# Universal Neural Machine Translation for Extremely Low Resource Languages

**Jiatao Gu**[†*] **Hany Hassan**[‡] **Jacob Devlin**[□*] **Victor O.K. Li**[†]
[†]The University of Hong Kong [‡]Microsoft Research
{jiataogu, vli}@eee.hku.hk hanyh@microsoft.com
[□]Google Research
jacobdevlin@google.com

## Abstract

In this paper, we propose a new universal machine translation approach focusing on languages with a limited amount of parallel data. Our proposed approach utilizes a transfer-learning approach to share lexical and sentence level representations across multiple source languages into one target language. The lexical part is shared through a Universal Lexical Representation to support multilingual word-level sharing. The sentence-level sharing is represented by a model of experts from all source languages that share the source encoders with all other languages. This enables the low-resource language to utilize the lexical and sentence representations of the higher resource languages. Our approach is able to achieve 23 BLEU on Romanian-English WMT2016 using a tiny parallel corpus of 6k sentences, compared to the 18 BLEU of strong baseline system which uses multilingual training and back-translation. Furthermore, we show that the proposed approach can achieve almost 20 BLEU on the same dataset through fine-tuning a pre-trained multi-lingual system in a zero-shot setting.

## 1 Introduction

Neural Machine Translation (NMT) (Bahdanau et al., 2015) has achieved remarkable translation quality in various on-line large-scale systems (Wu et al., 2016; Devlin, 2017) as well as achieving state-of-the-art results on Chinese-English translation (Hassan et al., 2018). With such large systems, NMT showed that it can scale up to immense amounts of parallel data in the order of tens of millions of sentences. However, such data is not widely available for all language pairs and domains. In this paper, we propose a novel universal multi-lingual NMT approach focusing mainly on low resource languages to overcome the limitations of NMT and leverage the capabilities of multi-lingual NMT in such scenarios.

Our approach utilizes multi-lingual neural translation system to share lexical and sentence level representations across multiple source languages into one target language. In this setup, some of the source languages may be of extremely limited or even zero data. The lexical sharing is represented by a universal word-level representation where various words from all source languages share the same underlaying representation. The sharing module utilizes monolingual embeddings along with seed parallel data from all languages to build the universal representation. The sentence-level sharing is represented by a model of language experts which enables low-resource languages to utilize the sentence representation of the higher resource languages. This allows the system to translate from any language even with tiny amount of parallel resources.

We evaluate the proposed approach on 3 different languages with tiny or even zero parallel data. We show that for the simulated “zero-resource" settings, our model can consistently outperform a strong multi-lingual NMT baseline with a tiny amount of parallel sentence pairs.

## 2 Motivation

Neural Machine Translation (NMT) (Bahdanau et al., 2015; Sutskever et al., 2014) is based on Sequence-to-Sequence encoder-decoder model along with an attention mechanism to enable better handling of longer sentences (Bahdanau et al., 2015). Attentional sequence-to-sequence models are modeling the log conditional probability of the

[*]This work was done while the authors at Microsoft.


Figure 1: BLEU scores reported on the test set for Ro-En. The amount of training data effects the translation performance dramatically using a single NMT model.

translation $Y$ given an input sequence $X$. In general, the NMT system $\theta$ consists of two components: an encoder $\theta_e$ which transforms the input sequence into an array of continuous representations, and a decoder $\theta_d$ that dynamically reads the encoder's output with an attention mechanism and predicts the distribution of each target word. Generally, $\theta$ is trained to maximize the likelihood on a training set consisting of $N$ parallel sentences:

$$\begin{aligned}\mathcal{L}(\theta) &= \frac{1}{N}\sum_{n=1}^{N}\log p\left(Y^{(n)}|X^{(n)};\theta\right)\\ &= \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T}\log p\left(y_t^{(n)}|y_{1:t-1}^{(n)}, f_t^{\text{att}}(h_{1:T_s}^{(n)})\right)\end{aligned} \tag{1}$$

where at each step, $f_t^{\text{att}}$ builds the attention mechanism over the encoder's output $h_{1:T_s}$. More precisely, let the vocabulary size of source words as $V$

$$h_{1:T_s} = f^{\text{ext}}\left[e_{x_1}, ..., e_{x_{T_s}}\right], \quad e_x = E^I(x) \tag{2}$$

where $E^I \in \mathbb{R}^{V\times d}$ is a look-up table of source embeddings, assigning each individual word a unique embedding vector; $f^{\text{ext}}$ is a sentence-level feature extractor and is usually implemented by a multi-layer bidirectional RNN (Bahdanau et al., 2015; Wu et al., 2016), recent efforts also achieved the state-of-the-art using non-recurrence $f^{\text{ext}}$, e.g. ConvS2S (Gehring et al., 2017) and Transformer (Vaswani et al., 2017).

