# Neural Interlingua for Zero-shot Machine Translation

**Jonatan Hellgren**
University of Gothenburg
gusheljot@student.gu.se

**Rode Grönkvist**
University of Gothenburg
gusgroro@student.gu.se

## Abstract

We create a multilingual encoder-decoder model for neural machine translation with an explicit interlingua layer and train it to translate between English, German and Swedish. This lets our encoders and decoders work for multiple language pairs. We achieve reasonable performance on encoding and decoding the same language as well as the bilingual translation tasks the model was trained on, but we fail to achieve zero-shot translation between a pair of languages that was not included in training.

## 1 Introduction

In recent years, the use of neural networks has become common for machine translation tasks, giving excellent performance with a variety of architectures. The most common approach is using an encoder-decoder architecture, often with attention or other enhancements. Neural machine translation (NMT) also has issues, however, one of which is the need for large amounts of parallel training data in the languages between which translation is to take place. For resource-poor languages this is a large barrier to good performance. Another issue is that a multilingual neural machine translation network using the common encoder-decoder architecture will have its training cost scale quadratically with the amount of languages it is trained on, since a unique encoder-decoder pair must be trained for each language pair.

A desired goal of multilingual NMT is thus to achieve what is known as zero-shot translation, which means direct translation of a language pair on which the network has not been trained. For instance, a network trained to translate French $\leftrightarrow$ English and German $\leftrightarrow$ English would also be able to translate French $\leftrightarrow$ German. This would reduce the time and resources needed to train a multilingual translator significantly. A similar approach that should be distinguished from zero-shot translation is pivot translation, in which the network translates into a pivot language in between the source and target languages. This approach is simpler to implement than zero-shot translation but has several drawbacks: error propagation from multiple translations, doubled computing cost during translation, and possible problems arising from the pivot language being linguistically dissimilar from the source and target languages.

In this report, we investigate an approach to zero-shot translation that uses the idea of a trained neural interlingua. Interlingua refers to some language-independent representation of the information in the text to be translated, that an NMT network theoretically encodes source text into and decodes target text from. The idea of this approach is that the meaning-content of a sentence should be the same in any language, and that an interlingual representation of the sentence could thus ideally be independent of the specific encoder used to arrive at it or the decoder used to decode it. If such a representation could be achieved, one would simply need to train one encoder and one decoder per language, into and from the interlingua. This would make zero-shot translation possible, and might also enhance non-zero-shot translation by transfer learning from other language combinations. Additionally, adding new languages to this kind of NMT system would be simple, as only one new encoder/decoder pair would need to be trained.

There is also some linguistic and philosophical interest in how good an interlingua can be at representing meaning from different languages. Are there statements that can only be made in certain languages, and that cannot be translated? If that is the case, to what extent are languages funda-

mentally different in their capacity to represent information, and what languages are more different? These are questions for which an answer could be approached by developing neural interlingua.

## 2 Related work

Though the idea of a neural interlingua is rather clear, this concept can be implemented in several different ways in practice. Although neural interlingua models are a very recent development, several approaches have been proposed.

Our work is based on a paper by Lu et al. (2018), in which the authors implement a neural interlingua by means of an intermediate LSTM layer in a normal multilingual encoder-decoder architecture. To our knowledge, this is the first work that incorporates an explicit neural interlingua for multilingual NMT, and thus also the first work where a neural interlingua is used for zero-shot translation. Their model is based on of the encoder-decoder model with attention introduced by Bahdanau et al. (2015) and adapted to the multilingual NMT context by Firat et al. (2016). Our implementation differs from that of Lu et al. in that we use GRU layers as the basis of our encoders and decoders rather than LSTM layers.

A closely related implementation is that of Vázquez et al. (2019), who use a similar encoder-decoder setup but with a fixed-size self-attention layer as a neural interlingua rather than an LSTM layer. A more recent development that achieves much better results can be found in Liao et al. (2021), who instead base their entire NMT network on a Transformer architecture, with language-specific encoders and decoders that achieve an interlingua by parameter sharing between some of the top-most layers of the language-specific encoders.

## 3 Models

We are here going to describe two MT models, one bilingual and one multilingual. We will evaluate our multilingual model by taking a look at the BLEU-score when performing a zero-shot translation from Swedish to English. The bilingual model will be trained on the translating from German to English and its BLEU-score will be used as a reference to determine how well the multilingual model performed.

