.

Lang.	*Bitext Quality: High → Low*						
	RTV-1	RTV-2	RTV-3	RTV-4	RTV-5	Random	RTV-ALL
de-en	**73.0**	67.9	65.8	64.5	63.1	37.8	70.9
de-fr	78.9	74.2	70.8	69.5	67.3	60.6	**79.4**
en-de	**64.4**	59.7	58.1	56.6	57.2	36.5	62.5
en-es	**77.0**	76.5	73.7	68.4	66.1	43.3	75.3
en-fr	**73.4**	70.5	67.9	65.7	65.5	47.8	68.3
en-ru	**53.1**	48.0	44.2	40.8	41.0	15.0	51.3
en-zh	**69.9**	59.6	66.1	60.1	61.3	48.2	67.6
es-en	**82.8**	82.4	79.6	74.2	72.3	44.4	81.1
fr-de	**80.9**	76.9	73.2	74.7	74.5	64.7	79.1
fr-en	**80.0**	79.0	74.2	72.6	71.6	50.1	79.4
ru-en	**72.7**	66.8	60.5	55.8	54.0	14.7	71.0
zh-en	**62.5**	58.0	54.1	50.9	49.3	13.6	61.3
avg.	**72.4**	68.3	65.7	62.8	61.9	39.7	70.6

Table 2: F_1 scores ($\times 100$) on the test set of the BUCC 2020 shared task (Rapp et al., 2020). We use the weakly supervised algorithm (Section 4.2). The best number in each row is bolded. RTV-1 is the same as RTV in Table 1.

induction performance (Table 2). In the table, RTV-1 refers to the bitext of the highest quality and RTV-5 refers to the ones of the lowest quality, in terms of the margin score (Eq 1).[14] We also add a random pseudo bitext baseline (Random), where all the bitext are randomly sampled from each language pair, as well as using all retrieved sentence pairs that have scores larger than 1 (RTV-ALL).

In general, the lexicon induction performance of RTV correlates well with the quality of bitext. Even using the bitext of the lowest quality (RTV-5), it is still able to induce reasonably good bilingual lexicon, outperforming the best numbers reported by BUCC 2020 participants (Table 1) on average. However, RTV achieves poor performance with random bitext (Table 2), indicating that it is only robust to a reasonable level of noise. While this is a lower-bound on bitext quality, even random bitext does not lead to 0 F_1 since the model may align any co-occurrences of correct word pairs even when they appear in unrelated sentences.

Word alignment quality. We compare the lexicon induction performance using the same set of constructed bitext (RTV) and different word aligners (Table 3). According to Sabet et al. (2020), `SimAlign` outperforms `fast_align` in terms of word alignment. We observe that such a trend translates to resulting lexicon induction performance well: a significantly better word aligner can usually lead to a better induced lexicon.

Bitext quantity. We investigate how the BLI performance changes when the quantity of bitext changes (Figure 2). We use CRISS with nucleus sampling (GEN-N) to create different amount of bitext of the same quality. We find that with only 1% of the bitext (160K sentence pairs on average) used by GEN-N, our weakly-supervised framework outperforms the previous state of the art (BUCC;

[14] See Appendix C for examples from each tier.

Languages	SimAlign	fast_align
de-en	**73.0**	69.7
de-fr	**78.9**	69.1
en-de	**64.4**	61.2
en-es	**77.0**	72.8
en-fr	**73.4**	68.5
en-ru	**53.1**	50.7
en-zh	**69.9**	66.0
es-en	**82.8**	79.8
fr-de	**80.9**	75.8
fr-en	**80.0**	77.3
ru-en	**72.7**	70.2
zh-en	**62.5**	60.2
average	**72.4**	68.4

Table 3: F_1 scores ($\times 100$) on the BUCC 2020 test set. Models are trained with the retrieval–based bitext (RTV), in the weakly-supervised setting (Section 4.2. The best number in each row is bolded.