**Extremely Low-Resource NMT** Both $\theta_e$ and $\theta_d$ should be trained to converge using parallel training examples. However, the performance is highly correlated to the amount of training data. As shown in Figure. 1, the system cannot achieve reasonable translation quality when the number of the parallel examples is extremely small ($N \approx 13k$ sentences, or not available at all $N = 0$).

**Multi-lingual NMT** Lee et al. (2017) and Johnson et al. (2017) have shown that NMT is quite efficient for multilingual machine translation. Assuming the translation from $K$ source languages into one target language, a system is trained with maximum likelihood on the mixed parallel pairs $\{X^{(n,k)}, Y^{(n,k)}\}_{k=1...K}^{n=1...N_k}$, that is

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{k=1}^{K}\sum_{n=1}^{N_k}\log p\left(Y^{(n,k)}|X^{(n,k)};\theta\right) \tag{3}$$

where $N = \sum_{k=1}^{K} N_k$. As the input layer, the system assumes a multilingual vocabulary which is usually the union of all source language vocabularies with a total size as $V = \sum_{k=1}^{K} V_k$. In practice, it is essential to shuffle the multilingual sentence pairs into mini-batches so that different languages can be trained equally. Multi-lingual NMT is quite appealing for low-resource languages; several papers highlighted the characteristic that make it a good fit for that such as Lee et al. (2017), Johnson et al. (2017), Zoph et al. (2016) and Firat et al. (2016). Multi-lingual NMT utilizes the training examples of multiple languages to regularize the models avoiding over-fitting to the limited data of the smaller languages. Moreover, the model transfers the translation knowledge from high-resource languages to low-resource ones. Finlay, the decoder part of the model is sufficiently trained since it shares multilingual examples from all languages.

### 2.1 Challenges

Despite the success of training multi-lingual NMT systems; there are a couple of challenges to leverage them for zero-resource languages:

**Lexical-level Sharing** Conventionally, a multi-lingual NMT model has a vocabulary that represents the union of the vocabularies of all source languages. Therefore, the multi-lingual words do not practically share the same embedding space since each word has its own representation. This does not pose a problem for languages with sufficiently large amount of data, yet it is a major limitation for extremely low resource languages since most of the vocabulary items will not have enough, if any, training examples to get a reliably trained models.

A possible solution is to share the surface form of all source languages through sharing sub-units

such as subwords (Sennrich et al., 2016b) or characters (Kim et al., 2016; Luong and Manning, 2016; Lee et al., 2017). However, for an arbitrary low-resource language we cannot assume significant overlap in the lexical surface forms compared to the high-resource languages. The low-resource language may not even share the same character set as any high-resource language. It is crucial to create a shared semantic representation across all languages that does not rely on surface form overlap.

**Sentence-level Sharing** It is also crucial for low-resource languages to share source sentence representation with other similar languages. For example, if a language shares syntactic order with another language it should be feasible for the low-resource language to share such representation with another high recourse language. It is also important to utilize monolingual data to learn such representation since the low or zero resource language may have monolingual resources only.

## 3 Universal Neural Machine Translation

We propose a Universal NMT system that is focused on the scenario where minimal parallel sentences are available. As shown in Fig. 2, we introduce two components to extend the conventional multi-lingual NMT system (Johnson et al., 2017): Universal Lexical Representation (ULR) and Mixture of Language Experts (MoLE) to enable both word-level and sentence-level sharing, respectively.

### 3.1 Universal Lexical Representation (ULR)

As we highlighted above, it is not straightforward to have a universal representation for all languages. One potential approach is to use a shared source vocabulary, but this is not adequate since it assumes significant surface-form overlap in order being able to generalize between high-resource and low-resource languages. Alternatively, we could train monolingual embeddings in a shared space and use these as the input to our MT system. However, since these embeddings are trained on a monolingual objective, they will not be optimal for an NMT objective. If we simply allow them to change during NMT training, then this will not generalize to the low-resource language where many of the words are unseen in the parallel data. Therefore, our goal is to create a shared embedding space which (a) is trained towards NMT rather than a monolingual objective, (b) is not based on lexical surface forms, and (c) will generalize from the high-resource languages to the low-resource language.

We propose a novel representation for multilingual embedding where each word from any language is represented as a probabilistic mixture of universal-space word embeddings. In this way, semantically similar words from different languages will naturally have similar representations. Our method achieves this utilizing a discrete (but probabilistic) "universal token space", and then learning the embedding matrix for these universal tokens directly in our NMT training.

**Lexicon Mapping to the Universal Token Space** We first define a discrete universal token set of size $M$ into which all source languages will be projected. In principle, this could correspond to any human or symbolic language, but all experiments here use English as the basis for the universal token space. As shown in Figure 2, we have multiple embedding representations. $E^Q$ is language-specific embedding trained on monolingual data and $E^K$ is universal tokens embedding. The matrices $E^K$ and $E^Q$ are created beforehand and are not trainable during NMT training. $E^U$ is the embedding matrix for these universal tokens which is learned during our NMT training. It is worth noting that shaded parts in Figure2 are trainable during NMT training process.

