Self-Augmented In-Context Learning for Unsupervised Word Translation

Anonymous ACL submission

Abstract

Recent work has shown that, while large language models (LLMs) demonstrate strong word translation or bilingual lexicon induction (BLI) capabilities in few-shot setups, they still cannot match the performance of 'traditional' mapping-based approaches in the unsupervised scenario where no seed translation pairs are available, especially for lower-resource languages. To address this challenge with LLMs, we propose **s**elf-**a**ugmented **i**n-context **l**earning (SAIL) for unsupervised BLI: starting from a zero-shot prompt, SAIL iteratively induces a set of high-confidence word translation pairs for in-context learning (ICL) from an LLM, which it then reapplies to the same LLM in the ICL fashion. Our method shows substantial gains over zero-shot prompting of LLMs on two established BLI benchmarks spanning a wide range of language pairs, also outperforming mapping-based baselines across the board. In addition to achieving state-of-the-art unsupervised BLI performance, we also conduct comprehensive analyses on SAIL and discuss its limitations.

1 Introduction and Motivation

The task of word translation (WT), also known as bilingual lexicon induction (BLI), aims to automatically induce lexica of words with the same or similar meaning in different languages, thus bridging the lexical gap between languages. Even in the era of large language models (LLMs), BLI still has wide applications in machine translation and cross-lingual transfer learning (Sun et al., 2021; Zhou et al., 2021; Wang et al., 2022; Ghazvininejad et al., 2023; Jones et al., 2023). A particular BLI setup, termed *(fully) unsupervised BLI*, is especially compelling because it is not only more technically challenging but is also used as a pivotal component towards unsupervised machine translation (Lample et al., 2018; Artetxe et al., 2018b; Marchisio et al., 2020; Chronopoulou et al., 2021).

Until recently, BLI approaches have predominantly relied on learning cross-lingual word embedding (CLWE) mappings: these are known as MAPPING-BASED approaches and are developed based on static or decontextualized word embeddings (WEs) (Patra et al., 2019; Grave et al., 2019; Li et al., 2022a; Yu et al., 2023). The trend in BLI has also recently shifted towards exploring autoregressive LLMs, which have become the cornerstone of modern NLP techniques (Brown et al., 2020; Ouyang et al., 2022; Touvron et al., 2023a) with success in many real-world tasks (Kasneci et al., 2023; Wu et al., 2023; Thirunavukarasu et al., 2023). Concerning BLI, Li et al. (2023) first show that prompting LLMs with gold-standard WT pairs as in-context examples (few-shot in-context learning: ICL) outperforms all existing BLI approaches in the supervised and semi-supervised BLI setups (where typically 1K∼5K gold-standard WT pairs are available for training or ICL), while zero-shot prompting still falls behind traditional MAPPING-BASED approaches for the fully unsupervised BLI setup, especially for lower-resource languages.

In this work, we thus aim at improving unsupervised BLI with LLMs. To this end, we propose the **s**elf-**a**ugmented **i**n-context **l**earning (SAIL) method for unsupervised BLI with LLMs. The key idea is to first iteratively retrieve a set of high-confidence WT pairs by zero-shot prompting LLMs and then use the gradually refined bilingual lexicon for BLI inference in an ICL fashion (§2). Our extensive experiments show that SAIL establishes new state-of-the-art unsupervised BLI performance on two standard BLI benchmarks. We also conduct thorough analyses of its key components, providing further insight into its inner workings (§3-§4).

2 Methodology

Unsupervised BLI: Task Preliminaries. We assume a pair of two languages: a source language

L^x with its vocabulary $\mathcal{X}$ and a target language L^y with vocabulary $\mathcal{Y}$. In a typical, standard BLI setup the vocabulary of each language contains the most frequent $200,000$ word types in the language (Glavaš et al., 2019; Li et al., 2022a). Given a source word $w^x \in \mathcal{X}$, the unsupervised BLI task then aims to infer its translation in L^y, without any word-level parallel data (i.e., seed translation pairs from a lexicon) available for training or ICL.[1]

Zero-Shot Prompting. Li et al. (2023) have proposed to prompt autoregressive LLMs for the BLI task, where the input word w^x is embedded into a predefined text template. We adopt the pool of templates provided by Li et al. (2023) and conduct template search for each LLM on a randomly chosen language pair. As an example, the zero-shot template for LLAMA-2_{7B} is as follows:[2]

'The L^x word w^x in L^y is:',

where L^x, L^y, and w^x are placeholders for the source language, target language, and the query word in the source language (e.g., L^x = Hungarian, w^x = macska, L^y = Catalan).

