Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu,
Zhifeng Chen, Nikhil Thorat
`melvinp,schuster,qvl,krikun,yonghui,zhifengc,nsthorat@google.com`

Fernanda Viégas, Martin Wattenberg, Greg Corrado,
Macduff Hughes, Jeffrey Dean

Abstract

We propose a simple, elegant solution to use a single Neural Machine Translation (NMT) model to translate between multiple languages. Our solution requires no change in the model architecture from our base system but instead introduces an artificial token at the beginning of the input sentence to specify the required target language. The rest of the model, which includes encoder, decoder and attention, remains unchanged and is shared across all languages. Using a shared wordpiece vocabulary, our approach enables Multilingual NMT using a single model without any increase in parameters, which is significantly simpler than previous proposals for Multilingual NMT. Our method often improves the translation quality of all involved language pairs, even while keeping the total number of model parameters constant. On the WMT'14 benchmarks, a single multilingual model achieves comparable performance for English→French and surpasses state-of-the-art results for English→German. Similarly, a single multilingual model surpasses state-of-the-art results for French→English and German→English on WMT'14 and WMT'15 benchmarks respectively. On production corpora, multilingual models of up to twelve language pairs allow for better translation of many individual pairs. In addition to improving the translation quality of language pairs that the model was trained with, our models can also learn to perform implicit bridging between language pairs never seen explicitly during training, showing that transfer learning and zero-shot translation is possible for neural translation. Finally, we show analyses that hints at a universal interlingua representation in our models and show some interesting examples when mixing languages.

1 Introduction

Neural Machine Translation (NMT) [22, 2, 5] is an end-to-end approach to machine translation that has rapidly gained adoption in many large-scale settings [24]. Almost all such systems are built for a single language pair — so far there has not been a sufficiently simple and efficient way to handle multiple language pairs using a single model without making significant changes to the basic NMT architecture.

In this paper we introduce a simple method to translate between multiple languages using a single model, taking advantage of multilingual data to improve NMT for all languages involved. Our method requires no change to the traditional NMT model architecture. Instead, we add an artificial token to the input sequence to indicate the required target language. All other parts of the system as described in [24] — encoder, decoder, attention, and shared wordpiece vocabulary — stay exactly the same. We call our system Multilingual GNMT since it is an extension of [24]. This method has several attractive benefits:

- **Simplicity**: Since no changes are made to the architecture of the model, scaling to more languages is trivial — any new data is simply added, possibly with over- or under-sampling such that all languages are appropriately represented, and used with a new token if the target language changes. This also simplifies production deployment since it can cut down the total number of models necessary when dealing with multiple languages. Note that at Google, we support a total of over 100 languages as source and target, so theoretically 100^2 models would be necessary for the best possible translations between

all pairs, if each model could only support a single language pair. Clearly this would be problematic in a production environment.

- **Low-resource language improvements**: In a multilingual NMT model, all parameters are implicitly shared by all the language pairs being modeled. This forces the model to generalize across language boundaries during training. <mark>It is observed that when language pairs with little available data and language pairs with abundant data are mixed into a single model, translation quality on the low resource language pair is significantly improved.</mark>

- **Zero-shot translation**: A surprising benefit of modeling several language pairs in a single model is that the model implicitly learns to translate between language pairs it has never seen (zero-shot translation) — a working example of transfer learning within neural translation models. For example, a multilingual NMT model trained with Portuguese→English and English→Spanish examples can generate reasonable translations for Portuguese→Spanish although it has not seen any data for that language pair. We show that the quality of zero-shot language pairs can easily be improved with little additional data of the language pair in question.

In the remaining sections of this paper we first discuss related work and explain our multilingual system architecture in more detail. Then, we go through the different ways of merging languages on the source and target side in increasing difficulty (many-to-one, one-to-many, many-to-many), and discuss the results of a number of experiments on WMT benchmarks, as well as on some of Google's production datasets. We present results from transfer learning experiments and show how implicitly-learned bridging (zero-shot translation) performs in comparison to explicit bridging (i.e., first translating to a common language like English and then translating from that common language into the desired target language) as typically used in machine translation systems. We describe visualizations of the new system in action, which provide early evidence of shared semantic representations (interlingua) between languages. Finally we also show some interesting applications of mixing languages with examples: code-switching on the source side and weighted target language mixing, and suggest possible avenues for further exploration.