Therefore, each source word $e_x$ is represented as a mixture of universal tokens $M$ of $E^U$.

$$e_x = \sum_{i=1}^{M} E^U(u_i) \cdot q(u_i|x) \tag{4}$$

where $E^U$ is an NMT embedding matrix, which is learned during NMT training.

The mapping $q$ projects the multilingual words into the universal space based on their semantic similarity. That is, $q(u|x)$ is a distribution based on the distance $D_s(u, x)$ between $u$ and $x$ as:

$$q(u_i|x) = \frac{e^{D(u_i,x)/\tau}}{\sum_{u_j} e^{D(u_j,x)/\tau}} \tag{5}$$

where $\tau$ is a temperature and $D(u_i, x)$ is a scalar score which represents the similarity between source word $x$ and universal token $u_i$:

$$D(u, x) = E^K(u) \cdot A \cdot E^Q(x)^T \tag{6}$$

where $E^K(u)$ is the "key" embedding of word $u$, $E^Q(x)$ is the "query" embedding of source word $x$.

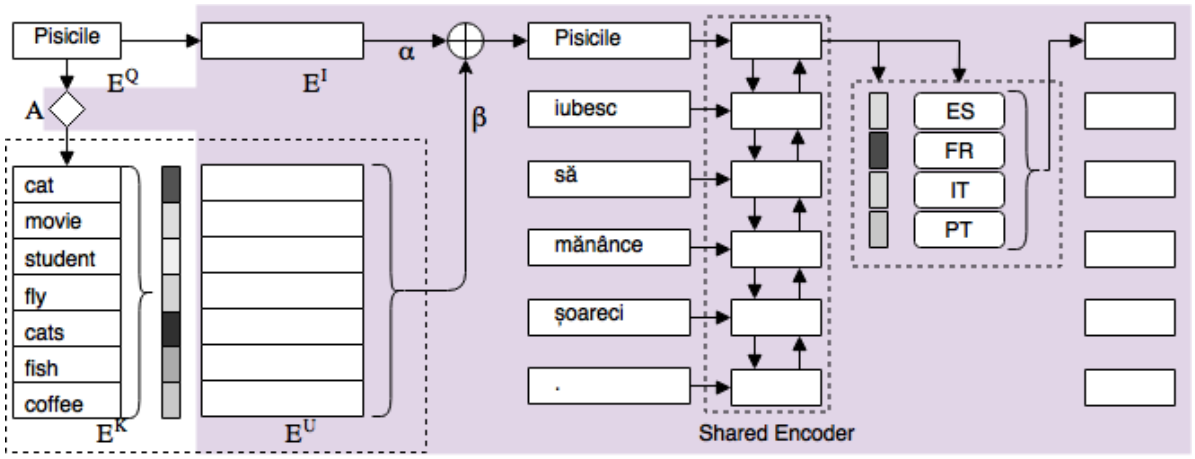


Figure 2: An illustration of the proposed architecture of the ULR and MoLE. Shaded parts are trained within NMT model while unshaded parts are not changed during training.

The transformation matrix $A$, which is initialized to the identity matrix, is learned during NMT training and shared across all languages.

This is a key-value representation, where the queries are the monolingual language-specific embedding, the keys are the universal tokens embeddings and the values are a probabilistic distribution over the universal NMT embeddings. This can represent unlimited multi-lingual vocabulary that has never been observed in the parallel training data. It is worth noting that the trainable transformation matrix $A$ is added to the query matching mechanism with the main purpose to tune the similarity scores towards the translation task. $A$ is shared across all languages and optimized discriminatively during NMT training such that the system can fine-tune the similarity score $q()$ to be optimal for NMT.

**Shared Monolingual Embeddings** In general, we create one $E^Q$ matrix per source language, as well as a single $E^K$ matrix in our universal token language. For Equation 6 to make sense and generalize across language pairs, all of these embedding matrices must live in a similar semantic space. To do this, we first train off-the-shelf monolingual word embeddings in each language, and then learn one projection matrix per source language which maps the original monolingual embeddings into $E^K$ space. Typically, we need a list of *source word - universal token* pairs (seeds $S_k$) to train the projection matrix for language $k$. Since vectors are normalized, learning the optimal projection is equivalent to finding an orthogonal transformation $O_k$ that makes the projected word vectors as close as to its corresponded universal tokens:

$$\max_{O_k} \sum_{(\tilde{x},\tilde{y})\in S_k} \left(E^{Q_k}(\tilde{x}) \cdot O_k\right) \cdot E^K(\tilde{y})^T \quad \text{s.t. } O_k^T O_k = I, \quad k = 1, ..., K \tag{7}$$

which can be solved by SVD decomposition based on the seeds (Smith et al., 2017). In this paper, we chose to use a short list of seeds from automatic word-alignment of parallel sentences to learn the projection. However, recent efforts (Artetxe et al., 2017; Conneau et al., 2018) also showed that it is possible to learn the transformation without any seeds, which makes it feasible for our proposed method to be utilized in purely zero parallel resource cases.