The deterministic beam search (with beam size of n as a hyper-parameter) is adopted to generate n output text pieces in the final beam, ranked by their sequence scores.[3] For each of the n outputs, the first word in the generated output following the input sequence is extracted as a candidate answer. After filtering out those candidate answers not in $\mathcal{Y}$, the candidate L^y word with the highest associated sequence score is returned as the final word translation prediction.

Limitations of Zero-Shot Prompting. The above zero-shot approach for unsupervised BLI, proposed by Li et al. (2023), comes with several limitations. First, the template does not stipulate the output format and thus parsing the output text may not be as straightforward as expected. Put simply, LLM's prediction may not be the first word in the generated sequence. Second, the LLM may not fully 'understand' the input template and sometimes may tend not to generate words for lower-resource languages. For the *supervised* BLI setup, where a dictionary of gold standard translation pairs is assumed and available, few-shot in-context learning can substantially improve final BLI performance (Li et al., 2023), since it not only provides examples of the desired output format but also helps LLMs 'understand' the BLI task. However, the availability of such a seed dictionary is not assumed in the *unsupervised* BLI task variant, and the key idea of this work is to derive and iteratively refine a seed dictionary by prompting LLMs.

SAIL: Self-Augmented In-Context Learning for Unsupervised BLI. We thus propose to facilitate and improve unsupervised BLI by **S1)** using zero-shot prompting to retrieve $\mathcal{D}_h$, a set of high-confidence translation pairs, and then **S2)** leveraging these pairs as 'self-augmented' in-context examples for few-shot prompting to further iteratively refine $\mathcal{D}_h$ (across 0 to $N_{it}-1$ iterations, where N_{it} is a hyper-parameter denoting total times of $\mathcal{D}_h$ inference in S1 and S2), and finally **S3)** conducting few-shot learning with the final, N_{it}-th self-created seed lexicon $\mathcal{D}_h$ for BLI inference on the test set.

Deriving High-Confidence Pairs. For both steps S1 and S2 outlined above, we start with the most frequent N_f words in L^x since representations of less frequent words are considered to be much noisier in general (Artetxe et al., 2018a). For each w^x, we conduct $L^x \rightarrow L^y$ translation: we refer to this predicted word as $\hat{w}^y$. We then propose to conduct *word back-translation*, translating $\hat{w}^y$ from L^y back into L^x. The word pair $(w^x, \hat{w}^y)$ is considered a high-confidence pair only if w^x is also the output word of the back-translation step.[4] We denote the set of all high-confidence pairs from the L^x words as $\mathcal{D}_h^x$. Likewise, we also start from the most frequent N_f words in L^y and symmetrically derive $\mathcal{D}_h^y$. Finally, we update the high-confidence dictionary with $\mathcal{D}_h = \mathcal{D}_h^x \cup \mathcal{D}_h^y$.

Few-Shot Prompting with High-Confidence Pairs. Step S1 of SAIL relies on zero-shot prompting, but all the subsequent iterations in S2 and S3 apply few-shot prompting/ICL with the 'self-augmented' high-confidence translation pairs $\mathcal{D}_h$. Following Li et al. (2023), we adopt 5-shot prompting, and again conduct template search on the BLI task with a single, randomly selected language pair.[5] The in-context examples, $(w_i^x, w_i^y) \in$

[1] Again following prior work, when w^x has multiple ground truth translations in L^y, a prediction is considered correct if it is any of the ground truth answers.

[2] The full list of templates used for other LLMs are presented in Table in Appendix .

[3] We use $n = 5$ following Li et al. (2023).

[4] Earlier MAPPING-BASED approaches have retrieved high-confidence pairs through ranking cross-lingual word similarity scores (e.g., cosine similarity) to refine CLWE mappings (Artetxe et al., 2018a; Li et al., 2022a); in a sense, our work renovates and revitalises the idea with LLMs.

[5] The decoding and output parsing strategy is the same as in zero-shot prompting.

$\mathcal{D}_h, 1 \leq i \leq 5$, are retrieved where the w_i^x words are the nearest neighbours of the input word w^x in L^x's static word embedding space. The few-shot template for LLAMA-2_{7B} is then as follows:

> 'The L^x word w_1^x in L^y is w_1^y. The L^x word w_2^x in L^y is w_2^y. ... The L^x word w^x in L^y is'.