### 3.1 Bilingual model

This model is a encoder-decoder model with attention and does the following computation,

$$p(y_i|y_{<i}, x) = \text{Dec}(\text{Enc}(x), y_{i-1}, h^t_{i-1}).$$

The encoder function starts with embedding the source sentence $x$,

$$B = \text{Emb}_s(x),$$

where $B$ is a $e^b \times L_x$ sized matrix where $e^b$ is the size of the embedding and $L_X$ is the length of the sentence $x$. Then the embedded sentence is passed through a Bi-directional Gated Recurrent Unit,

$$E_{.,i} = \text{Enc}(x) = \text{BiGRU}(B_{.,i}, h_{.,i-1}),$$

where $h^s_{.,i-1}$ is the $i-1$th row of the hidden units for the previous iteration and $E$ is a $e^s \times L_x$ dimensional matrix where $e^s$ is the size of the encoders output. GRU units are used instead of the LSTM units used by Lu et al. as they should be more efficient and training cost is a concern for us.

Then when we have our encoded representation of our source sentence we let the decoder function decode it into the target language,

$$p(y_i|y_{<i}, x) = \text{Dec}(I, y_{i-1}, h_{i-1})_{.,i}$$

$$= \text{softmax}(W[GRU([y_{i-1}, c_i], h_{i-1}), c_i] + \theta),$$

where $W$ and $\theta$ are the weights and bias for a final dense linear layer. The context vector $c_i$ is computed using attention ,

$$c_i = \sum_{j=1}^{L_x} \alpha_{ij} E_{.,j}$$

where $\alpha_{ij}$ is the normalized attention weights used in Bahdanau et al. (2015).

### 3.2 Multilingual model

We implemented a more simple multilingual model then most of the described models from literature, it is inspired by the model described in Lu et al. (2018). Our model consists of one interlingua component and encoder-decoder pairs for each language, each encoder includes an embedding layer. To get our model to translate from a source language to another target language we will have to use the encoder for the source language, the decoder for the target language and the interlingua component which is commonly shared

| Source | Target |
|---|---|
| Swedish | German |
| German | Swedish |
| German | English |
| English | German |
| English | English |
| German | German |
| Swedish | Swedish |

Table 1: Here we can see all the language pair we will train our model on.

for each translation. In Table 1 we can see which language pairs we are going to train our model to translate on.

When translating a sentence $x$ from the source language $s$ to a sentence $y$ in the target language $t$ our model does the following computation,

$$p(y_i|y_{<i}, x) = \text{Dec}_t(\text{Inter}(\text{Enc}_s(x)), y_{i-1}, h^t_{i-1}).$$

The encoder function takes the input sentence $x$ and outputs a encoded sentence representation, by first embedding the sentence,

$$B^s = \text{Emb}_s(x),$$

then passing the embedded sentence through a Bi-directional Gated Recurrent Unit,

$$E^s_{.,i} = \text{Enc}_s(x) = \text{BiGRU}(B^s_{.,i}, h^s_{.,i-1}),$$

where $h^s_{.,i-1}$ is the $i-1$th row of the hidden units for the previous iteration. $E^s$ is a $e^s \times L_x$ dimensional matrix where $e^s$ is the size of the encoders output and $L_x$ is the length of the source sentence.

The interlingua function takes the encoders output as input and returns a interlingua representation,

$$I_{.,i} = \text{Inter}(E^s_{.,i}) = W^I[\text{GRU}(c^I_i, h^I_{i-1})] + \theta^I,$$

where the context vector $c^I_i$ is computed using attention described in Bahdanau, $W^I$ are the weights and $\theta$ are the biases for a dense linear layer. I is a $e^i \times L_i$ sized matrix where $e^i$ is the size of the interlingua components output and $L_i$ is a fixed number equal to the maximal sentence length (in our implementation this value is 25).

Finally the decoder takes the interlingua representation and translates it to the target language,

$$p(y_i|y_{<i}, x) = \text{Dec}_t(I, y_{i-1}, h^t_{i-1})_{.,i}$$
$$= \text{softmax}(W^t[GRU([y_{i-1}, c^t_i], h^t_{i-1}), c^t_i] + \theta^t),$$

here $c^t_i$ is context vector that is computed with attention similarly as for the interlingua function.

| | Bilingual | Multilingual |
|---|---|---|
| embedding size | 128 | 128 |
| vocabulary size | 10000 | 10000 |
| batch size | 64 | 64 |
| max sentence length | 25 | 25 |
| learning rate | 2e-4 | 2e-4 |
| weight decay | 0 | 0 |
| encoder dim | 128 | 128 |
| encoder depth | 2 | 2 |
| interlingua dim | 0 | 128 |
| interlingua depth | 0 | 1 |
| decoder dim | 128 | 128 |
| decoder depth | 1 | 1 |

Table 2: In this table we can see the hyper-parameters we used for training our models.

### 3.3 Training of the models

We trained our models on data from the European Parliament collected from OPUS. During training we applied teacher forcing. We trained both models for 37 epoch using the parameters that can be found in Table 2. With bilingual translations we used $200'000$ sentences and for monolingual translations we used $10'000$ sentences, this is to give the encoder and decoders different task to mix up the learning a bit but still keeping the focus on translation between languages.