It is worth noting that $O_k$ is a language-specific matrix which maps the monolingual embeddings of each source language into a similar semantic space as the universal token language.

**Interpolated Embeddings** Certain lexical categories (e.g. function words) are poorly captured by Equation 4. Luckily, function words often have very high frequency, and can be estimated robustly from even a tiny amount of data. This motivates an interpolated $e_x$ where embeddings for very frequent words are optimized directly and not through the universal tokens:

$$\alpha(x)E^I(x) + \beta(x)\sum_{i=1}^{M} E^U(u_i) \cdot q(u_i|x) \tag{8}$$

Where $E^I(x)$ is a language-specific embedding of word $x$ which is optimized during NMT training. In general, we set $\alpha(x)$ to 1.0 for the top $k$ most frequent words in each language, and 0.0 otherwise,

where $k$ is set to 500 in this work. It is worth noting that we do not use an absolute frequency cutoff because this would cause a mismatch between high-resource and low-resource languages, which we want to avoid. We keep $\beta(x)$ fixed to 1.0.

**An Example** To give a concrete example, imagine that our target language is English (En), our high-resource auxiliary source languages are Spanish (Es) and French (Fr), and our low-resource source language is Romanian (Ro). En is also used for the universal token set. We assume to have 10M+ parallel Es-En and Fr-En, and a few thousand in Ro-En. We also have millions of monolingual sentences in each language.

We first train word2vec embeddings on monolingual corpora from each of the four languages. We next align the Es-En, Fr-En, and Ro-En parallel corpora and extract a seed dictionary of a few hundred words per language, e.g., `gato` $\rightarrow$ `cat`, `chien` $\rightarrow$ `dog`. We then learn three matrices $O_1, O_2, O_3$ to project the Es, Fr and Ro embeddings ($E^{Q_1}, E^{Q_2}, E^{Q_3}$), into En ($E^K$) based on these seed dictionaries. At this point, Equation 5 should produce *reasonable* alignments between the source languages and En, e.g., $q(\texttt{horse}|\texttt{magar}) = 0.5$, $q(\texttt{donkey}|\texttt{magar}) = 0.3$, $q(\texttt{cow}|\texttt{magar}) = 0.2$, where `magar` is the Ro word for `donkey`.

### 3.2 Mixture of Language Experts (MoLE)

As we paved the road for having a universal embedding representation; it is crucial to have a language-sensitive module for the encoder that would help in modeling various language structures which may vary between different languages. We propose a Mixture of Language Experts (MoLE) to model the sentence-level universal encoder. As shown in Fig. 2, an additional module of mixture of experts is used after the last layer of the encoder. Similar to (Shazeer et al., 2017), we have a set of expert networks and a gating network to control the weight of each expert. More precisely, we have a set of expert networks as $f_1(h), ..., f_K(h)$ where for each expert, a two-layer feed-forward network which reads the output hidden states $h$ of the encoder is utilized. The output of the MoLE module $h'$ will be a weighted sum of these experts to replace the encoder's representation:

$$h' = \sum_{k=1}^{K} f_k(h) \cdot \text{softmax}(g(h))_k, \qquad (9)$$

where an one-layer feed-forward network $g(h)$ is used as a gate to compute scores for all the experts.

In our case, we create one expert per auxiliary language. In other words, we train to only use expert $f_i$ when training on a parallel sentence from auxiliary language $i$. Assume the language $1...K-1$ are the auxiliary languages. That is, we have a multi-task objective as:

$$\mathcal{L}^{\text{gate}} = \sum_{k=1}^{K-1} \sum_{n=1}^{N_k} \log \left[\text{softmax}\left(g(h)\right)_k\right] \qquad (10)$$

We do not update the MoLE module for training on a sentence from the low-resource language. Intuitively, this allows us to represent each token in the low-resource language as a context-dependent mixture of the auxiliary language experts.

## 4 Experiments

We extensively study the effectiveness of the proposed methods by evaluating on three "almost-zero-resource" language pairs with variant auxiliary languages. The vanilla single-source NMT and the multi-lingual NMT models are used as baselines.

### 4.1 Settings

**Dataset** We empirically evaluate the proposed Universal NMT system on 3 languages – Romanian (Ro) / Latvian (Lv) / Korean (Ko) – translating to English (En) in near zero-resource settings. To achieve this, single or multiple auxiliary languages from Czech (Cs), German (De), Greek (El), Spanish (Es), Finnish (Fi), French (Fr), Italian (It), Portuguese (Pt) and Russian (Ru) are jointly trained. The detailed statistics and sources of the available parallel resource can be found in Table 1, where we further down-sample the corpora for the targeted languages to simulate zero-resource.