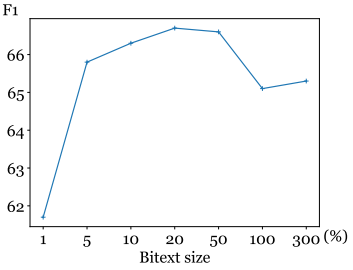
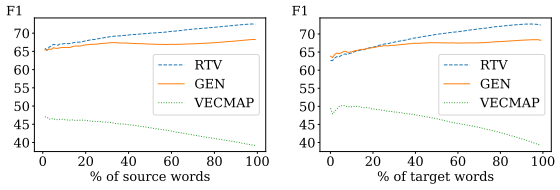


Figure 2: F_1 scores ($\times 100$) on the BUCC 2020 test set, produced by our weakly-supervised framework using different amount of bitext generated by CRISS with nucleus sampling. 100% is the same as GEN-N in Table 1. For less than 100%, we uniformly sample the corresponding amount of bitext; for greater, we generate multiple translations for each source sentence.

Table 1). The model reaches its best performance using 20% of the bitext (3.2M sentence pairs on average) and then drops slightly with even more bitext. This is likely because more bitext introduces more candidates word pairs.

Dependence on word frequency of GEN vs. RTV. We observe that retrieval-based bitext construction (RTV) works significantly better than generation-based ones (GEN and GEN-N), in terms of lexicon induction performance (Table 1). To further investigate the source of such difference, we compare the performance of the RTV and GEN as a function of source word frequency or target word frequency, where the word frequency are computed from the lower-cased Wikipedia corpus. In Figure 3, we plot the F_1 of RTV and GEN when the most frequent $k\%$ of words are considered. When all words are considered RTV outperform GEN for



(a) (b)

Figure 3: Average F_1 scores ($\times 100$) with our weakly-supervised framework across the 12 language pairs (Table 1) on the filtered BUCC 2020 test set. Results on entries with (a) the k% most frequent source words, and (b) the k% most frequent target words.

11 of 12 language pairs except de-fr. In 6 of 12 language pairs, GEN does better than RTV for high frequency source words. As more lower frequency words are included, GEN eventually does worse than RTV. This helps explain why the combined model GEN-RTV is even better since GEN can have an edge in high frequency words over RTV. The trend that $F_1(\text{RTV}) - F_1(\text{GEN})$ increases as more lower frequency words are included seems true for all language pairs (Appendix A).

On average and for the majority of language pairs, both methods do better on low-frequency source words than high-frequency ones (Figure 3a), which is consistent with the findings by BUCC 2020 participants (Rapp et al., 2020).

VECMAP. While BLI through bitext construction and word alignment clearly achieves superior performance than that through vector rotation (Table 1), we further show that the gap is larger on low-frequency words (Figure 3).

7.3 Ground-truth Analysis

Following the advice of Kementchedjhieva et al. (2019) that some care is needed due to the incompleteness and biases of the evaluation, we perform manual analysis of selected results. For Chinese–English translations, we uniformly sample 20 wrong lexicon entries according to the evaluation for both GEN-RTV and weakly-supervised VECMAP. Our judgments of these samples are shown in Table 4. For GEN-RTV, 18/20 of these sampled errors are actually acceptable translations, whereas for VECMAP, only 11/20 are acceptable. This indicates that the improvement in quality may be partly limited by the incompleteness of the reference lexicon and the ground truth performance of our method might be even better. The same analysis for English–Chinese is in Appendix B.

GEN-RTV			VECMAP		
倉庫	depot	✓	申明	endorsing	✗
浪費	wasting	✓	條件	preconditions	?
背面	reverse	✓	移動	moving	✓
嘴巴	mouths	✓	天津	shanghai	✗
可笑	laughable	✓	個案	cases	✓
隱藏	conceal	✓	百合	peony	✗
虔誠	devout	✓	申報	filing	✓
純淨	purified	?	車廂	carriages	✓
截止	deadline	✓	海草	seaweed	✓
對外	foreign	?	履歷	résumé	✓
鍾	clocks	✓	收容所	asylums	✓
努力	effort	✓	開幕	soft-opened	✗
艦	ships	✓	有形	intangible	✗
州	states	✓	小刀	penknife	✓
受傷	wounded	✓	黑山	carpathian	✓
滑動	sliding	✓	象徵	symbolise	✓
毒理學	toxicology	✓	精華	fluff-free	✗
推翻	overthrown	✓	同謀	conspirator	✓
穿	wore	✓	籌碼	bargaining	✗
禮貌	courteous	✓	刮刀	rollers	✗