3 Experimental Setup

BLI Data and LLMs. We adopt two standard BLI benchmarks: **1)** 5 languages from XLING (Glavaš et al., 2019) including German (DE), English (EN), French (FR), Italian (IT), and Russian (RU); their combinations result in 20 BLI directions; **2)** 3 lower-resource languages including Bulgarian (BG), Catalan (CA), and Hungarian (HU) from PanLex-BLI (Vulić et al., 2019), which result in 6 BLI directions.[6] For both benchmarks, a test set of 2K WT pairs is provided for each BLI direction. We experiment with four open-source LLMs: LLAMA $_{7B}$, LLAMA-2_{7B}, LLAMA $_{13B}$, and LLAMA-2_{13B} (Touvron et al., 2023a,b). Li et al. (2023) found that 4 other families of LLMs, including mT5, mT0, mGPT and XGLM, underperform LLAMA; we thus skip these LLMs in our work.

Implementation Details and BLI Evaluation. As mentioned in §2, our hyper-parameter and template search are conducted on a single, randomly selected language pair, which is DE-FR, following Li et al. (2023). Batch size is set to 1. We adopt $N_{it} = 1$, $N_f = 5,000$ in our main experiments (§4.1) and then investigate their influence on BLI performance and the effectiveness of our proposed word back-translation in our further analyses (§4.2). Half-precision floating-point format (`torch.float16`) is adopted for all our SAIL and ZERO-SHOT experiments. Since our method does *not* imply any randomness, all results are from single runs. For evaluation, we adopt the standard *top-1 accuracy* as prior work.

Baselines. We adopt two established MAPPING-BASED baselines. **1)** VECMAP is a representative unsupervised BLI approach and features a self-learning mechanism that refines linear maps for deriving CLWEs (Artetxe et al., 2018a). **2)** CONTRASTIVEBLI learns CLWEs with a two-stage contrastive learning framework and is the strongest MAPPING-BASED approach for supervised and semi-supervised BLI tasks on our two benchmarks (Li et al., 2022a); however, it does not support unsupervised setup. We extend CONTRASTIVEBLI to unsupervised BLI by initialising the initial map with the unsupervised VECMAP method. The CONTRASTIVEBLI C1 variant based on static WEs and its stronger C2 variant combining static and decontextualised WEs are both used as our baselines. In addition, we report **3)** ZERO-SHOT prompting with each of our LLMs as baselines following Li et al. (2023).

[6] The two datasets are also used in many recent BLI works (Sachidananda et al., 2021; Aboagye et al., 2022; Li et al., 2022a,b; Vulić et al., 2020, 2023; Li et al., 2023).

[Unsupervised BLI]	DE	EN	FR	IT	RU	AVG.
	MAPPING-BASED					
VECMAP	44.14	51.7	51.51	51.03	34.36	46.55
CONTRASTIVEBLI (C1)	44.72	52.12	52.29	51.77	35.5	47.28
CONTRASTIVEBLI (C2)	46.02	53.32	53.26	52.99	37.26	48.57
	ZERO-SHOT					
LLAMA $_{7B}$	41.94	50.16	48.25	46.91	40.04	45.46
LLAMA-2_{7B}	43.91	52.7	50.68	48.23	42.8	47.66
LLAMA $_{13B}$	45.39	53.35	52.39	50.58	41.74	48.69
LLAMA-2_{13B}	47.12	55.02	51.31	52.02	43.09	49.71
	SAIL (Ours)					
LLAMA $_{7B}$	51.39	61.92	58.92	56.94	50.7	55.97
LLAMA-2_{7B}	53.81	64.12	61.09	59.96	53.77	58.55
LLAMA $_{13B}$	55.35	64.84	62.49	61.27	54.5	59.69
LLAMA-2_{13B}	**57.69**	**67.0**	**64.11**	**63.18**	**57.04**	**61.8**

Table 1: Main results on the 20 XLING BLI directions. For each language, the average accuracy scores over 8 BLI directions (i.e., going from and going to other 4 languages) is reported. See also Appendix E.