2 Related Work

Interlingual translation is a classic method in machine translation [16, 10]. Despite its distinguished history, most practical applications of machine translation have focused on individual language pairs because it was simply too difficult to build a single system that translates reliably from and to several languages.

Neural Machine Translation [22, 2, 5] is a promising end-to-end learning approach to machine translation which quickly was extended to multilingual machine translation in various ways. One early attempt is the work in [12] which proposed multilingual training in a multitask learning setting. <mark>Their model is a basic encoder-decoder network for multilingual NMT, in this case without an attention mechanism.</mark> To make proper use of multilingual data, they extend their model with multiple encoders and decoders, one for each supported source and target language. Similarly in [6], the authors modify an attention-based encoder-decoder approach to perform multilingual NMT by adding a separate decoder and attention mechanism for each target language. In [3] the authors incorporate multiple modalities other than text into the encoder-decoder framework.

Several other approaches have been proposed for multilingual training, especially for low-resource language pairs. For instance, in [26] a form of multi-source translation was proposed where the model has multiple different encoders and different attention mechanisms for each source language. However, this work requires the presence of a multi-way parallel corpus between all the languages involved, which is difficult to obtain in practice. Most closely related to our approach is [7] in which the authors propose multi-way multilingual NMT using a single shared attention mechanism but multiple encoders/decoders for each source/target language. Recently in [11] a CNN-based character-level encoder was proposed which is shared across multiple source languages. However, this approach can only perform translations into a single target language.

Our approach is related to the multitask learning framework [4]. Despite its promise, this framework has seen limited practical success in real world applications. In speech recognition, there have been many successful reports of modeling multiple languages using a single model (see [17] for an extensive reference and references therein). Multilingual language processing has also shown to be successful in domains other than translation [9, 23].

In an approach similar to ours in spirit, but with a very different purpose, the NMT framework has been extended to control the politeness level of the target translation by adding a special token to the source sentence in [20].

Our method has an additional benefit not seen in other systems: It gives the system the ability to perform zero-shot translation, meaning the system can translate from a source language to a target language without having seen explicit examples from this specific language pair during training. Zero-shot translation was the direct goal of [8]. Although they were not able to achieve this direct goal, they were able to do what they call "zero-resource" translation by using their pre-trained multi-way multilingual model and later fine-tuning it with pseudo-parallel data generated by the model.

To the best of our knowledge, our work is the first to validate the use of true multilingual translation using an end-to-end model, and is incidentally also already used in a production setting. It is also the first work to demonstrate the possibility of zero-shot translation, a successful example of transfer learning in machine translation, without any additional steps.

3 System Architecture for Multilingual Translation

The multilingual model architecture (see Figure 1) is identical to Google's Neural Machine Translation (GNMT) system [24] (with the optional addition of direct connections between encoder and decoder layers which we have used for some of our experiments, see description of Figure 1) and we refer to that paper for a detailed description.
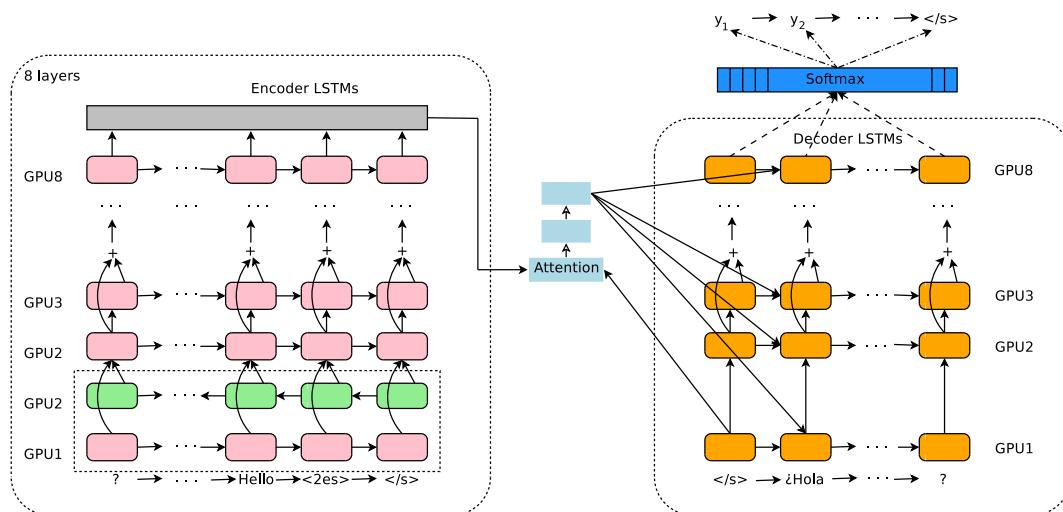