Table 4: Manually labeled acceptability judgments for random 20 error cases made by GEN-RTV (left) and VECMAP (right). ✓ and ✗ denote acceptable and unacceptable translation respectively. **?** denotes word pairs that may be acceptable in rare or specific contexts.

Data Source	Precision	Recall	F_1
MUSE	93.4	**78.8**	**85.5**
GEN-RTV	**96.6**	71.9	82.5

Table 5: Comparison of Chinese-English lexicons against manually labeled ground truth. The best number in each column is bolded.

Furthermore, we randomly sample 200 source words from the MUSE zh-en test set, and compare the quality between MUSE translation and those predicted by GEN-RTV. This comparison is MUSE-favored since only MUSE source words are included. Concretely, we take the union of word pairs, construct the new ground-truth by manual judgments (i.e., removing unacceptable pairs), and evaluate the F_1 score against the constructed ground-truth (Table 5). The overall gap of 3 F_1 means that a higher quality benchmark is necessary to resolve further improvements over GEN-RTV. The word pairs and judgments are included in the supplementary material (Section F).

8 Word Alignment Results

We evaluate different word alignment methods (Table 6) on existing word alignment datasets,[15] following Sabet et al. (2020). We investigate four language pairs: German–English (de-en), English–French (en-fr), English–Hindi (en-hi) and Romanian–English (ro-en). We find that the CRISS-based `SimAlign` already achieves competitive performance with the state-of-the-art method (Garg et al., 2019) which requires real bitext for training. By ensembling the *argmax* and *itermax* CRISS-based `SimAlign` results (Section 5), we set the new state of the art of word alignment without using any bitext supervision.

However, by substituting the CRISS-based `SimAlign` in the BLI pipeline with our aligner, we obtain an average F_1 score of 73.0 for GEN-RTV, which does not improve over the result of 73.3 achieved by CRISS-based `SimAlign` (Table 1), indicating that further effort is required to take the advantage of the improved word aligner.

Model	de-en	en-fr	en-hi	ro-en
`GIZA++`†	0.22	0.09	0.52	0.32
`fast_align`†	0.30	0.16	0.62	0.32
Garg et al. (2019)	0.16	0.05	N/A	0.23
Zenkel et al. (2019)	0.21	0.10	N/A	0.28
`SimAlign` (Sabet et al., 2020)				
XLM-R-*argmax*†	0.19	0.07	0.39	0.29
mBART-*argmax*	0.20	0.09	0.45	0.29
CRISS-*argmax*∗	0.17	0.05	0.32	0.25
CRISS-*itermax*∗	0.18	0.08	0.30	0.23
MLP (ours)∗	**0.15**	**0.04**	**0.28**	**0.22**

Table 6: Average error rate (AER) for word alignment (lower is better). The best numbers in each column are bolded. Models in the top section require ground-truth bitext, while those in the bottom section do not. ∗: models that involve unsupervised bitext construction. †: results copied from Sabet et al. (2020).

9 Discussion

We present a direct and effective framework for BLI with unsupervised bitext mining and word alignment, which sets a new state of the art on the task. From the perspective of pretrained multilingual models (Conneau et al., 2019; Liu et al., 2020; Tran et al., 2020, *inter alia*), our work shows that they have successfully captured information about word translation that can be extracted using similarity based alignment and refinement. Although BLI is only about word types, it strongly benefits from contextualized reasoning at the token level.