4 Results and Discussion

4.1 Main Results

Results on the Two BLI Benchmarks are summarised in Tables 1 and 2 respectively, with full BLI scores pear each individual language pair in Tables 6 and 7 in Appendix E. As the main findings, **1)** our SAIL shows consistent gains against ZERO-SHOT for each of the 4 LLMs, showing the effectiveness of the proposed approach; **2)** while ZERO-SHOT still lags behind MAPPING-BASED on PanLex-BLI's lower-resource languages, applying SAIL outperforms MAPPING-BASED across the board. The only exception is that CONTRASTIVEBLI (C2) still has a slight edge over SAIL with the weakest LLM overall, LLAMA $_{7B}$. **3)** Among the 4 LLMs, LLAMA-2_{13B} presents the strongest BLI capability.

Variance and Statistical Significance. The whole SAIL method does *not* imply any variance due to randomness: it does not rely on any actual LLM fine-tuning; we adopt deterministic beam search; the deterministic nearest neighbour retrieval is used for deriving IC examples. Here, we report the sta-

[Unsupervised BLI]	BG	CA	HU	AVG.
	MAPPING-BASED			
VECMAP	37.22	36.27	36.89	36.8
CONTRASTIVEBLI (C1)	36.7	35.86	37.82	36.79
CONTRASTIVEBLI (C2)	38.87	38.48	40.54	39.3
	ZERO-SHOT			
LLAMA $_{7B}$	27.9	28.87	27.18	27.98
LLAMA-2$_{7B}$	28.2	27.21	26.92	27.45
LLAMA $_{13B}$	27.49	30.61	28.2	28.77
LLAMA-2$_{13B}$	29.08	32.38	30.53	30.66
	SAIL (Ours)			
LLAMA $_{7B}$	37.02	37.63	36.29	36.98
LLAMA-2$_{7B}$	40.06	40.51	40.22	40.27
LLAMA $_{13B}$	41.71	42.76	42.07	42.18
LLAMA-2$_{13B}$	**45.4**	**46.26**	**44.88**	**45.51**

Table 2: Main results on 6 PanLex-BLI BLI directions. For each language, the average accuracy scores over 4 BLI directions (i.e., going from and going to other 2 languages) is reported. See also Appendix E.

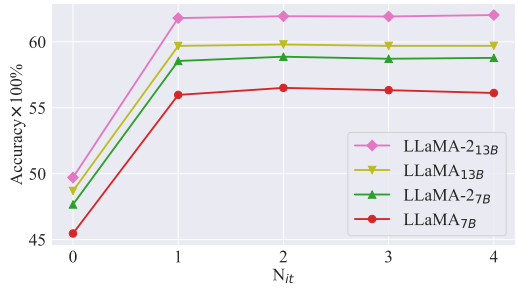

Figure 1: Top-1 accuracy ($\times 100\%$) averaged over 20 XLING BLI directions with respect to N_{it}. Setting $N_{it} = 0$ refers to the ZERO-SHOT baseline.

tistical significance with χ^2 tests. When comparing SAIL and ZERO-SHOT (both with LLAMA-2$_{13B}$), the p-value is $1.1e$-251 on 20 XLING BLI directions and $2.7e$-109 on 6 PanLex-BLI BLI directions. We then compare SAIL (with LLAMA-2$_{13B}$) against CONTRASTIVEBLI (C2) which is the strongest mapping-based baseline: the p-values are $3.1e$-300 and $7.8e$-20 respectively. These show that our findings are strongly statistically significant.

4.2 Further Analyses

Impact of N_{it}. Figure 1 shows the influence of the number of iterations N_{it} on the average BLI scores on XLING. When $N_{it} = 1$, where only step S1 is executed (see §2), SAIL already approaches (almost) its optimal performance. Further refining the $\mathcal{D}_h$ for more iterations (step S2) only leads to small fluctuations in BLI performance, which we deem not worth the increased computational cost. Figure 3 (Appendix B) with results on PanLex-BLI shows a similar trend.

Impact of N_f. We then study the impact of the frequency threshold N_f on the average BLI performance with a subset of XLING spanning DE-FR, EN-RU and RU-FR, each in both directions. The results in Figure 2 reveal that even with $N_f = 1,000$, the BLI performance is boosted substantially when compared against the ZERO-SHOT baseline (i.e., when $N_f = 0$). When we further increase N_f, the accuracy score still increases slowly, and the gain seems negligible with $N_f \geq 5000$: i.e., increasing N_f again may not be worth the extra computational cost.

