# Dict-NMT: Bilingual Dictionary based NMT for Extremely Low Resource Languages

**Anonymous ACL submission**

## Abstract

Neural Machine Translation (NMT) models have been effective on large bilingual datasets. However, the existing methods and techniques show that the model's performance is highly dependent on the number of examples in training data. For many languages, having such amount of corpora is a far fetched dream. Taking inspiration from monolingual speakers exploring new languages using bilingual dictionaries, we investigate the applicability of bilingual dictionaries for languages with extremely low, or no bilingual corpus. In this paper, we explore methods using bilingual dictionaries with an NMT model to improve translations for extremely low resource languages. We extend this work for multilingual systems, exhibiting zero-shot property. We present a detailed analysis of effects of quality of dictionary, training dataset size, language family, etc., on the translation quality. Results on multiple low-resource test languages show a clear advantage of our bilingual dictionary-based method over the baselines.

## 1 Introduction

With the growing interest in improving automatic translation systems, deep learning-based models have played a significant role. They have a ubiquitous influence on such solutions. Neural Machine translation has been ruling the roost in recent times both in academia as well as in industries. It has outperformed other translation methods, and even human translators for some languages (Bojar et al., 2016) (Bentivogli et al., 2016) (Barrault et al., 2020). The encoder-decoder framework of NMT models allow them to transfer the semantic and syntactic information more precisely.

One of the major challenges for such languages is training corpora of sufficient size. Such models need for large bilingual or monolingual datasets, where usually it ranges between 1-50 million parallel sentences. For the extremely low resourced languages, dataset with size lesser than 20 thousand parallel sentences, NMT models have not been that successful (Östling and Tiedemann, 2017). The standard approach to this problem has mostly relied on techniques such as transfer learning (Zoph et al., 2016), and data augmentation approaches such as back-translation (Sennrich et al., 2015) (Przystupa and Abdul-Mageed, 2019) and data diversification (Nguyen et al., 2019).

The use of prior knowledge sources for translation of low-resource languages, such as bilingual dictionaries, is still under-explored. The work of (Duan et al., 2020) and (Nag et al., 2020) are closest to ours and both utilize bilingual dictionary, but they use additional large monolingual corpus while we use an extremely small test language's bilingual corpus or no bilingual corpora of the test language at all. The existing approaches mostly depend on the availability of some additional corpora like target monolingual corpus and target-to-source model for back-translation, sister language for transfer learning or additional computations as in data diversification. One of the most common and widely available prior knowledge resources across low resourced languages is the bilingual dictionary which has shown potential in NMT in recent

times.

In our work, we explore the use of bilingual dictionary for translation of extremely low languages. Any meaningful translation requires us to address the points as illustrated in Figure 1. In extremely low resource languages, NMT falls behind given the lack of enormous quantity of data required to train them properly. We study the potential of assisting the NMT models with the contextual dictionary transformation. Our proposed method involves the

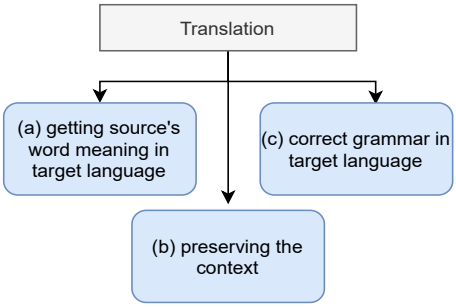



Figure 1: Translation (a) word mapping task, which can be partially, or completely achieved with bilingual dictionary lookup. (b) is about the association a word has with its surrounding words, which in turn affects its alignment. (c) is the transformation of syntactic features of the source language to the source language.

use of bilingual dictionary for addressing the points (Figure 1 (a) and (b)) and an NMT model for (Figure 1 (c)), i.e., we use an NMT model to transform a distorted sentence into a meaningful sentence within the same language. Using this method, we propose two simple frameworks which can be extremely useful for languages with extremely less or no corpus available but having a bilingual dictionary. Summarizing the contributions of our paper as follows:

- We introduce a simple and effective method for incorporating a bilingual dictionary in a neural machine translation task.
- We propose a one-to-one bilingual dictionary based NMT model for extremely low-resource languages.
- We propose a many-to-one NMT model capable of translating for languages it has never seen in the training sets.