Figure 1: The model architecture of the Multilingual GNMT system. In addition to what is described in [24], our input has an artificial token to indicate the required target language. In this example, the token "<2es>" indicates that the target sentence is in Spanish, and the source sentence is reversed as a processing step. For most of our experiments we also used direct connections between the encoder and decoder although we later found out that the effect of these connections is negligible (however, once you train with those they have to be present for inference as well). The rest of the model architecture is the same as in [24].

To be able to make use of multilingual data within a single system, we propose one simple modification to the input data, which is to introduce an artificial token at the beginning of the input sentence to indicate the target language the model should translate to. For instance, consider the following English→Spanish pair of sentences:

```
Hello, how are you? -> ¿Hola como estás?
```

It will be modified to:

```
<2es> Hello, how are you? -> ¿Hola como estás?
```

to indicate that Spanish is the target language. Note that we don't specify the source language – the model will learn this automatically. Not specifying the source language has the potential disadvantage that words with the same spelling but different meaning from different source languages can be ambiguous to translate, but the advantage is that it is simpler and we can handle input with code-switching. We find that in almost all cases context provides enough language evidence to produce the correct translation.

After adding the token to the input data, we train the model with all multilingual data consisting of multiple language pairs at once, possibly after over- or undersampling some of the data to adjust for the relative ratio of the language data available. We use a shared wordpiece model [18, 21] across all the source and target data used for training, usually with 32,000 word pieces. All training is carried out as in our general NMT pipeline as described in [24] and implemented in TensorFlow [1].

In summary, this approach is the simplest and most elegant among the alternatives that we are aware of. During training and inference, we only need to add one additional token to each sentence of the source data to specify the desired target language.

4 Experiments and Results

In this section, we apply our proposed method to train multilingual models in several different configurations. Since we can have models with either single or multiple source/target languages we test three interesting cases:

- many source languages to one target language (many to one),

- one source language to many target languages (one to many), and

- many source languages to many target languages (many to many).

We will also show results and discuss benefits of bringing together many (un)related languages in a single large scale model trained on production data. Finally, we will present our findings on zero-shot translation where the model learns to translate between pairs of languages for which no explicit parallel examples existed in the training data, and show results of experiments where adding additional data improves zero-shot translation quality further.

4.1 Datasets, Training Protocols and Evaluation Metrics

We train our models on the WMT'14 English(En)→French(Fr) dataset, the WMT'14 English→German(De) data, as well as on some of Google's internal production datasets. On WMT En→Fr, the training set contains 36M sentence pairs. The WMT En→De training set contains 5M sentence pairs. In both cases, we use newstest2014 as the test sets to compare against previous work [14, 19, 25, 24]. For WMT Fr→En and De→En we use newstest2014 and newstest2015 as test sets. Despite training on WMT'14 data, which is somewhat smaller than WMT'15, we test our De→En model on newstest2015, similar to [13]. The combination of newstest2012 and newstest2013 is used as the development set.