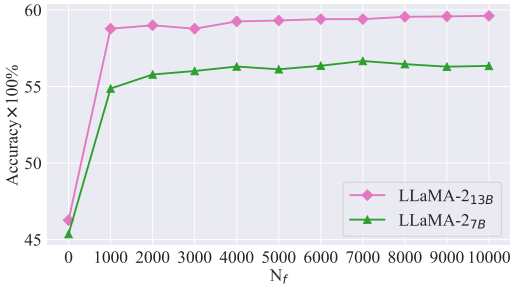

Figure 2: Top-1 accuracy on XLING with respect to N_f. $N_f = 0$ yields the ZERO-SHOT baseline.

	ZERO-SHOT	SAIL (w/o back translation)	SAIL
LLAMA-2$_{7B}$	45.36	52.9	**56.12**
LLAMA-2$_{13B}$	46.26	55.1	**59.31**

Table 3: BLI results on XLING, demonstrating the usefulness of back-translation when constructing $\mathcal{D}_h$.

Impact of Word Back-Translation. The back-translation step aims to improve the quality of $\mathcal{D}$. Here, we experiment with the ablated version of SAIL without back translation on the same XLING subset as before. The results in Table 3 clearly demonstrate the effectiveness of proposed word back-translation: the p-values (χ^2 tests) are $8.8e$-7 and $1.0e$-10 respectively for LLAMA-2$_{7B}$ and LLAMA-2$_{13B}$ when comparing SAIL variants with and without the back-translation component.

5 Conclusion

We proposed Self-Augmented In-Context Learning (SAIL) to improve unsupervised BLI with LLMs. The key idea is to iteratively retrieve a set of high-confidence word translation pairs by prompting LLMs and then leverage the retrieved pairs as in-context examples for unsupervised BLI. Our experiments on two standard BLI benchmarks showed that the proposed SAIL method substantially outperforms established MAPPING-BASED and ZERO-SHOT BLI baselines. The code will be available at [ANONYMOUS-URL].

Limitations

The main limitation of this work, inherited from prior work as well (Li et al., 2023) is that the scope of our languages is constrained to the languages supported (or 'seen') by the underlying LLMs. For example, LLAMA-2 is reported to support only around 27 natural languages (Touvron et al., 2023b). This limitation could be mitigated if more advanced LLMs that support more languages are available in the future. It might also be feasible to adapt existing LLMs to more languages by fine-tuning on their monolingual corpora potentially combined with modern cross-lingual transfer learning techniques, whereas such adaptations of LLMs to unseen languages extend way beyond this work focused on the BLI task.

In addition, compared to the ZERO-SHOT baseline, our SAIL framework organically requires more computational time and budget, as reported in Table 5 of Appendix D.

Moreover, the SAIL framework is proposed and evaluated for the unsupervised BLI task. This work does not discuss if and how adapted variants of SAIL could also be applied to other NLP tasks beyond BLI. Further, the SAIL method should be equally applicable in weakly supervised BLI setups (Vulić et al., 2019) where a tiny set of available seed word translations (e.g., 50-500 word pairs) can be assumed to seed the iterative procedure. We leave this to future work.

References

Prince Osei Aboagye, Jeff Phillips, Yan Zheng, Junpeng Wang, Chin-Chia Michael Yeh, Wei Zhang, Liang Wang, and Hao Yang. 2022. Normalization of language embeddings for cross-lingual alignment. In *International Conference on Learning Representations*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. 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. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018b. Unsupervised neural machine translation. In *International Conference on Learning Representations*.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Alexandra Chronopoulou, Dario Stojanovski, and Alexander Fraser. 2021. Improving the lexical ability of pretrained language models for unsupervised neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 173–180, Online. Association for Computational Linguistics.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based phrase-level prompting of large language models for machine translation. *arXiv preprint arXiv:2302.07856*.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. 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. Association for Computational Linguistics.

Edouard Grave, Armand Joulin, and Quentin Berthet. 2019. Unsupervised alignment of embeddings with wasserstein procrustes. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1880–1890. PMLR.

Alex Jones, Isaac Caswell, Ishank Saxena, and Orhan Firat. 2023. Bilex rx: Lexical data augmentation for massively multilingual machine translation. *arXiv preprint arXiv:2303.15265*.

Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proceedings of the International Conference on Learning Representations*.