We provide a brief description of our method in Section 2. We discuss the usage of the bilingual dictionary, tokenizer, and the NMT model. We explore the applicability of our proposed method in two settings, extremely less corpus and no corpus available for concerned language. In Section 3, we describe our one-to-one translation framework useful for translation in extremely low resource setting. We provide a detailed analysis with comparison among translation quality, dataset size and dictionary quality. In Section 4, we provide detailed information about our proposed many-to-one translation framework, which shows zero-shot property. We summarize and conclude our results and contributions in Section 5.

## 2 Dict-NMT: Assisting NMT model with bilingual dictionary

We propose a simple yet effective method of translation, dict-NMT, using an NMT model with the help of the respective languages' bilingual dictionary. We use a bilingual dictionary as word-to-word translator to convert words from the source language to an intermediate sequence. This distorted sentence in the target language is then fed to an NMT model, here Transformer, to learn the relation between the intermediate sequence and ground truth (Figure 2). This opens up doors for various frameworks for translation. One simple way is to apply this method to a one-to-one translation system (Section 3). Furthermore, one can also devise a many-to-one translation framework (Section 4), where the NMT model is trained on word-to-word translations from various languages. This generalised model can then be used even for languages which were not used in the training data. Other possible ways include fine-tuning the generalised model on a specific language. Other data augmentation methods, such as backtranslation and data diversification, are also applicable to our proposed method. Another possible way of augment-

ing data is by adding intra-shuffled (i.e., words within a sentence $s$ are shuffled), noisy (replace tokens in $s$ with random tokens with some probability) sentences from target language to the training data. We leave these methods for future work.

### 2.1 Bilingual Dictionary

A dictionary is a map of words from the source language to the target language, where the mapping can be one to many. Here, we consider mappings that are word to word and not word to phrase.

First, using the dictionary, we change the source language sentences into an intermediate sequence. This step would reduce the workload on our NMT model from learning the word meanings from the available small dataset. If a word in the source sentence is present in the dictionary then it is converted accordingly in intermediate sequence, otherwise, it remains unchanged in the intermediate sequence, i.e., we consider the word to be in the target language space. When using the dictionary, there might exist multiple target language words as meaning for a source language word. We settle this problem of polysemy by selecting the word most similar to the previous word's dictionary translation (using target language's pre-trained word embeddings). This would help us to preserve the contextual information.

More precisely, for any source language $S$ and target language $T$ with a bilingual dictionary $D_{S \to T}$, the first step is to translate the text in $S$ word-to-word to $T$ using $D_{S \to T}$. In case the mapping is not available for any word $w$ in $S$, it is mapped to itself and is considered a random noise in $T$. For the case of polysemous words, we take the help of word embeddings of $T$. We select the word (in $T$) most similar to the previous word's dictionary translation (using target language's pre-trained word embeddings). For instance, given is a sentence $s = \{s_1, s_2, ..., s_n\}$ in $S$ with word-to-word translation $t = \{t_1, t_2, ..., t_n\}$ in $T$. For any $s_i (i > 1)$ having dictionary translations $D_{S \to T}(s_i) = \{t_i^1, t_i^2, ..., t_i^m\}$, we

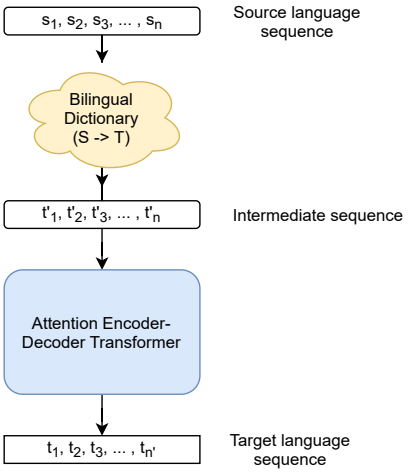



Figure 2: Our proposed method involving bilingual dictionary for NMT.

select its translation as

$$t_i = \underset{t \in D_{S \to T}(s_i)}{argmax}\ similarity(t_j, t)$$

where $j = \max\{j' < i | s_j \in Dom(D_{S \to T})\}$. We randomly select translation for the first occurring polysemous word. Here, for our experiments, we assume that the target language is a popular one, thus, decent word embeddings for $T$ exist. This method would help us to preserve the contextual information. However, if the first randomly selected word is erroneous, the trailing polysemous words might have incorrect translations.

