Williams College's Submission for the Coco4MT 2023 Shared Task

Alex Root asr6@williams.edu
Mark Hopkins mh24@williams.edu
Department of Computer Science, Williams College, Williamstown, MA, 01267

Abstract

Professional translation is expensive. As a consequence, when developing a translation system in the absence of a pre-existing parallel corpus, it is important to strategically choose sentences to have professionally translated for the training corpus. In our contribution to the Coco4MT 2023 Shared Task, we explore how sentence embeddings can be leveraged to choose an impactful set of sentences to translate. Based on six language pairs of the JHU Bible corpus, we demonstrate that a technique based on SimCSE embeddings outperforms a competitive suite of baselines.

1 Introduction

It has become increasingly possible to train decent translation models with small, high-quality parallel corpora. For instance, Maillard et al. (2023) recently showed that just 6000 professionally-translated sentences made a big impact on the quality of trained translation models for 39 low-resource languages.

This raises a research question. Suppose we want to develop a model that translates between a high-resource language and a low-resource language (or possibly a specialized domain of a moderately-resourced language). How should one select a "seed" dataset to have professionally translated from the high-resource language into the low-resource language? This is the subject of the Coco4MT 2023 Shared Task.

For our contribution to this shared task, we focus on *model-agnostic* approaches. In contrast to approaches that leverage model uncertainty to select sentences to professionally translate (Bhatnagar et al., 2022), model-agnostic approaches (Zhao et al., 2020) select a training corpus based exclusively on the distribution and content of sentences in the high-resource language. While ignoring model uncertainty may result in lower-quality data selection for a particular model, it has the potential advantage of broader applicability, as the data selection is not tied to a specific model architecture.

2 Task Description

Define a *translation model* t as a function that maps source-language documents to target-language documents. Each translation pair $(x, t(x))$ (where x is a source-language document) is assumed to have a non-negative real-valued quality $q(x, t(x))$. Given a distribution $P_{\mathcal{X}}$ over source-language documents, the quality of translation model t is the expected translation quality:

$$q(t) = E_{P_{\mathcal{X}}}[q(x, t(x))]$$

A *parallel corpus* is a set of pairs (x, y), where y is a target-language translation of source-language document x. A trainer τ takes a parallel corpus as its input, and outputs (possibly nondeterministically) a translation model t.

The Coco4MT 2023 Shared Task is about optimizing parallel corpus construction for training translation models. We have access to a monolingual corpus of documents $X = \{x_1, ..., x_n\}$ in a high-resource language, assumed to be drawn i.i.d. from document distribution $P_{\mathcal{X}}$. We also have access to a cost function $c : X \to \mathbb{R}^+$ that maps each document $x_i \in X$ to the positive real cost $c(x_i)$ of obtaining a professional translation for document x_i. For monolingual corpus X and a subset $I \subseteq \{1, ..., n\}$ of selected document ids, let $Z_{X,I} = \{(x_i, y_i) \mid i \in I\}$ be the parallel corpus we would obtain (i.e. y_i is the professional translation of document x_i) from commissioning translations for the documents associated with the selected ids. The cost of building parallel corpus $Z_{X,I}$ is therefore:

$$c(Z_{X,I}) = \sum_{i \in I} c(x_i)$$

The goal of the task is to construct a parallel corpus $Z_{X,I}$ that produces a translation model of maximal expected quality, subject to the constraint that the construction cost $c(Z_{X,I})$ is less than a specified budget B. In other words, we want to compute:

$$\hat{I} = \underset{\substack{I \subseteq \{1,...,n\}: \\ c(Z_{X,I}) \leq B}}{\text{argmax}} \ E[q(\tau(Z_{X,I}))]$$

where $E[q(\tau(Z_{X,I}))]$ is the expected quality of the translation model produced by trainer τ, when trained on parallel corpus $Z_{X,I}$.

The official Coco4MT 2023 Shared Task uses a uniform cost function, i.e. $c(x_i) = 1$ for all $x_i \in X$. In other words, all documents have the same translation cost. The budget $B = 0.2 * |X|$ is 20% of the documents in monolingual corpus X. We will also experiment with a token-based translation cost, i.e. where $c(x_i)$ equals the number of tokens in document x_i[1]. The corresponding budget B will be 20% of the tokens in corpus X.