Yaoyiran Li, Anna Korhonen, and Ivan Vulić. 2023. On bilingual lexicon induction with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9577–9599, Singapore. Association for Computational Linguistics.

Yaoyiran Li, Fangyu Liu, Nigel Collier, Anna Korhonen, and Ivan Vulić. 2022a. Improving word translation via two-stage contrastive learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4353–4374, Dublin, Ireland. Association for Computational Linguistics.

Yaoyiran Li, Fangyu Liu, Ivan Vulić, and Anna Korhonen. 2022b. Improving bilingual lexicon induction with cross-encoder reranking. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4100–4116, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sasha Luccioni, Victor Schmidt, Alexandre Lacoste, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. In *NeurIPS 2019 Workshop on Tackling Climate Change with Machine Learning*.

Kelly Marchisio, Kevin Duh, and Philipp Koehn. 2020. When does unsupervised machine translation work? In *Proceedings of the Fifth Conference on Machine Translation*, pages 571–583, Online. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R. Gormley, and Graham Neubig. 2019. Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 184–193, Florence, Italy. Association for Computational Linguistics.

Vin Sachidananda, Ziyi Yang, and Chenguang Zhu. 2021. Filtered inner product projection for crosslingual embedding alignment. In *International Conference on Learning Representations*.

Jimin Sun, Hwijeen Ahn, Chan Young Park, Yulia Tsvetkov, and David R. Mortensen. 2021. Cross-cultural similarity features for cross-lingual transfer learning of pragmatically motivated tasks. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2403–2414, Online. Association for Computational Linguistics.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ivan Vulić, Goran Glavaš, Fangyu Liu, Nigel Collier, Edoardo Maria Ponti, and Anna Korhonen. 2023. Probing cross-lingual lexical knowledge from multilingual sentence encoders. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2089–2105, Dubrovnik, Croatia. Association for Computational Linguistics.

Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Do we really need fully unsupervised cross-lingual embeddings? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4407–4418, Hong Kong, China. Association for Computational Linguistics.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Xinyi Wang, Sebastian Ruder, and Graham Neubig. 2022. Expanding pretrained models to thousands more languages via lexicon-based adaptation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 863–877, Dublin, Ireland. Association for Computational Linguistics.

Zongxiao Wu, Yizhe Dong, Yaoyiran Li, and Baofeng Shi. 2023. Unleashing the power of text for credit

default prediction: Comparing human-generated and ai-generated texts. *Available at SSRN 4601317*.

Shenglong Yu, Wenya Guo, Ying Zhang, and Xiaojie Yuan. 2023. Cd-bli: Confidence-based dual refinement for unsupervised bilingual lexicon induction. In *Natural Language Processing and Chinese Computing*, pages 379–391, Cham. Springer Nature Switzerland.

Yucheng Zhou, Xiubo Geng, Tao Shen, Wenqiang Zhang, and Daxin Jiang. 2021. Improving zero-shot cross-lingual transfer for multilingual question answering over knowledge graph. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5822–5834, Online. Association for Computational Linguistics.

A Languages

Family	Language	Code
Germanic	English	EN
	German	DE
Romance	Catalan	CA
	French	FR
	Italian	IT
Slavic	Bulgarian	BG
	Russian	RU
Uralic	Hungarian	HU

Table 4: Languages used in our experiments with their ISO 639-1 codes.

B Impact of N_{it} with PanLex-BLI

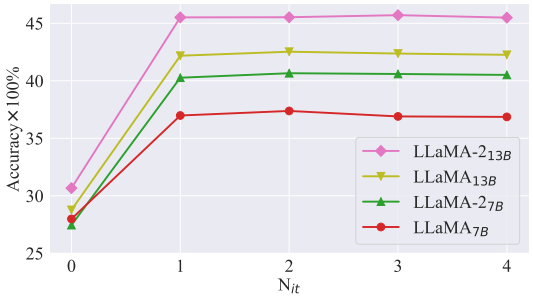

Figure 3: Top-1 accuracy ($\times 100\%$) averaged over 6 PanLex-BLI BLI directions with respect to N_{it}. Setting $N_{it} = 0$ refers to the ZERO-SHOT baseline.

C Templates

Li et al. (2023) provide the suggested (carefully searched) templates for LLAMA $_{7B}$ and LLAMA $_{13B}$, which we directly adopt in our work. For LLAMA-2_{7B} and LLAMA-2_{13B}, we conduct template search following Li et al. (2023) on a single language pair DE-FR in both directions.