It also requires additional large amount of monolingual data to obtain the word embeddings for each language, where we use the latest Wikipedia dumps [5] for all the languages. Typically, the monolingual corpora are much larger than the parallel corpora. For validation and testing, the standard validation and testing sets are utilized for each targeted language.

[1] http://www.statmt.org/wmt16/translation-task.html
[2] https://sites.google.com/site/koreanparalleldata/
[3] http://www.statmt.org/europarl/
[4] http://opus.lingfil.uu.se/MultiUN.php (subset)
[5] https://dumps.wikimedia.org/

| | Zero-Resource Translation | | | Auxiliary High-Resource Translation | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| source | Ro | Ko | Lv | Cs | De | El | Es | Fi | Fr | It | Pt | Ru |
| corpora | WMT16[1] | KPD[2] | Europarl v8[3] | | | | | | | | | UN [4] |
| size | 612k | 97k | 638k | 645k | 1.91m | 1.23m | 1.96m | 1.92m | 2.00m | 1.90m | 1.96m | 11.7m |
| subset | 0/6k/60k | 10k | 6k | / | | | | | | | | 2.00m |

Table 1: Statistics of the available parallel resource in our experiments. All the languages are translated to English.

**Preprocessing** All the data (parallel and monolingual) have been tokenized and segmented into subword symbols using byte-pair encoding (BPE) (Sennrich et al., 2016b). We use sentences of length up to 50 subword symbols for all languages. For each language, a maximum number of $40,000$ BPE operations are learned and applied to restrict the size of the vocabulary. We concatenate the vocabularies of all source languages in the multilingual setting where special a "language marker " have been appended to each word so that there will be no embedding sharing on the surface form. Thus, we avoid sharing the representation of words that have similar surface forms though with different meaning in various languages.

**Architecture** We implement an attention-based neural machine translation model which consists of a one-layer bidirectional RNN encoder and a two-layer attention-based RNN decoder. All RNNs have 512 LSTM units (Hochreiter and Schmidhuber, 1997). Both the dimensions of the source and target embedding vectors are set to 512. The dimensionality of universal embeddings is also the same. For a fair comparison, the same architecture is also utilized for training both the vanilla and multilingual NMT systems. For multilingual experiments, $1 \sim 5$ auxiliary languages are used. When training with the universal tokens, the temperature $\tau$ (in Eq. 6) is fixed to $0.05$ for all the experiments.

**Learning** All the models are trained to maximize the log-likelihood using Adam (Kingma and Ba, 2014) optimizer for 1 million steps on the mixed dataset with a batch size of 128. The dropout rates for both the encoder and the decoder is set to 0.4. We have open-sourced an implementation of the proposed model. [6]

### 4.2 Back-Translation

We utilize back-translation (BT) (Sennrich et al., 2016a) to encourage the model to use more information of the zero-resource languages. More concretely, we build the synthetic parallel corpus by translating on monolingual data[7] with a trained translation system and use it to train a backward direction translation model. Once trained, the same operation can be used on the forward direction. Generally, BT is difficult to apply for zero resource setting since it requires a reasonably good translation system to generate good quality synthetic parallel data. Such a system may not be feasible with tiny or zero parallel data. However, it is possible to start with a trained multi-NMT model.

### 4.3 Preliminary Experiments

**Training Monolingual Embeddings** We train the monolingual embeddings using `fastText`[8] (Bojanowski et al., 2017) over the Wikipedia corpora of all the languages. The vectors are set to 300 dimensions, trained using the default setting of skip-gram . All the vectors are normalized to norm 1.

**Pre-projection** In this paper, the pre-projection requires initial word alignments (seeds) between words of each source language and the universal tokens. More precisely, for the experiments of Ro/Ko/Lv-En, we use the target language (En) as the universal tokens; `fast_align`[9] is used to automatically collect the aligned words between the source languages and English.

## 5 Results

We show our main results of multiple source languages to English with different auxiliary languages in Table 2. To have a fair comparison, we use only 6k sentences corpus for both Ro and Lv with all the settings and 10k for Ko. It is obvious that applying both the universal tokens and mixture of experts modules improve the overall translation quality for all the language pairs and the improvements are additive.

To examine the influence of auxiliary languages, we tested four sets of different combinations of auxiliary languages for Ro-En and two sets for Lv-En.

[6] https://github.com/MultiPath/NA-NMT/tree/universal_translation

[7] We used News Crawl provided by WMT16 for Ro-En.