In addition to WMT, we also evaluate the multilingual approach on some Google-internal datasets representing a wide spectrum of languages with very distinct linguistic properties: English↔Japanese(Ja), English↔Korean(Ko), English↔Spanish(Es), and English↔Portuguese(Pt). These datasets are two to three decimal orders of magnitude larger than the WMT datasets.

Our training protocols are mostly identical to those described in [24] and we refer the reader to the detailed description in that paper. We find that some multilingual models take a little more time to train than single language pair models, likely because each language pair is seen only for a fraction of the training process. Depending on the number of languages a full training can take up to 10M steps and 3 weeks to converge (on roughly 100 GPUs). We use larger batch sizes with a slightly higher initial learning rate to speed up the convergence of these models.

We evaluate our models using the standard BLEU score metric and to make our results comparable to [24, 22, 14, 25], we report tokenized BLEU score as computed by the `multi-bleu.pl` script, which can be downloaded from the public implementation of Moses.[1]

Since all language pairs do not have the same amount of training data, we explore two strategies when building multilingual models: one where we oversample the data from all language pairs to be of the same size as the largest language pair and another where we mix the data as is without any change. On our WMT models we report results using both of these strategies. For the production models, we always balance the data such that the ratios are equal. A benefit of the way we share all the components of the model is that the mini-batches can contain data from different language pairs during training and inference. This is much simpler than having separate batches per language pair and devising a scheduling strategy to handle updates during training, and simpler to deploy in production when dealing with multiple languages.

4.2 Many to One

In this section we explore having multiple source languages and a single target language — the simplest way of combining language pairs. Since there is only a single target language no additional source token is required. We perform three sets of experiments:

- The first set of experiments is on the WMT datasets, where we combine German→English and French→English to train a multilingual model. Our baselines are two single language pair models: German→English and French→English trained independently. We perform these experiments once with oversampling such that the amount of data per language pair is equal, and once without.

- The second set of experiments is on production data where we combine Japanese→English and Korean→English, with oversampling. The baselines are two single language pair models: Japanese→English and Korean→English trained independently.

- Finally, the third set of experiments is on production data where we combine Spanish→English and Portuguese→English, with oversampling. The baselines are again two single language pair models trained independently.

All of the multilingual and single language pair models have the same total number of parameters as the baseline NMT models trained on a single language pair (using 1024 nodes and 8 LSTM layers as described in [24] and a shared wordpiece model vocabulary of 32k). This is presumably unfair to the multilingual models as the number of parameters available per language pair is reduced by a factor of N compared to the single language pair models, if N is the number of language pairs combined in the multilingual model. The multilingual model also has to handle a combined vocabulary. We chose to keep the number of parameters constant for all models to simplify experimentation.

Table 1: Many to One: BLEU scores on various data sets for single language pair and multilingual models.

Model	Single	Multi	Diff
WMT German→English (oversampling)	30.43	30.59	+0.16
WMT French→English (oversampling)	35.50	35.73	+0.23
WMT German→English (no oversampling)	30.43	30.54	+0.11
WMT French→English (no oversampling)	35.50	36.77	+0.27
Prod Japanese→English	23.41	23.87	+0.46
Prod Korean→English	25.42	25.47	+0.05
Prod Spanish→English	38.00	38.73	+0.73
Prod Portuguese→English	44.40	45.19	+0.79

The results are presented in Table 1. For all experiments the multilingual models outperform the baseline single systems despite the above mentioned disadvantage with respect to the number of parameters available per language pair. One possible hypothesis explaining the gains is that the model has been shown more

---

[1] http://www.statmt.org/moses/

English data on the target side, and that the source languages belong to the same language families, so the model has learned useful generalizations.

For the WMT experiments, we obtain a maximum gain of +1.27 BLEU for French→English. Note that the results on both the WMT test sets are better than other published state-of-the-art results for a single model, to the best of our knowledge. On the production experiments, we see that the multilingual models outperform the baseline single systems by as much as +0.8 BLEU.

4.3 One to Many

In this section, we explore the application of our method when there is a single source language and multiple target languages. Here we need to prepend the input with an additional token to specify the target language. We perform three sets of experiments almost identical to the previous section except that the source and target languages have been reversed.