3 Baselines

3.1 Simple Baselines

We experimented with three simple baselines:

- **longest**: choose ids corresponding to the longest documents in corpus X, until the budget is exhausted.
- **random**: choose ids uniformly at random, until the budget is exhausted.
- **weighted random**: choose ids from a categorical distribution where each document is weighted by token length, until the budget is exhausted.

3.2 Decay Logarithm Frequency (delfy)

As an additional baseline, we also implemented an algorithm from Zhao et al. (2020) called **de**cay **l**ogarithm **f**requenc**y** (**delfy**). The **delfy** algorithm attempts to choose a subset of documents that are *representative* of distribution $P_{\mathcal{X}}$ without being overly *redundant* (the intuition is that it is wasteful to commission translations of similar documents, even if these documents

[1] Since English is always the source language (and since punctuation arguably doesn't contribute to the translation cost of a sentence), we simply use whitespace-based tokenization in our experiments.

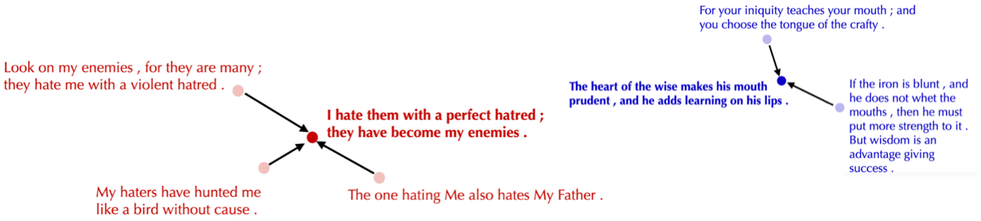
Look on my enemies , for they are many ; they hate me with a violent hatred .
I hate them with a perfect hatred ; they have become my enemies .
My haters have hunted me like a bird without cause .
The one hating Me also hates My Father .
For your iniquity teaches your mouth ; and you choose the tongue of the crafty .
The heart of the wise makes his mouth prudent , and he adds learning on his lips .
If the iron is blunt , and he does not whet the mouths , then he must put more strength to it . But wisdom is an advantage giving success .

Figure 1: A visualization of our document selection approach. We embed each document using SimCSE (Gao et al., 2021) and then choose the most central documents to include in our training set.

are in a high-probability region of the document space). The **delfy** algorithm consists of K successive rounds of document selection, such that each round exhausts $\frac{1}{K}$ of the budget. During each round, the unselected documents are ranked according a formula that balances frequency in the overall corpus with infrequency in the documents selected during previous rounds. Then the top-ranked documents are selected. Following Zhao et al. (2020), we run $K = 20$ selection rounds. We refer the reader to Zhao et al. (2020) for details of the ranking formula.

4 Our Approach

Inspired by the **delfy** algorithm, we investigated other ways to balance representativeness with redundancy. Specifically, we focused on using document embeddings to select the documents to translate. Document embeddings seek to cluster documents based on lexical and semantic similarity. For instance, if we apply the popular document embedding model SimCSE (Gao et al., 2021) to the JHU Bible Corpus (McCarthy et al., 2020), we obtain clusters like the ones shown in Figure 1. For instance, the most similar sentence to *My haters have hunted me like a bird without cause* is *I hate them with a perfect hatred; they have become my enemies.*

The intuition behind our approach was to balance representativeness and redundancy by choosing representatives from each cluster of embeddings. We embedded each document using SimCSE (Gao et al., 2021), then determined the nearest neighbor (based on cosine similarity) of each document. In other words, we compiled the following set of document pairs:

$$\mathsf{nearest}(X) = \left\{ \left(x_i, \operatorname*{argmin}_{x_j \in X \setminus \{x_i\}} \mathsf{sim}(x_i, x_j) \right) \right\}$$

where $X = \{x_1, ...x_n\}$ is the high-resource language corpus (as defined in Section 2), and sim is the cosine similarity function. Then we ranked the high-resource documents based on how many times they were the nearest neighbor of another document:

$$\mathsf{centrality}(x_i) = \min(2, |\{(x_j, x_i) \in \mathsf{nearest}(X)\}|)$$

Based on this ranking, we selected documents from the high-resource corpus until the budget was exhausted. When using a uniform translation cost (i.e. when all documents had the same translation cost), ties were broken based on sentence length[2]. When using a token-based translation cost, ties were broken randomly.

We experimented with alternative approaches to selecting cluster representatives (e.g. centroids of k-nearest neighbor clustering), but these all underperformed the simple method described above.

[2] When two sentences had the same centrality, the longer sentence was preferred.