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

[9] https://github.com/clab/fast_align

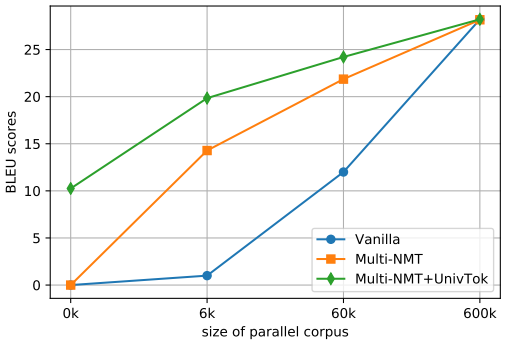


Figure 3: BLEU score vs corpus size

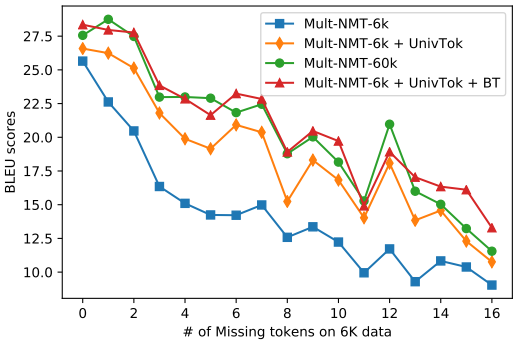


Figure 4: BLEU score vs unknown tokens

| Src | Aux | Multi | +ULR | + MoLE |
|---|---|---|---|---|
| Ro | Cs De El Fi | | 18.02 | 18.37 |
| | Cs De El Fr | | 19.48 | 19.52 |
| | De El Fi It | | 19.11 | 19.33 |
| | Es Fr It Pt | 14.83 | 20.01 | **20.51** |
| Lv | Es Fr It Pt | 7.68 | 10.86 | 11.02 |
| | Es Fr It Pt Ru | 7.88 | 12.40 | **13.16** |
| Ko | Es Fr It Pt | 2.45 | 5.49 | **6.14** |

Table 2: Scores over variant source languages (6k sentences for Ro & Lv, and 10k for Ko). "Multi" means the Multi-lingual NMT baseline.

| Models | BLEU |
|---|---|
| Vanilla | 1.21 |
| Multi-NMT | 14.94 |
| Closest Uni-Token Only | 5.83 |
| Multi-NMT + ULR + ($A$=$I$) | 18.61 |
| Multi-NMT + ULR | **20.01** |
| Multi-NMT + BT | 17.91 |
| Multi-NMT + ULR + BT | **22.35** |
| Multi-NMT + ULR + MoLE | 20.51 |
| Multi-NMT + ULR + MoLE + BT | **22.92** |
| Full data (612k) NMT | **28.34** |

Table 3: BLEU scores evaluated on test set (6k), compared with ULR and MoLE. "vanilla" is the standard NMT system trained only on Ro-En training set

It shows that Ro performs best when the auxiliary languages are all selected in the same family (Ro, Es, Fr, It and Pt are all from the Romance family of European languages) which makes sense as more knowledge can be shared across the same family. Similarly, for the experiment of Lv-En, improvements are also observed when adding Ru as additional auxiliary language as Lv and Ru share many similarities because of the geo-graphical influence even though they don't share the same alphabet.

We also tested a set of Ko-En experiments to examine the generalization capability of our approach on non-European languages while using languages of Romance family as auxiliary languages. Although the BLEU score is relatively low, the proposed methods can consistently help translating less-related low-resource languages. It is more reasonable to have similar languages as auxiliary languages.

### 5.1 Ablation Study

We perform thorough experiments to examine effectiveness of the proposed method; we do ablation study on Ro-En where all the models are trained based on the same Ro-En corpus with 6k sentences.

As shown in Table 3, it is obvious that 6k sentences of parallel corpora completely fails to train a vanilla NMT model. Using Multi-NMT with the assistance of 7.8M auxiliary language sentence pairs, Ro-En translation performance gets a substantial improvement which, however, is still limited to be usable. By contrast, the proposed ULR boosts the Multi-NMT significantly with +5.07 BLEU, which is further boosted to +7.98 BLEU when incorporating sentence-level information using both MoLE and BT. Furthermore, it is also shown that ULR works better when a trainable transformation matrix $A$ is used (4th vs 5th row in the table). Note that, although still $5 \sim 6$ BLEU scores lower than the full data ($\times 100$ large) model.

We also measure the translation quality of simply training the vanilla system while replacing each token of the Ro sentence with its closet universal token in the projected embedding space, considering we are using the target languages (En) as

the universal tokens. Although the performance is much worse than the baseline Multi-NMT, it still outperforms the vanilla model which implies the effectiveness of the embedding alignments.

**Monolingual Data** In Table. 3, we also showed the performance when incorporating the monolingual Ro corpora to help the UniNMT training in both cases with and without ULR. The back-translation improves in both cases, while the ULR still obtains the best score which indicates that the gains achieved are additive.