[15] `http://www-i6.informatik.rwth-aachen.de/goldAlignment` (de-en); `https://web.eecs.umich.edu/~mihalcea/wpt` (en-fr and ro-en); `https://web.eecs.umich.edu/~mihalcea/wpt05` (en-hi)

Acknowledgment

We thank Chau Tran for help with pretrained CRISS models, as well as Mikel Artetxe, Kevin Gimpel, Karen Livescu, Jiayuan Mao and anonymous reviewers for their valuable feedback on this work.

References

Sadaf Abdul-Rauf and Holger Schwenk. 2009. On the use of comparable corpora to improve SMT performance. In *Proc. of EACL*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proc. of EMNLP*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proc. of ACL*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. Bilingual lexicon induction through unsupervised machine translation. In *Proc. of ACL*.

Mikel Artetxe and Holger Schwenk. 2019a. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proc. of ACL*.

Mikel Artetxe and Holger Schwenk. 2019b. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *TACL*, 7:597–610.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *TACL*, 5:135–146.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL-HLT*.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. In *Proc. of ICLR*.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proc. of NAACL-HLT*.

Cristina Espana-Bonet, Adám Csaba Varga, Alberto Barrón-Cedeno, and Josef van Genabith. 2017. An empirical analysis of nmt-derived interlingual embeddings and their use in parallel sentence identification. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1340–1350.

Sarthak Garg, Stephan Peitz, Udhyakumar Nallasamy, and Matthias Paulik. 2019. Jointly learning to align and translate with transformer models. In *Proc. of EMNLP*.

Mandy Guo, Qinlan Shen, Yinfei Yang, Heming Ge, Daniel Cer, Gustavo Hernandez Abrego, Keith Stevens, Noah Constant, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Effective parallel corpus mining using bilingual sentence embeddings. In *Proc. of WMT*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *Proc. of ICLR*.

Yedid Hoshen and Lior Wolf. 2018. Non-adversarial unsupervised word translation. In *Proc. of EMNLP*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Trans. on Big Data*.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Herve Jegou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proc. of EMNLP*.

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

Phillip Keung, Julian Salazar, Yichao Lu, and Noah A. Smith. 2020. Unsupervised bitext mining and translation via self-trained contextual embeddings. *TACL*, 8:828–841.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proc. of ICLR*.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proc. of ICLR*.

Martin Laville, Amir Hazem, and Emmanuel Morin. 2020. TALN/LS2N participation at the BUCC shared task: Bilingual dictionary induction from comparable corpora. In *Proc. of Workshop on Building and Using Comparable Corpora*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *arXiv preprint arXiv:2001.08210*.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Jan-Thorsten Peter, Arne Nix, and Hermann Ney. 2017. Generating alignments using target foresight in attention-based neural machine translation. *The Prague Bulletin of Mathematical Linguistics*, 108(1):27–36.

Reinhard Rapp, Pierre Zweigenbaum, and Serge Sharoff. 2020. Overview of the fourth BUCC shared task: Bilingual dictionary induction from comparable corpora. In *Proc. of Workshop on Building and Using Comparable Corpora*.

Philip Resnik. 1999. Mining the web for bilingual text. In *Proc. of ACL*.

Masoud Jalili Sabet, Philipp Dufter, and Hinrich Schutze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of EMNLP*.

Holger Schwenk. 2018. Filtering and mining parallel data in a joint multilingual space. In *Proc. of ACL*.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. *arXiv preprint arXiv:1907.05791*.

Silvia Severini, Viktor Hangya, Alexander Fraser, and Hinrich Schütze. 2020. LMU bilingual dictionary induction system with word surface similarity scores for BUCC 2020. In *Proc. of Workshop on Building and Using Comparable Corpora*.

Lei Shi, Cheng Niu, Ming Zhou, and Jianfeng Gao. 2006. A DOM tree alignment model for mining parallel data from the web. In *Proc. of ACL*.

Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *Proc. of ICLR*.

Chau Tran, Yuqing Tang, Xian Li, and Jiatao Gu. 2020. Cross-lingual retrieval for iterative self-supervised training. In *Proc. of NeurIPS*.