### 2.2 Tokenizer

Tokenization is the process of breaking the given text into smaller chunks. Since the model input and output are in the same language, we share the tokenizer for both of them. The intermediate representation might consist words from foreign language. Thus, instead of using the traditional whitespace tokenizer and giving all such words a <OOV> token, we use subspace tokenizer to handle the large amount of out-of-vocabulary words. This way, the noise created by the tokens of foreign language, would help the model being more robust.

### 2.3 NMT Model

Since both intermediate sequence and target belong to the same language, the NMT

**Algorithm 1:** Dict-NMT. $D_i = (S_i, T)$ is a set of parallel sentences from language $S_i$ to $T$. Corresponding to the language pair $(S_i, T)$, we have a bilingual dictionary $B_i$ where $B_i(s)$ is a word-to-word translation of $s$. We train the model $M$ on $\{D_i\}_{i=1}^n$ using $\{B_i\}_{i=1}^n$.

**Procedure** `Train` ($\{D_i\}_{i=1}^n$, $\{B_i\}_{i=1}^n$, $p$)**:**
  M *randomly initialised NMT model.*
  Train M on `create_dataset`($\{D_i\}_{i=1}^n$, $\{B_i\}_n$, p) until it converges
  **return** M
**Function** `create_dataset` ($\{D_i\}_{i=1}^n$, $\{B_i\}_n$, $p$)**:**
  $D' = \phi$
  **for** $D \in \{D_i\}_{i=1}^n$ **do**
    **for** *(s, t)* $\in D$ **do**
      **if** $count_T(B_i(s)) \geq p$ **then**
        $D'$ = $D' \cup$ ($B_i(s)$, t)
  `/* count`$_T$`(B`$_s$`) = % of words in` $s$ `having dictionary translations */`
  *shuffle* $D'$
  **return** $D'$

model is relieved from learning the word meanings. The model will now try to focus mostly on learning the grammar for the target language space. The NMT model learns the mappings from the source invariant representations coming from various languages to the target language and tries to generalise which would be beneficial for unknown languages.

Our proposed method can be applied to any NMT model. For our experiments, we use the state-of-the-art Transformer (Vaswani et al., 2017) model. Since, the intermediate sentence, i.e., the input for the Transformer, in itself does not make any sense, the attention mechanism helps to understand the dependencies of words through the whole sequence. The encoder-decoder framework allows us to find the meaning of the words not translated by the dictionary while preserving the context.

## 3 Dict-NMT for one-to-one translation

We propose a dictionary based one-to-one translation framework for extremely low resource settings. Given a language pair $(S, T)$, we train an NMT model on the word-to-word dictionary translations of $S$, and T ($i = 1$ in Algo 1).

### 3.1 Experimental Settings

We extensively check the effectiveness of the dictionary (by varying the dictionary percentage) across five European languages' translation tasks as well as the size of the bilingual corpora. We keep Transformer as our baseline model. We use 4 layer Transformer with 100 embedding/hidden units, and 400 feed-forward filter size. We tie source and target embeddings. We keep batch size 32, epochs 50, dropout 0.1 and optimizer Adam. In this work, we use the pre-trained BERT WordPiece tokenizer (Devlin et al., 2019), a subword tokenizer.

#### 3.1.1 Bilingual dictionary

For our experiments, we use the publicly available Facebook MUSE's[1] bilingual dictionary, which consists of 110 large-scale ground-truth bilingual dictionaries (Conneau et al., 2017). For preserving the context while dictionary translation, we

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

use the Fasttext embeddings (Bojanowski et al., 2017).

#### 3.1.2 Data

For our experiments, we consider Europarl v7 parallel corpus (Koehn, 2002) for Pt-En, Sv-En, Nl-En, Pt-En, and Fr-En language pairs. Here we selected English as our target language in all the cases. The intuition behind this is that any bilingual dictionary for an extremely low resource language would be created by taking a commonly used language, so that it can be of practical use. As English is one such language, we tried our experiments with it.