**Corpus Size** As shown in Fig. 3, we also evaluated our methods with varied sizes – 0k[10], 6k, 60k and 600k – of the Ro-En corpus. The vanilla NMT and the multi-lingual NMT are used as baselines. It is clear in all cases that the performance gets better when the training corpus is larger. However, the multilingual with ULR works much better with a small amount of training examples. Note that, the usage of ULR universal tokens also enables us to directly work on a "pure zero" resource translation with a shared multilingual NMT model.

**Unknown Tokens** One explanation on how ULR help the translation for almost zero resource languages is it greatly cancel out the effects of missing tokens that would cause out-of-vocabularies during testing. As in Fig. 4, the translation performance heavily drops when it has more "unknown" which cannot be found in the given 6k training set, especially for the typical multilingual NMT. Instead, these "unknown" tokens will naturally have their embeddings based on ULR projected universal tokens even if we never saw them in the training set. When we apply back-translation over the monolingual data, the performance further improves which can almost catch up with the model trained with 60k data.

### 5.2 Qualitative Analysis

**Examples** Figure 5 shows some cherry-picked examples for Ro-En. Example (a) shows how the lexical selection get enriched when introducing ULR (Lex-6K) as well as when adding Back Translation (Lex-6K-BT). Example (b) shows the effect of using romance vs non-romance languages as the supporting languages for Ro. Example (c) shows the importance of having a trainable $A$ as have been discussed; without trainable $A$ the model confuses "india" and "china" as they may have close representation in the mono-lingual embeddings.

**Visualization of MoLE** Figure 6 shows the activations along with the same source sentence with various auxiliary languages. It is clear that MoLE is effectively switching between the experts when dealing with zero-resource language words. For this particular example of Ro, we can see that the system is utilizing various auxiliary languages based on their relatedness to the source language. We can approximately rank the relatedness based of the influence of each language. For instance, the influence can be approximately ranked as $Es \approx Pt > Fr \approx It > Cs \approx El > De > Fi$, which is interestingly close to the grammatical relatedness of Ro to these languages. On the other hand, Cs has a strong influence although it does not fall in the same language family with Ro, we think this is due to the geo-graphical influence between the two languages since Cs and Ro share similar phrases and expressions. This shows that MoLE learns to utilize resources from similar languages.

### 5.3 Fine-tuning a Pre-trained Model

All the described experiments above had the low resource languages jointly trained with all the auxiliary high-resource languages, where the training of the large amount of high-resource languages can be seen as a sort of regularization. It is also common to train a model on high-resource languages first, and then fine-tune the model on a small resource language similar to transfer learning approaches (Zoph et al., 2016). However, it is not trivial to effectively fine-tune NMT models on extremely low resource data since the models easily over-fit due to over-parameterization of the neural networks.

In this experiment, we have explored the fine-tuning tasks using our approach. First, we train a Multi-NMT model (with ULR) on {Es, Fr, It, Pt}-En languages only to create a zero-shot setting for Ro-En translation. Then, we start fine-tuning the model with $6k$ parallel corpora of Ro-En, with and without ULR. As shown in Fig. 7, both models improve a lot over the baseline. With the help of ULR, we can achieve a BLEU score of around 10.7 (also shown in Fig. 3) for Ro-En translation with "zero-resource" translation. The BLEU score can further improve to almost 20 BLEU after 3 epochs of training on $6k$ sentences using ULR. This is almost 6 BLEU higher than the best score of the

[10]For 0k experiments, we used the pre-projection learned from 6k data. It is also possible to use unsupervised learned dictionary.

| | | |
|---|---|---|
| (a) | Source | situatia este putin diferita atunci cand sunt analizate separat raspunsurile barbatilor si ale femeilor . |
| | Reference | the situation is slightly different when responses are analysed separately for men and women . |
| | Mul-6k | the situation is less different when it comes to issues of men and women . |
| | Mul-60k | the situation is at least different when it is weighed up separately by men and women . |
| | Lex-6k | the situation is somewhat different when we have a separate analysis of women 's and women 's responses . |
| | Lex-6k +BT | the situation is slightly different when it is analysed separately from the responses of men and women . |

| | | |
|---|---|---|
| (b) | Source | ce nu stim este in cat timp se va intampla si cat va dura . |
| | Reference | what we don ' t know is how long all of that will take and how long it will last . |
| | Lex (Romance) | what we do not know is how long it will be and how long it will take . |
| | Lex (Non-Rom) | what we know is as long as it will happen and how it will go |

| | | |
|---|---|---|
| (c) | Source | limita de greutate pentru acestea dateaza din anii ' 80 , cand air india a inceput sa foloseasca grafice cu greutatea si inaltimea ideale . |
| | Reference | he weight limit for them dates from the ' 80s , when air india began using ideal weight and height graphics . |
| | Lex (A = I) | the weight limit for these dates back from the 1960s , when the chinese air began to use physiars with weight and the right height . |
| | Lex | the weight limit for these dates dates from the 1980s , when air india began to use the standard of its standard and height . |

Figure 5: Three sets of examples on Ro-En translation with variant settings.