Table 2 summarizes the results when performing translations into multiple target languages. We see that the multilingual models are comparable to, and in some cases outperform, the baselines, but not always. We obtain a large gain of +0.9 BLEU for English→Spanish. Unlike the previous set of results, there are less significant gains in this set of experiments. This is perhaps due to the fact that the decoder has a more difficult time translating into multiple target languages which may even have different scripts e.g., Japanese and Korean, which are combined into a single shared wordpiece vocabulary. Also, as mentioned above, the multilingual models are at a disadvantage because they have the same number of parameters as the single language pair models.

Table 2: One to Many: BLEU scores on various data sets for single language pair and multilingual models.

Model	Single	Multi	Diff
WMT English→German (oversampling)	24.67	24.97	+0.30
WMT English→French (oversampling)	38.95	36.84	-2.11
WMT English→German (no oversampling)	24.67	22.61	-2.06
WMT English→French (no oversampling)	38.95	38.16	-0.79
Prod English→Japanese	23.66	23.73	+0.07
Prod English→Korean	19.75	19.58	-0.17
Prod English→Spanish	34.50	35.40	+0.90
Prod English→Portuguese	38.40	38.63	+0.23

We observe that oversampling helps the smaller language pair (En→De) at the cost of lower quality for the larger language pair (En→Fr). The model without oversampling achieves better results on the larger language compared to the smaller one as expected. We expect that there is a setting in between these two extremes that will benefit both the languages and finding that setting is the scope of future work. We also find that this effect is more prominent on smaller datasets (WMT) and much less so on our production datasets.

4.4 Many to Many

In this section, we report on experiments when there are multiple source languages and multiple target languages within a single model — the most difficult setup. Since multiple target languages are given, the input needs to be prepended with the target language token as above.

The results are presented in Table 3. We see that the multilingual production models with the same model size and vocabulary size as the single language models are quite close to the baselines. In some cases they even outperforms the baseline model while in others they are worse. We believe this effect is because we do not go through all available training data and each language pair sees only one fourth of the data seen by the baselines.

On the WMT datasets, we once again explore the impact of oversampling the smaller language pairs. We notice a similar trend to the previous section in which oversampling helps the smaller language pairs at the expense of the larger ones, while not oversampling seems to have the reverse effect.

Table 3: Many to Many: BLEU scores on various data sets for single language pair and multilingual models.

Model	Single	Multi	Diff
WMT English→German (oversampling)	24.67	24.49	-0.18
WMT English→French (oversampling)	38.95	36.23	-2.72
WMT German→English (oversampling)	30.43	29.84	-0.59
WMT French→English (oversampling)	35.50	34.89	-0.61
WMT English→German (no oversampling)	24.67	22.61	-2.06
WMT English→French (no oversampling)	38.95	38.16	-0.79
WMT German→English (no oversampling)	30.43	29.22	-1.21
WMT French→English (no oversampling)	35.50	35.93	+0.43
Prod English→Japanese	23.66	23.12	-0.54
Prod English→Korean	19.75	19.73	-0.02
Prod Japanese→English	23.41	22.86	-0.55
Prod Korean→English	25.42	24.76	-0.66
Prod English→Spanish	34.50	34.69	+0.19
Prod English→Portuguese	38.40	37.25	-1.15
Prod Portuguese→English	44.40	44.02	-0.38
Prod Spanish→English	38.00	37.65	-0.35

Although there are some significant losses in quality from training many languages jointly using a model with the same total number of parameters as the single language pair models, these models reduce the total complexity involved in training and productionization. Additionally, these multilingual models have more interesting advantages as will be discussed in more detail in the sections below.

4.5 Large Scale Experiments

This section shows the result of combining 12 production language pairs into a single multilingual model using the same number of parameters as the single language pair models. As above, the input needs to be prepended with the target language token. We oversample the examples from the smaller language pairs to balance the data as explained above.

Table 4: Large scale experiments: BLEU scores on various data sets for single language pair and multilingual models.