We filter each sentence such that it contains at most 80 tokens. We use these data in low resource setting, i.e. we use only 2K, 8K, 16K and 20K data size for each language. We create these datasets according to the percentage of words from each sentences available in the corresponding dictionary, precisely we did this for 50%-80% (Table 1). For each data size 0.05% is the test set and rest we use as training set.

### 3.2 Results and Analysis

We perform intensive experiments on the effectiveness of training data size and dictionary coverage on the performance of the translation system. Table 2 shows comparison between the baseline (bilingual dictionary based word-to-word translation) and our proposed method. The best scores for each language pair, along with the dictionary coverage are reported in the table. The best result is chosen over the dictionary coverage ($50\% - 80\%$), i.e. least percentage of words in each sentence available in the bilingual dictionary, and varied dataset size ($2K - 20K$). We report arithmetic mean of scores on 3 different datasets sampled from the same large data. The scores show a significant increase from simple word-to-word translations ($4.8 - 8.6$). We performed experiments for three language pairs with simple transformer as well. However, due to very less data, the model seemed to struggle considerably. For Pt-En, Sv-En, and Fr-En language pairs, best scores on $20K$ dataset came to be 1.56, 1.39, and 1.37 respectively. This shows there is a clear advantage of using the proposed method for extremely low resource languages.

In Figure 3, we have heat-maps of BLEU

| Dict % | Ro-En (399.37K) | Pt-En (1.96M) | Fr-En (2M) | It-En (1.9M) | Es-En (1.97M) |
|---|---|---|---|---|---|
| 50 | 104K | 438.6K | 1.4M | 941K | 1.5M |
| 60 | 22.6K | 77.7K | 536.4K | 202.6K | 631.6K |
| 70 | 4K | 11.3K | 106.2K | 26.8K | 111.4K |
| 80 | 975 | 2.5K | 17.7K | 4.8K | 15.9K |
| 90 | 244 | 937 | 2.6K | 1.9K | 2.5K |
| 100 | 230 | 920 | 2K | 1.8K | 2K |

Italic

| Dict % | Da-En (1.97M) | Sv-En (1.86M) | De-En (1.92M) | Nl-En (2M) |
|---|---|---|---|---|
| 50 | 1.3M | 1.7M | 1.9M | 1.7M |
| 60 | 506.6K | 1.3M | 1.7M | 1M |
| 70 | 107.9K | 563.7K | 1M | 31.7K |
| 80 | 20.2K | 124.9K | 317.5K | 55K |
| 90 | 2.6K | 9.2K | 28.2K | 4.7K |
| 100 | 1.9K | 1.8K | 3.3K | 1.9K |

Germanic

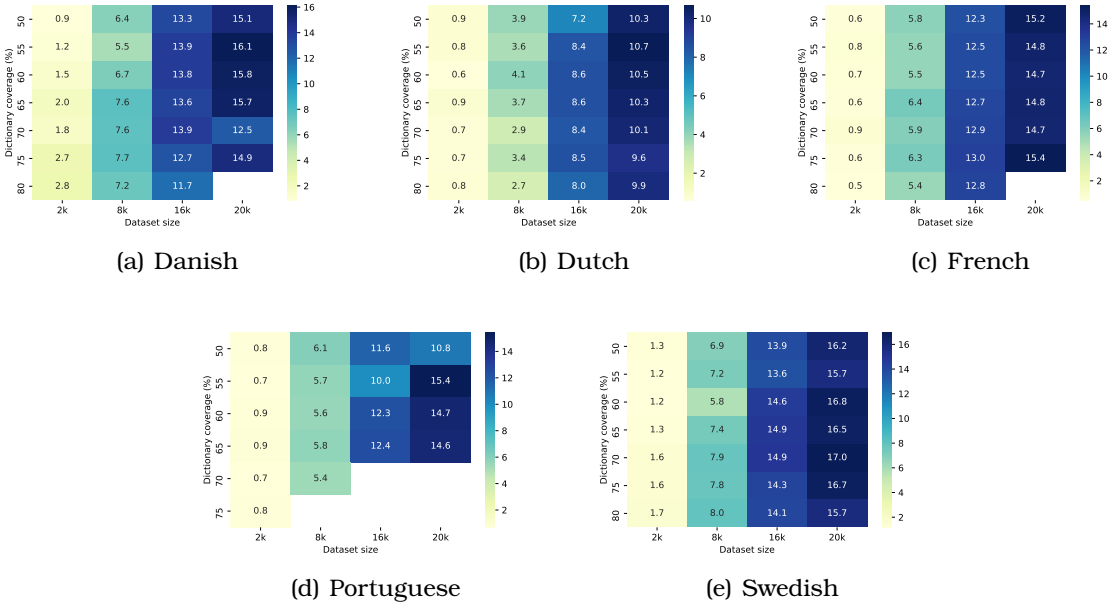
| Dict % | Bg-En (406.9K) | Cz-En (646.6K) | Pl-En (632.57K) | Sl-En (623.49M) | Sk-En (640.72K) |
|---|---|---|---|---|---|
| 50 | 33.2K | 136.8K | 137.8K | 64.6K | 180.4K |
| 60 | 6.4K | 36.6K | 36.5K | 16K | 49.2K |
| 70 | 1K | 9.5K | 8.6K | 3.9K | 11.3K |
| 80 | 233K | 3.1K | 2.7K | 1.6K | 3.5K |
| 90 | 58K | 1.4K | 1.1K | 1.1K | 1.4K |
| 100 | 57K | 1.3K | 1K | 1.1K | 1.3K |