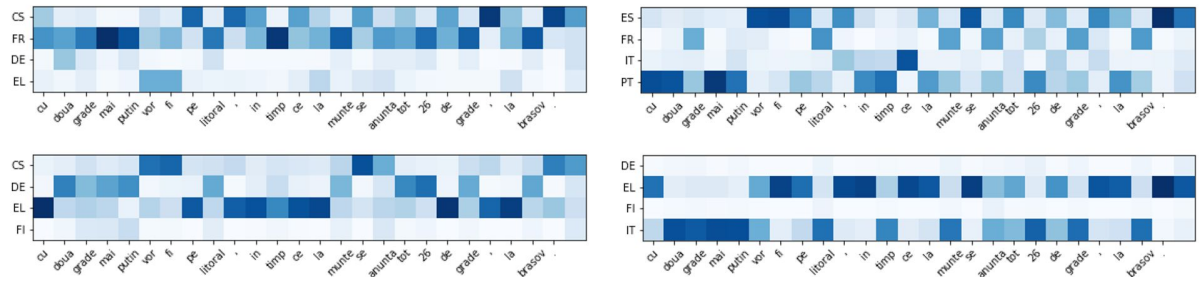


Figure 6: The activation visualization of mixture of language experts module on one randomly selected Ro source sentences trained together with different auxiliary languages. Darker color means higher activation score.

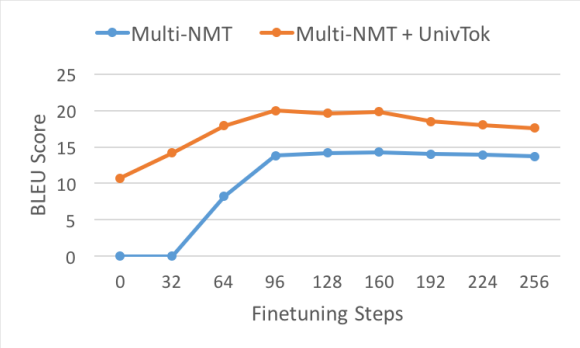


Figure 7: Performance comparison of Fine-tuning on 6K RO sentences.

baseline. It is worth noting that this fine-tuning is a very efficient process since it only takes less than 2 minutes to train for 3 epochs over such tiny amount of data. This is very appealing for practical applications where adapting a per-trained system on-line is a big advantage. As a future work, we will further investigate a better fine-tuning strategy such as meta-learning (Finn et al., 2017) using ULR.

## 6 Related Work

Multi-lingual NMT has been extensively studied in a number of papers such as Lee et al. (2017), Johnson et al. (2017), Zoph et al. (2016) and Firat et al. (2016). As we discussed, these approaches have significant limitations with zero-resource cases. Johnson et al. (2017) is more closely related to our current approach, our work is extending it to overcome the limitations with very low-resource languages and enable sharing of lexical and sentence representation across multiple languages.

Two recent related works are targeting the same problem of minimally supervised or totally unsupervised NMT. Artetxe et al. (2018) proposed a totally unsupervised approach depending on multi-lingual embedding similar to ours and dual-learning and reconstruction techniques to train the model from mono-lingual data only. Lample et al. (2018) also proposed a quite similar approach while utilizing adversarial learning.

## 7 Conclusion

In this paper, we propose a new universal machine translation approach that enables sharing resources between high resource languages and extremely low resource languages. Our approach is able to achieve 23 BLEU on Romanian-English WMT2016 using a tiny parallel corpus of 6k sentences, compared to the 18 BLEU of strong multi-lingual baseline system.

## References


Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *ACL*.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *Proceedings of International Conference on Learning Representations (ICLR)*. Vancouver, Canada.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of International Conference on Learning Representations (ICLR)*.

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.

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

Jacob Devlin. 2017. Sharp models on dull hardware: Fast and accurate neural machine translation decoding on the cpu. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, pages 2810–2815.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning (ICML)*.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of International Conference on Machine Learning (ICML)*.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation. *CoRR* abs/1803.05567.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics* 5:339–351.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M. Rush. 2016. Character-aware neural language models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. AAAI Press, AAAI'16, pages 2741–2749.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *Proceedings of International Conference on Learning Representations (ICLR)*. Vancouver, Canada.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2017. Fully character-level neural machine translation without explicit segmentation. *TACL* 5:365–378.

Minh-Thang Luong and Christopher D. Manning. 2016. Achieving open vocabulary neural machine translation with hybrid word-character models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 1054–1063.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh neural machine translation systems for wmt 16. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 371–376.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 1715–1725.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *Proceedings of International Conference on Learning Representations (ICLR)*.

Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *Proceedings of International Conference on Learning Representations (ICLR)*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*.

Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Ł. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *ArXiv e-prints* .

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1568–1575.