Model	Single	Multi	Diff
Prod English→Japanese	23.66	21.10	-2.56
Prod English→Korean	19.75	18.41	-1.34
Prod Japanese→English	23.41	21.62	-1.79
Prod Korean→English	25.42	22.87	-2.55
Prod English→Spanish	34.50	34.25	-0.25
Prod English→Portuguese	38.40	37.35	-1.05
Prod Portuguese→English	44.40	42.53	-1.87
Prod Spanish→English	38.00	36.04	-1.96
Prod English→German	26.43	23.15	-3.28
Prod English→French	35.37	34.00	-1.37
Prod German→English	31.77	31.17	-0.60
Prod French→English	36.47	34.40	-2.07

The results are summarized in Table 4. We find that the multilingual model is reasonably close to the best single models and in some cases even achieves comparable quality. It is remarkable that a single model with $278M$ parameters can do what 12 models with a total of $3.33B$ parameters would have done. The multilingual model also requires one twelfth of the training time and computing resources to converge. Another important

point is that since we only train for a little longer than the single models, the individual language pairs can see as low as one twelfth of the data in comparison to their single language pair models. Again we note that the comparison below is somewhat unfair for the multilingual model and we expect a larger model trained on all available data will likely achieve comparable or better quality than the baselines.

In summary, multilingual NMT enables us to group languages with little or no loss in quality while having the benefits of better training efficiency, smaller number of models, and easier productionization.

4.6   Zero-Shot Translation

An interesting benefit of our approach is that we can perform zero-shot translation between a language pair for which no explicit training data has been seen.

To demonstrate this we will use two multilingual models — a model trained with examples from two different language-pairs, Portuguese→English and English→Spanish (Model 1), and a model trained with examples from four different language-pairs, English↔Portuguese and English↔Spanish (Model 2). We show that both of these models can generate reasonably good quality Portuguese→Spanish translations without ever having seen Portuguese→Spanish data during training. To our knowledge this is the first demonstration of true multilingual zero-shot translation. As with the previous multilingual models, both of these models perform comparable to or even slightly better than the baseline single language pair models. Note that besides the pleasant fact that zero-shot translation works at all it has also the advantage of halving decoding speed as no explicit bridging through a third language is necessary when translating from Portuguese to Spanish.

Table 5 summarizes our results for the Portuguese→Spanish translation experiments. Rows (a) and (b) report the performance of the phrase-based machine translation (PBMT) system and the NMT system through bridging (translation from Portuguese to English and translating the resulting English sentence to Spanish). It can be seen that the NMT system outperforms the PBMT system by close to 2 BLEU points. Note that Model 1 and Model 2 can be bridged within themselves to perform Portuguese→Spanish translation. We do not report these numbers since they are similar to the performance of bridging with two individual single language pair NMT models. For comparison, we built a single NMT model on all available Portuguese→Spanish parallel sentences (see (c) in Table 5).

Table 5: Portuguese→Spanish BLEU scores using various models.

	Model	BLEU
(a)	PBMT bridged	28.99
(b)	NMT bridged	30.91
(c)	NMT Pt→Es	31.50
(d)	Model 1 (Pt→En, En→Es)	21.62
(e)	Model 2 (En↔{Es, Pt})	24.75
(f)	Model 2 + incremental training	31.77

The most interesting observation is that both Model 1 and Model 2 can perform zero-shot translation with reasonable quality (see (d) and (e)). It should be noted that Model 2 outperforms Model 1 by close to 3 BLEU points. In other words, the addition of Spanish on the source side and Portuguese on the target side helps Portuguese→Spanish zero-shot translation. We believe that this is possible only because our shared architecture enables the model to learn an interlingua between all these languages. We explore this hypothesis in more detail in Section 5.

Finally we incrementally train Model 2 with a small amount of true Portuguese→Spanish parallel data (an order of magnitude less than Table 5 (c)) and obtain the best quality. Thus we show that our multilingual model can perform zero-shot translation by implicit bridging and it also outperforms a strong NMT baseline when incrementally trained with a small amount of data. This result is especially significant for non-English low-resource language pairs where it might be easier to obtain parallel data with English but much harder to obtain parallel data for language pairs where neither the source nor the target language is English. We explore the effect of parallel data in more detail in Section 4.7.