Slavic

Table 1: Dataset size after filtering sentences containing at least Dict% of dictionary words.

| Language | Dict % | BLEU | |
|---|---|---|---|
| | | Base | w D |
| Pt-En | 55% | 6.9 | **15.4** |
| Sv-En | 70% | 8.4 | **17.0** |
| Nl-En | 65% | 5.9 | **10.7** |
| Fr-En | 65% | 8.6 | **15.4** |
| Da-En | 65% | 7.7 | **16.1** |

Table 2: Results: Best BLEU score for each language. w D is “with dictionary", i.e. our proposed method, Base is “dictionary baseline" which is simply word-to-word translation . We get the best scores for maximum data size (i.e., 20K). Training set dictionary coverage (Dict %) is given for each corresponding score. The BLEU scores are calculated using SacreBLEU’s corpus_bleu (Post, 2018)




(a) Danish (b) Dutch (c) French

(d) Portuguese (e) Swedish

Figure 3: BLEU Scores for One-to-One translation method. We report the scores on test data with $80\%$ dictionary coverage, as it was maximum in every case.

scores for different languages calculated over different datasets. The x-axis shows the size of the dataset and the y-axis shows the dictionary percentage. We can have the following observations from the maps,

- **BLEU VS Dataset size:** The model clearly benefits from increasing the dataset size in an extremely low resource setting. There is a direct correlation between the score and the number of training examples.

- **BLEU VS Training data dictionary coverage:** Dictionary coverage can be seen as inversely proportional to the amount of noise generated by the untranslated words from the dictionary. The best scores for each column of any map are always somewhere in the middle (except $20K$ dataset of French). We suspect this behavior is linked to finding the correct balance of noise and generalisation. With more noise (less dictionary coverage), the method seems to get more robust, however, it underfits when it is exposed to too many of them.

## 4 Dict-NMT for many-to-one translation

A conventional idea for a many-to-one model would involve mapping the source text to a common representation space which would then further be used by the model to generate the translations. Fixing the target language, we can create a common representation for any given source language by translating the source text word-to-word into the target language using the bilingual dictionary. This is similar to how we humans translate any foreign language with the help of a bilingual dictionary.

We propose a many-to-one translation framework, which, just using a bilingual dictionary, can translate for languages which are not present in the training phase- absolute zero-shot translation. Given a test language pair $(S, T)$, we train an NMT model on dictionary based word-to-word translations of language pairs $\{(S_i, T)\}_{i=1}^{n}$, where $S \neq S_i$ for $i = 1, .., n$. Our goal is to make the model invariant of source language. We achieve this via adding word-to-word dictionary translations from various languages coming from different families (Algo 1).

### 4.1 Experimental Settings

We perform a comprehensive study on the effect of dataset size, no. of languages, inclusion of test language family, and dictionary coverage in test set, on the translation quality. We perform our experiments on European languages with English as the target language. We keep the tokenizer, NMT model and its hyperparameters similar to that of previous experiment's setting. We use Facebook MUSE's bilingual dictionary for this experiment as well.