Thomas Zenkel, Joern Wuebker, and John DeNero. 2019. Adding interpretable attention to neural translation models improves word alignment. *arXiv preprint arXiv:1901.11359*.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Adversarial training for unsupervised bilingual lexicon induction. In *Proc. of ACL*.

Appendices

A Language-Specific Analysis

While Figure 3 shows the average trend of F_1 scores with respect to the portion of source words or target words kept, we present such plots for each language pair in Figure 4 and 5. The trend of each separate method is inconsistent, which is consistent to the findings by BUCC 2020 participants (Rapp et al., 2020). However, the conclusion that RTV gains more from low-frequency words still holds for most language pairs.

B Acceptability Judgments for en → zh

GEN-RTV			VECMAP		
southwestern	西南部	✓	spiritism	扶箕	✗
subject	話題	✓	danny	john	✗
screenwriter	劇作家	?	hubbard	威廉斯	✗
preschool	學齡前	✓	swizz	incredible	✗
palestine	palestine	✗	viewing	觀賞	?
strengthening	强化	✓	prohibition	禁令	✓
zero	0	✓	tons	滿載	✗
insurance	保險公司	✗	pascal	帕斯卡	✓
lines	線路	✓	claudia	christina	✗
suburban	市郊	✓	massive	巨大	✓
honorable	尊貴	?	equity	估值	✗
placement	置入	✓	sandy	沙質	✓
lesotho	萊索托	✓	fwd	不過後	✗
shanxi	shanxi	✗	taillight	煞車燈	?
registration	注冊	✓	horoscope	生辰八字	✗
protestors	抗議者	✓	busan	仁川	✗
shovel	剷	✓	hiding	躲藏	✓
side	一方	✓	entry	關時	✗
turbulence	湍流	✓	weekends	雙休日	?
omnibus	omnibus	✗	flagbearer	掌旗	✓

Table 7: Manually labeled acceptability judgments for random 20 error cases in English to Chinese translation made by GEN-RTV and VECMAP.

We present error analysis for the induced lexicon for English to Chinese translations (Table 7) using the same method as Table 4. In this direction, many of the unacceptable cases are copying English words as their Chinese translations, which is also observed by Rapp et al. (2020). This is due to an idiosyncrasy of the evaluation data where many English words are considered acceptable Chinese translations of the same words.

C Examples for Bitext in Different Sections

We show examples of mined bitext with different quality (Table 8), where the mined bitexts are divided into 5 sections with respect to the similarity-based margin score (Eq 1). The Chinese sentences are automatically converted to traditional Chinese alphabets using `chinese_converter`,[16] to keep consistent with the MUSE dataset.

Based on our knowledge about these languages, we see that the RTV-1 mostly consists of correct translations. While the other sections of bitext are of less quality, sentences within a pair are highly related or can be even partially aligned; therefore our bitext mining and alignment framework can still extract high-quality lexicon from such imperfect bitext.

D Results: P@1 on the MUSE Dataset

Precision@1 (P@1) is a widely applied metric to evaluate bilingual lexicon induction (Smith et al., 2017; Lample et al., 2018; Artetxe et al., 2019, *inter alia*), therefore we compare our models with existing approaches in terms of P@1 as well (Table 9). Our fully unsupervised method with retrieval-based bitext outperforms the previous state of the art (Artetxe et al., 2019) by 4.1 average P@1, and achieve competitive or superior performance on all investigated language pairs.

E Error analysis

To understand the remaining errors, we randomly sampled 400 word pairs from the induced lexicon and compare them to ground truth as and Google Translate via `=googletranslate(A1, "zh", "en")`. All error cases are included in Table 10. In overall precision, our induced lexicon is comparable to the output of Google translate API where there are 17 errors for GEN-RTV 14 errors for Google and 4 common errors.