Zero-Shot Template. LLAMA $_{7B}$, LLAMA-2_{7B} and LLAMA-2_{13B} share the same zero-shot template as introduced in §2. LLAMA-2_{13B}'s zero-shot template is as follows:

'Translate from L^x to L^y: w^x=>'.

Few-Shot Template.. We have introduced the few-shot template of LLAMA-2_{7B} in §2. The remaining three LLMs happen to share the same few-shot template, given as follows:

'The L^x word 'w_1^x' in L^y is w_1^y. The L^x word 'w_2^x' in L^y is w_2^y. ... The L^x word 'w^x' in L^y is'.

D Reproducibility Checklist

• **Source Code**: our code will be made publicly available at [ANONYMOUS-URL].

• **Hyperparameter Search**: N_{it} is selected from $\{1, 2, 3, 4\}$ and N_f from $\{1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000\}$.

• **Software**: Python 3.9.7, PyTorch 1.10.1, Transformers 4.28.1.

• **Computing Infrastructure**: our code is run with a single Nvidia 80GB A100 GPU.

• **Half-Precision Floating-Point Format**: as introduced in §3, our BLI inference relies on `torch.float16` for both our SAIL and the ZERO-SHOT baseline. We have verified that `fp16` can accelerate our computation with only negligible impact on the absolute BLI performance. Note that Li et al. (2023) did not specify `torch.float16` in their ZERO-SHOT experiments with LLAMA $_{7B}$ and LLAMA $_{13B}$, so the BLI scores reported are slightly different from ours.

• **Data, WEs, LLMs**: all the BLI data, WEs, LLMs and baseline codes are open-source and publicly available. The WEs for retrieving in-context examples are fastText WEs (Bojanowski et al., 2017): the version pretrained on Wikipedia[7] is used for XLING and the version pretrained with Wikipedia plus Common Crawl[8] is used for PanLex-BLI, as recommended by XLING and PanLex-BLI, respectively. The same WEs are used for our MAPPING-BASED baselines.

• **Baselines**: for every baseline, we use its recommended setup for unsupervised BLI and make sure the recommended setup achieves its own (near-)optimal performance. As introduced in §3, we extend CONTRASTIVEBLI to the unsupervised BLI setup. Specifically, we adopt the set of its hyperparameters recommended for weakly supervised BLI setup, which we found can also achieve strong unsupervised BLI performance.

• **Parameter Count and Runtime**: we report the number of parameters of each LLM and the GPU

[7] https://fasttext.cc/docs/en/pretrained-vectors.html

[8] https://fasttext.cc/docs/en/crawl-vectors.html

runtime for BLI inference on a single BLI direction DE→FR, which contains circa 2K word pairs, in Table 5.

• **Carbon Footprint**: our work consumes about 750 A100 GPU hours in total which we estimate causes the emission of 90kg CO_2 equivalents according to a publicly available 'machine learning emissions calculator' (Luccioni et al., 2019)[9].

E Full BLI Results

Table 6 shows detailed BLI scores for each BLI direction in the XLING dataset. Similarly, individual per-direction results on PanLex-BLI are presented in Table 7.

[9] https://mlco2.github.io/impact/#compute

LLM	Model ID	Parameter Count	Runtime: ZERO-SHOT	Runtime: SAIL
LLAMA $_{7B}$	"huggyllama/llama-7b"	$6,738,415,616$	5 min	40 min
LLAMA-2_{7B}	"meta-llama/Llama-2-7b-hf"	$6,738,415,616$	5 min	40 min
LLAMA $_{13B}$	"huggyllama/llama-13b"	$13,015,864,320$	6 min	49 min
LLAMA-2_{13B}	"meta-llama/Llama-2-13b-hf"	$13,015,864,320$	6 min	49 min

Table 5: LLMs adopted in our work with their huggingface.co model IDs, parameter count, and GPU runtime on a single BLI direction for ZERO-SHOT and SAIL respectively.