#### 4.1.1 Dataset

We perform our experiments on Europarl v7 parallel corpus, fixing English as our target language. We used languages from three families, namely, Italic (Romanian, Spanish, Portuguese, French, Italian), Slavic (Bulgarian, Czech, Polish, Slovene, Slovak) and Germanic (Danish, Swedish, German, Dutch). We use Romanian, Bulgarian, and Danish as our test languages. We analyse our results on training data, intra and inter combination of the language families with sizes $50K$, $150K$, $500K$, and $1M$. We test our experiments on $600$ sentences. $create\_dataset()$(Algo 1) shows how we created training data for our experiment. In our experiments, for a training set $create\_dataset(\{D_i\}_{i=1}^{n}, \{B_i\}_n, p)$ (Algo 1), we take equal number of sentences from all $n$ languages. The case of polysemy is handled the same way as it was in the previous experiment. We use the notation "All" for the combination of the above mentioned languages from all three language families (Italic, Germanic, and Slavic).

### 4.2 Results and Analysis

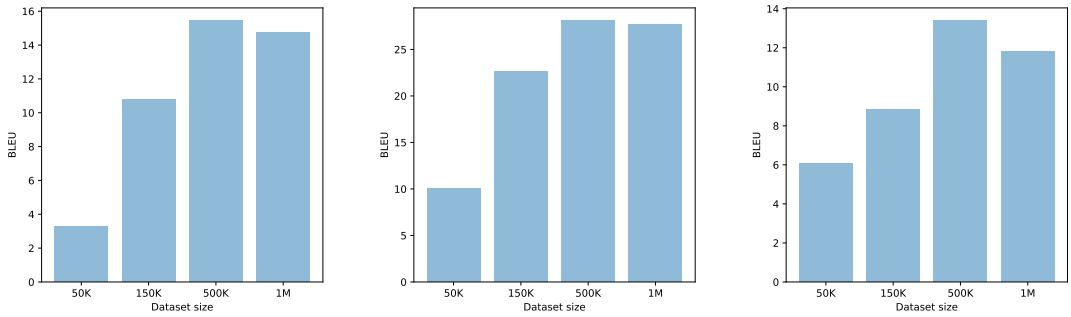
We present the best scores for three language pairs, Ro-En, Bg-En and Da-En in Table 3. We further compare the scores with word-to-word dictionary translations. We choose the best score over varied training data size ($50K$ - $1M$), Test set dictionary coverage ($0\%$ - $80\%$), and combination of language families. There is a significant difference in scores of baseline and our proposed method. Considering the fact that the training sample has no examples from test languages, the resultant score demonstrates the zero-shot property of the proposed method.

| Language | Dict % | BLEU | |
|---|---|---|---|
| | | Base | w D |
| Ro-En | 50% | 9.4 | **28.1** |
| Bg-En | 50% | 8.2 | **15.4** |
| Da-En | 50% | 7.7 | **13.4** |

Table 3: Results: Best BLEU score for each language. w D is "with dictionary", i.e. our proposed method, Base is "dictionary baseline" which is simply word-to-word translation . We get the best scores for "All" dataset (Germanic Italic Slavic) with $500K$ parallel sentences. Training set dictionary coverage (Dict %) is given for each corresponding score. The test set of Ro and Bg has atleast $80\%$ dictionary coverage, while Da has $70\%$ (the raw dataset was too little to make a decent test dataset with similar number of samples).