## 4 Results

We can see the results in Figure 1, as we can see our multilingual model achieved a BLEU-score of around 40-50 for monolingual translations task, around 10-15 for bilingual translation tasks and close to 0 for the zero-shot translation. When training our bilingual model the highest BLEU-score was just above 16.

## 5 Discussion

Judging by our results the model have clearly failed in creating the interlingua representation we desired, and thus the model didn't generalize well enough to get good results during zero-shot translation. However it is worth mentioning that we where still able to create encoder and decoder parts that could be used for multiple translation tasks without a major drop in performance, whereas a normal encoder-decoder NMT architecture only works for one specific language pair. So we can in fact draw a conclusion that the interlingua function does something, just not the thing we

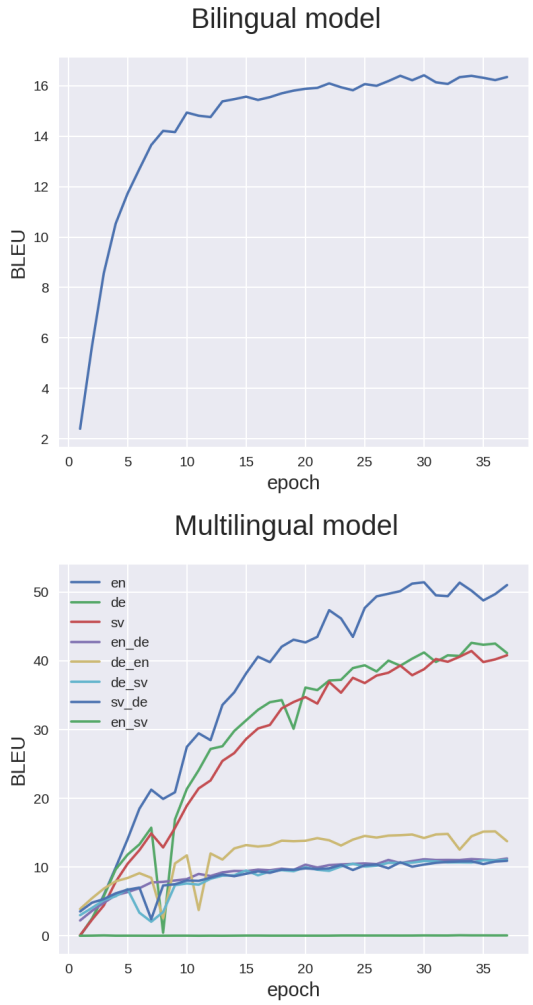


Figure 1: Here we can see the training history with respect to the BLEU score for both of our models. In the multilingual model the green line at the bottom labeled ”en_sv” are the results of zero-shot translation, and the first mentioned language is the source and the last corresponds to the target language. For the top three lines the results from monolingual translation, i.e. when the source and the target language is the same.

want it to do.

The task of creating a working interlingua that can perform zero-shot translations is difficult, this is mainly due to that we can not explicitly train the model on what we want it to be able to do. Tweaking the training loop might be required to get the model to generalize enough to create an interlingua representation. Another possibility is that simply more data is required, both in the form of more sentences and language-pairs.

We believe that the advantages of using explicit neural interlingua for NMT are so large that this approach will become much more common in large-scale NMT applications. The gains from transfer learning are a massive advantage when translating to and from low-resource languages Zero-shot interlingua NMT is already outperforming pivot translation on certain tasks and will only continue to grow more advantageous as interlingual architectures grow more sophisticated. The performance of pivot translation is by definition as good or worse than either of the bilingual translation pairs involved, while zero-shot translation can improve much beyond that, both hypothetically and, according to some recent papers, in practice. The gains from transfer learning are not limited to the multilingual setting. An interlingual representation could improve bilingual translation performance and generalizability.

## References

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. 3rd International Conference on Learning Representations, ICLR 2015 ; Conference date: 07-05-2015 Through 09-05-2015.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. https://doi.org/10.18653/v1/N16-1101 Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, California. Association for Computational Linguistics.

Junwei Liao, Yu Shi, Ming Gong, Linjun Shou, Hong Qu, and Michael Zeng. 2021. http://arxiv.org/abs/2102.06578 Improving zero-shot neural machine translation on language-specific encoders-decoders.

Yichao Lu, Phillip Keung, Faisal Ladhak, Vikas Bhardwaj, Shaonan Zhang, and Jason Sun. 2018. https://doi.org/10.18653/v1/W18-6309 A neural interlingua for multilingual machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 84–92, Brussels, Belgium. Association for Computational Linguistics.

Raúl Vázquez, Alessandro Raganato, Jörg Tiedemann, and Mathias Creutz. 2019. https://doi.org/10.18653/v1/W19-4305 Multilingual NMT with a language-independent attention bridge. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 33–39, Florence, Italy. Association for Computational Linguistics.