[16] https://pypi.org/project/chinese-converter/

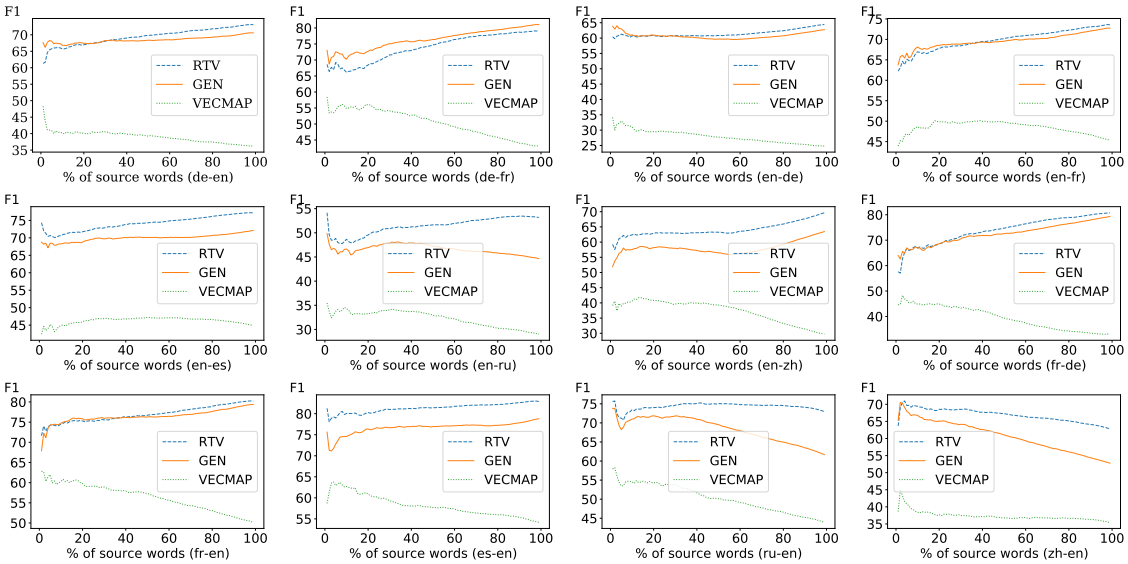


Figure 4: F_1 scores with respect to portion of source words kept for each investigated language pair, analogous to Figure 3a.

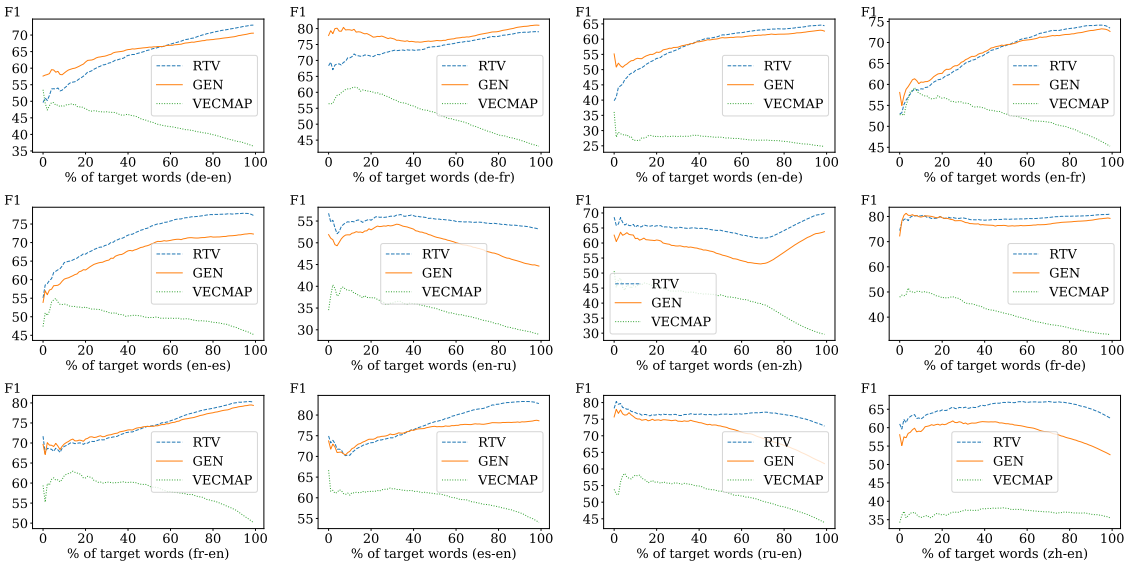


Figure 5: F_1 scores with respect to portion of target words kept for each investigated language pair, analogous to Figure 3b.