[Unsupervised BLI]	VECMAP	CONTRASTIVEBLI (C1)	CONTRASTIVEBLI (C2)	LLAMA $_{7B}$	LLAMA-2_{7B}	LLAMA $_{13B}$	LLAMA-2_{13B}	LLAMA $_{7B}$	LLAMA-2_{7B}	LLAMA $_{13B}$	LLAMA-2_{13B}
	MAPPING-BASED			ZERO-SHOT				SAIL (Ours)			
DE→FR	48.98	50.39	51.8	42.46	44.44	47.37	46.64	54.67	54.77	58.37	**61.5**
FR→DE	43.97	43.61	44.9	43.2	45.47	48.11	50.8	50.08	54.16	54.47	**56.29**
DE→IT	48.41	49.77	50.23	42.78	42.78	46.06	48.51	53.36	54.25	57.38	**59.05**
IT→DE	44.03	43.93	45.43	38.6	41.55	44.39	45.27	46.15	51.63	52.2	**52.92**
DE→RU	25.67	28.22	31.09	30.41	35.32	32.76	36.62	45.12	46.9	48.98	**51.59**
RU→DE	39.13	40.02	41.33	43.53	44.68	43.11	42.12	46.83	50.55	50.65	**53.9**
EN→DE	48.4	47.45	47.4	52.0	52.1	54.35	59.85	59.55	61.75	62.8	**65.05**
DE→EN	54.51	54.36	55.97	42.57	44.91	46.95	47.16	55.35	56.44	57.96	**61.24**
EN→FR	60.15	61.05	61.25	57.6	62.65	62.65	61.75	72.6	73.8	75.85	**76.35**
FR→EN	61.25	62.34	63.58	54.58	55.56	57.27	53.03	63.68	65.13	65.29	**66.63**
EN→IT	57.4	57.6	58.75	58.95	60.85	60.4	65.8	71.7	73.0	74.25	**77.6**
IT→EN	60.83	62.02	63.46	47.39	50.08	54.94	53.54	60.1	64.08	64.13	**65.43**
EN→RU	24.55	25.45	26.1	42.05	44.6	40.1	47.6	57.4	60.25	61.05	**63.75**
RU→EN	46.52	46.67	50.03	46.15	50.81	50.13	51.44	54.95	58.51	57.41	**59.93**
IT→FR	64.75	65.12	65.89	51.42	54.47	57.36	55.3	61.91	65.58	65.94	**68.17**
FR→IT	63.37	63.94	64.61	57.32	55.98	60.01	61.87	64.72	66.22	69.22	**69.53**
RU→FR	45.31	46.78	47.93	43.58	48.04	47.77	41.17	54.79	57.62	57.52	**60.29**
FR→RU	24.26	25.09	26.07	35.8	38.8	38.59	39.94	48.94	51.42	53.29	**54.11**
RU→IT	43.95	44.89	46.15	47.3	47.15	45.99	49.45	53.54	56.26	56.31	**59.25**
IT→RU	25.48	26.87	29.35	31.52	33.02	35.45	36.38	44.03	48.63	50.75	**53.49**
Avg.	46.55	47.28	48.57	45.46	47.66	48.69	49.71	55.97	58.55	59.69	**61.8**

Table 6: Full BLI results on 20 XLING BLI directions.

[Unsupervised BLI]	VECMAP	CONTRASTIVEBLI (C1)	CONTRASTIVEBLI (C2)	LLAMA $_{7B}$	LLAMA-2_{7B}	LLAMA $_{13B}$	LLAMA-2_{13B}	LLAMA $_{7B}$	LLAMA-2_{7B}	LLAMA $_{13B}$	LLAMA-2_{13B}
	MAPPING-BASED			ZERO-SHOT				SAIL (Ours)			
BG→CA	39.6	38.08	39.66	32.83	29.79	32.77	33.47	40.19	42.23	42.52	**47.9**
CA→HU	34.09	34.2	36.85	23.7	23.2	24.42	30.17	32.27	35.25	38.34	**39.83**
HU→BG	36.46	38.36	40.44	28.28	27.71	26.5	26.73	38.19	41.47	43.89	**46.66**
CA→BG	33.6	31.39	33.94	26.35	27.2	27.03	28.39	36.54	38.47	42.27	**45.67**
HU→CA	37.79	39.77	43.45	32.62	28.66	38.23	37.51	41.53	46.09	47.91	**51.65**
BG→HU	39.24	38.95	**41.44**	24.13	28.12	23.67	27.72	33.16	38.08	38.14	41.38
Avg.	36.8	36.79	39.3	27.98	27.45	28.77	30.66	36.98	40.27	42.18	**45.51**

Table 7: Full BLI results on 6 PanLex-BLI BLI directions.