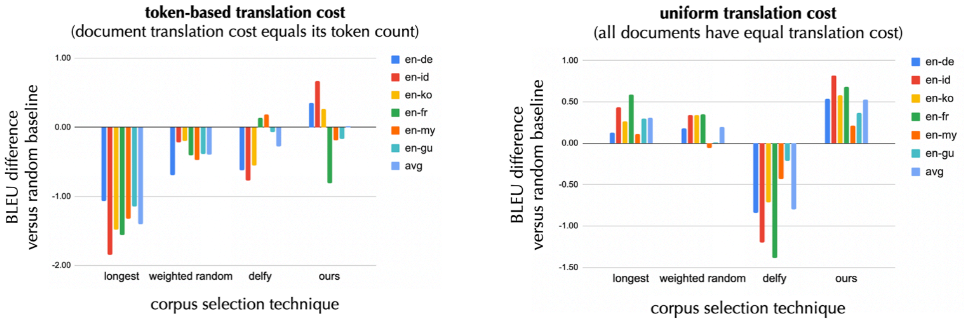

Figure 2: Results for token-based (left) and uniform (right) translation costs. For each language pair, we show the difference between the BLEU score of a model trained on documents selected by a given technique, versus a corresponding model trained on a randomly selected corpus. For each experiment, the budget is 20% of the high-resource training corpus (counted in terms of tokens and documents, respectively).

5 Experiments

We used the experimental setup provided by the Coco4LM 2023 Shared Task. The training corpus consisted of roughly 22k English sentences from the JHU Bible Corpus (McCarthy et al., 2020), along with translations in six other languages: German (de), Indonesian (id), Korean (ko), French (fr), Gujarati (gu), Burmese (my). Additionally, there was a development corpus of 3919 sentences and a test corpus of 8708 sentences (all from the biblical domain). Since none of the evaluated techniques used the non-English data for training[3], we evaluated on all six en-X language pairs (en-fr, en-gu, en-my, en-de, en-id, en-ko). For each corpus selection technique, we fine-tuned the mBART-50 model (Liu et al., 2020) on the documents[4] identified by that strategy and evaluated using BLEU (Papineni et al., 2002). We used the implementation and default parameters of mBART-50 provided by HuggingFace.

Results are shown in Figure 2. Our sentence embedding approach performed consistently across the two translation costs and six language pairs (with one exception: en-fr translation for the token-based translation cost). Surprisingly, the **delfy** algorithm consistently underperformed the random baseline, despite successes in other studies (Zhao et al., 2020). Perhaps this was due to the narrow domain (biblical) or the limited size of the training corpus.

6 Conclusion, Limitations, and Future Work

Sentence embeddings show potential as an instrument for selecting a good proxy dataset for a translation domain. However, it is important to acknowledge the limitations of this pilot study.

Domain Specificity: All conclusions were drawn on the basis of the JHU Bible Corpus.

Dataset Magnitude: All conclusions were drawn in the context of a 25k sentence corpus.

Budget: Our results are specific to a budget of 20% of the tokens/documents.

Sentence Embedding Method: We focused exclusively on SimCSE embeddings.

In future work, we would like to make more robust conclusions about our proposed technique by exploring a broader space of domains, dataset magnitudes, budgets, and sentence embeddings.

[3] The **delfy** algorithm used counts from the English corpus, and our approach used sentence embeddings from the English corpus.

[4] Most documents in this corpus were single sentences.

References

Bhatnagar, R., Ganesh, A., and Kann, K. (2022). CHIA: CHoosing instances to annotate for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7299–7315, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Gao, T., Yao, X., and Chen, D. (2021). SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Maillard, J., Gao, C., Kalbassi, E., Sadagopan, K. R., Goswami, V., Koehn, P., Fan, A., and Guzman, F. (2023). Small data, big impact: Leveraging minimal data for effective machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2740–2756, Toronto, Canada. Association for Computational Linguistics.

McCarthy, A. D., Wicks, R., Lewis, D., Mueller, A., Wu, W., Adams, O., Nicolai, G., Post, M., and Yarowsky, D. (2020). The Johns Hopkins University Bible corpus: 1600+ tongues for typological exploration. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2884–2892, Marseille, France. European Language Resources Association.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Zhao, Y., Zhang, H., Zhou, S., and Zhang, Z. (2020). Active learning approaches to enhancing neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1796–1806, Online. Association for Computational Linguistics.