zh-en RTV-1	許多自然的問題實際上是承諾問題。	Many natural problems are actually promise problems.
	寒冷氣候可能會帶來特殊挑戰。	Cold climates may present special challenges.
	很顯然,曾經在某個場合達成了其所不知道的某種協議。	I thought they'd come to some kind of an agreement.
	劇情發展順序與原作漫畫有些不同。	The plotline is somewhat different from the first series.
	他也創作過油畫和壁畫。	He also made sketches and paintings.
zh-en RTV-2	此節目被批評為宣揚偽科學和野史。	The book was criticized for misrepresenting nutritional science.
	威藍町體育運動場	Kawagoe Sports Park Athletics Stadium
	他是她的神聖醫師和保護者。	He's her protector and her provider.
	其後以5,000英鎊轉會到盧頓。	He later returned to Morton for £15,000.
	滬生和阿寶是小說的兩個主要人物。	Lawrence and Joanna are the play's two major characters.
zh-en RTV-3	一般上沒有會員加入到母政黨。	Voters do not register as members of political parties.
	曾任《紐約時報》書評人。	He was formerly an editor of "The New York Times Book Review" .
	48V微混系統主要由以下組件構成:	The M120 mortar system consists of the following major components:
	其後以5,000英鎊轉會到盧頓。	He later returned to Morton for £15,000.
	2月25日從香港抵達汕頭	and arrived at Hobart Town on 8 November.
zh-en RTV-4	1261年,拉丁帝國被推翻,東羅馬帝國復國。	The Byzantine Empire was fully reestablished in 1261.
	而這次航行也證明他的指責是正確的。	This proved that he was clearly innocent of the charges.
	並已經放出截面和試用版。	A cut-down version was made available for downloading.
	它重370克,由一根把和九根索組成。	It consists of 21 large gears and a 13 meters pendulum.
	派路在隊中的創造力可謂無出其右,功不可抹。	Still, the German performance was not flawless.
zh-en RTV-5	此要塞也用以鎮壓的部落。	that were used by nomads in the region.
	不過,這31次出場只有11次是首發。	In those 18 games, the visiting team won only three times.
	生於美國紐約州布魯克林。	He was born in Frewsburg, New York, USA.
	2014年7月14日,組團成為一員。	Roy joined the group on 4/18/98.
	盾上有奔走中的獅子。	Far above, the lonely hawk floating.
de-en RTV-1	Von 1988 bis 1991 lebte er in Venedig.	From 1988-1991 he lived in Venice.
	Der Film beginnt mit folgendem Zitat:	The movie begins with the following statement:
	Geschichte von Saint Vincent und den Grenadinen	History of Saint Vincent and the Grenadines
	Die Spuren des Kriegs sind noch allgegenwärtig.	Some signs of the people are still there.
	Saint-Paul (Savoie)	Saint-Paul, Savoie
de-en RTV-2	Nanderbarsche sind nicht brutpflegend.	Oxpeckers are fairly gregarious.
	Dort begegnet sie Raymond und seiner Tochter Sarah.	There she meets Sara and her husband.
	Armansperg wurde zum Premierminister ernannt.	Mansur was appointed the prime minister.
	Diese Arbeit wird von den Männchen ausgeführt.	Parental care is performed by males.
	August von Limburg-Stirum	House of Limburg-Stirum
de-en RTV-3	Es gibt mehrere Anbieter der Komponenten.	There are several components to the site.
	Doch dann werden sie von Piraten angegriffen.	They are attacked by Saracen pirates.
	Wird nicht die tiefste – also meist 6.	The shortest, probably five.
	Ihre Blüte hatte sie zwischen 1976 und 1981.	The crop trebled between 1955 and 1996.
	Er brachte Reliquien von der Hl.	Eulogies were given by the Rev.
de-en RTV-4	Gespielt wird meistens Mitte Juni.	It is played principally on weekends.
	Schuppiger Schlangenstern	Plains garter snake
	Das Artwork stammt von Dave Field.	The artwork is by Mike Egan.
	Ammonolyse ist eine der Hydrolyse analoge Reaktion,	Hydroxylation is an oxidative process.
	Die Pellenz gliedert sich wie folgt:	The Pellenz is divided as follows:
de-en RTV-5	Auch Nicolau war praktizierender Katholik.	Cassar was a practicing Roman Catholic.
	Im Jahr 2018 lag die Mitgliederzahl bei 350.	The membership in 2017 numbered around 1,000.
	Er trägt die Fahrgestellnummer TNT 102.	It carries the registration number AWK 230.
	Als Moderator war Benjamin Jaworskyj angereist.	Dmitry Nagiev appeared as the presenter.
	Benachbarte Naturräume und Landschaften sind:	Neighboring hydrographic watersheds are:

Table 8: Examples of bitext in different sections (Section 7.2). We see that tier 1 has majority parallel sentences whereas lower tiers have mostly similar but not parallel sentences.

	en-es		en-fr		en-de		en-ru		avg.
	→	←	→	←	→	←	→	←	
Nearest neighbor†	81.9	82.8	81.6	81.7	73.3	72.3	44.3	65.6	72.9
Inv. nearest neighbor (Dinu et al., 2015)†	80.6	77.6	81.3	79.0	69.8	69.7	43.7	54.1	69.5
Inv. softmax (Smith et al., 2017)†	81.7	82.7	81.7	81.7	73.5	72.3	44.4	65.5	72.9
CSLS (Lample et al., 2018)†	82.5	84.7	83.3	83.4	75.6	75.3	47.4	67.2	74.9
Artetxe et al. (2019)†	87.0	87.9	**86.0**	86.2	81.9	80.2	50.4	71.3	78.9
RTV (ours)	**89.9**	**93.5**	84.5	**89.5**	**83.0**	**88.6**	**54.5**	**80.7**	**83.0**
GEN (ours)	81.5	88.7	81.6	88.6	78.9	83.7	35.4	68.2	75.8

Table 9: P@1 of our lexicon inducer and previous methods on the standard MUSE test set (Lample et al., 2018), where the best number in each column is bolded. The first section consists of vector rotation–based methods, while Artetxe et al. (2019) conduct unsupervised machine translation and word alignment to induce bilingual lexicons. All methods are tested in the fully unsupervised setting. †: numbers copied from Artetxe et al. (2019).

src	GEN-RTV		Google Trans.
編劇	writers	<	screenwriter
可笑	laughing	<	ridiculous
極權	authoritarian	<	Totalitarian
押韻	couplets	<	rhyme
烙印	tattooed	<	brand
業主	homeowners	<	owner
安娜	grande	<	Anna
包頭	header	<	Baotou
編輯	editorial	<	edit
陣風	winds	<	gust
火柴	firewood	<	matches
盃	bowl	<	cup
武士道	samurai	<	Bushido
詩句	poem	<	verse
肚臍	belly	<	belly button
現代化	modern	<	modernization
感冒	flu	<	cold
協商	negotiate	>	Consult
納米	nanometer	>	Nano
類人猿	apes	>	Anthropoid
配件	accessories	>	Fitting
匯	aggregated	>	exchange
貸方	lenders	>	Credit
逆差	deficit	>	Trade deficit
如果	if	>	in case
附件	accessories	>	annex
實習	internship	>	practice
加冕	crowned	>	Crown
助理	assistant	>	assistant Manager
親和性	agreeableness	>	Affinity
國土	homeland	>	land
過境	crossings	✗	Transit
環流	circulation	✗	Circumfluence
羊群	sheep	✗	Herd

Table 10: all errors cases among 400 random outputs of GEN-RTV compared to both our judgement and Google translate for reference. >: GEN-RTV unacceptable while Google Trans acceptable. <: GEN-RTV acceptable while Google Trans unacceptable. ✗: both unacceptable.