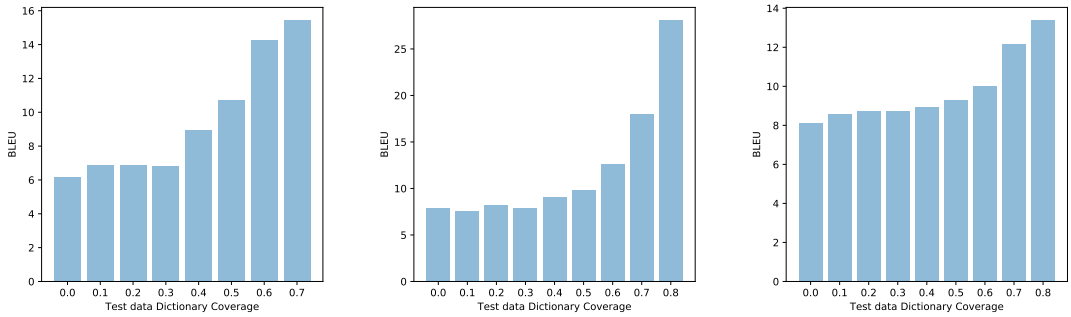
We perform experiments to test effect of dataset size, inclusion of test language family, and test data dictionary coverage.

- **BLEU VS Test data dictionary coverage:** From figure 5, it is evident that the scores increase with the increase in dictionary coverage of test data, i.e., the NMT model gets better assisted with more word-to-word translations in a given sentence.
- **BLEU VS Training set data size:** With increase in data size, the scores increase as well (Table 4). However, it tends to converge on the data size between $500K$ and $1M$.
- **Effect of addition of language families:** From Figure 6 it can be observed that the score stays the least for model trained just on Germanic family. There is a slight increase in score for Italic family. However, it increases significantly when we start combining language families together. We get the highest score for "All", which is a combination of all three language families. There is a slight decrease in score when we add sentences from two different languages.




(a) Bulgarian (b) Romanian (c) Danish

Figure 4: BLEU VS Training Dataset Size




(a) Bulgarian (b) Romanian (c) Danish

Figure 5: BLEU VS Test Data Dictionary Coverage. Here, the reported scores are for model trained on "All" dataset with size $500K$

We suspect less number of parameters of the model to be the reason behind such behaviour. For better generalisation on more number of languages, we believe larger NMT models would be beneficial.

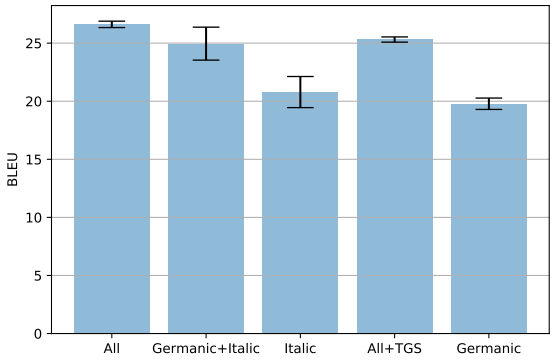



Figure 6: BLEU VS Effect of Test Family (Romanian) (TGS = Turkish + Greeek)

## 5 Conclusion

Using Europarl corpus, we showed that our method of incorporating bilingual dictionaries for NMT tasks can be quite effective. Given a dictionary, it not only works for languages with extremely low corpus, but also for languages with no parallel or monolingual corpus at all. We analyze the extent of improvement that can be done by varying dictionary percentage and with the range of size of datasets. This work can be extended by blending our method with other state-of-the-art approaches such as back translation, and transfer learning. We believe this work will motivate researchers to explore other possibilities of incorporating bilingual dictionaries for NMT in extremely low resource settings.

## References


Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. *arXiv preprint arXiv:1608.04631*.

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.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiangyu Duan, Baijun Ji, Hao Jia, Min Tan, Min Zhang, Boxing Chen, Weihua Luo, and Yue Zhang. 2020. Bilingual dictionary based neural machine translation without using parallel sentences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1570–1579.

Philipp Koehn. 2002. Europarl: A multilingual corpus for evaluation of machine translation. Draft.

Sreyashi Nag, Mihir Kale, Varun Lakshminarasimhan, and Swapnil Singhavi. 2020. Incorporating bilingual dictionaries for low resource semi-supervised neural machine translation. *arXiv preprint arXiv:2004.02071*.

Xuan-Phi Nguyen, Shafiq Joty, Wu Kui, Ai Ti Aw, Jiuxiang Gu, Jason Kuen, Shafiq Joty, Jianfei Cai, Vlad Morariu, Handong Zhao, et al. 2019. Data diversification: An elegant strategy for neural machine translation.

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

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Michael Przystupa and Muhammad Abdul-Mageed. 2019. Neural machine translation of low-resource and similar languages with backtranslation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 224–235.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.