4.7 Effect of Direct Parallel Data

In this section, we explore two ways of leveraging available parallel data to improve zero-shot translation quality, namely:

- Incrementally training the multilingual model on the additional parallel data for the zero-shot directions.

- Training a new multilingual model with all available parallel data mixed equally.

For our experiments, we use a baseline model which we call "Zero-Shot" trained on a combined parallel corpus of English↔{Belarusian, Russian, Ukrainian}. We trained a second model on the above corpus together with additional Russian↔{Belarusian, Ukrainian} data. We call this model "From-Scratch". Both models support four target languages, and are evaluated on our standard test sets. As done previously we oversample the data such that all language pairs are represented equally. Finally, we take the best checkpoint of the "Zero-Shot" model, and run incremental training on a small portion of the data used to train the "From-Scratch" model for a short period of time (3% of "Zero-Shot" model total training time). We call this model "Incremental".

As can be seen from Table 6, for the English↔X directions, all three models show comparable scores. On the Russian↔{Belarusian, Ukrainian} directions, the "Zero-Shot" model already achieves relatively high BLEU scores for all directions except one, without any explicit parallel data. This could be because these languages are linguistically related. In the "From-Scratch" column, we see that training a new model from scratch improves the zero-shot translation directions further. However, this strategy has a slightly negative effect on the English↔X directions because our oversampling strategy will reduce the frequency of the data from these directions. In the final column, we see that incremental training with direct parallel data recovers most of the BLEU score difference between the first two columns on the zero-shot language pairs. In summary, our shared architecture models the zero-shot language pairs quite well and hence enables us to easily improve their quality with a small amount of additional parallel data.

Table 6: BLEU scores for English↔{Belarusian, Russian, Ukrainian} models.

	Zero-Shot	From-Scratch	Incremental
English→Belarusian	16.85	17.03	16.99
English→Russian	22.21	22.03	21.92
English→Ukrainian	18.16	17.75	18.27
Belarusian→English	25.44	24.72	25.54
Russian→English	28.36	27.90	28.46
Ukrainian→English	28.60	28.51	28.58
Belarusian→Russian	56.53	82.50	78.63
Russian→Belarusian	58.75	72.06	70.01
Russian→Ukrainian	21.92	25.75	25.34
Ukrainian→Russian	16.73	30.53	29.92

5 Visual Analysis

The results of this paper — that training a model across multiple languages can enhance performance at the individual language level, and that zero-shot translation can be effective — raise a number of questions about how these tasks are handled inside the model, for example:

- Is the network learning some sort of shared representation, in which sentences with the same meaning are represented in similar ways regardless of language?

- Does the model operate on zero-shot translations in the same way as it treats language pairs it has been trained on?

One way to study the representations used by the network is to look at the activations of the network during translation. A starting point for investigation is the set of *attention vectors*, i.e., the activations in

the layer that connects the encoder and decoder networks (see the blue box in the center of Figure 1). A translation of a single sentence gives rise to a sequence of attention vectors. In this context, our original questions about shared representation can be studied by looking at how the vector sequences of different sentences relate. We could then ask for example:

- Do sentences cluster together depending on the source or target language?

- Or instead do sentences with similar meanings cluster, regardless of language?

We try to find answers to these questions by looking at lower-dimensional representations of internal embeddings that humans can more easily interpret.

5.1 Evidence for an Interlingua

Several trained networks indeed show strong visual evidence of a shared representation. For example, Figure 2 below was produced from a many-to-many model trained on English↔Japanese and English↔Korean. To visualize the model in action we began with a small corpus of 74 triples of semantically identical cross-language phrases. That is, each triple contained phrases in English, Japanese, and Korean with the same underlying meaning. To compile these triples, we searched a ground-truth database for English sentences which were paired with both Japanese and Korean translations.

We then applied the trained model to translate each sentence of each triple into the two other possible languages. Performing this process yielded six new sentences based on each triple, for a total of $74 * 6 = 444$ total translations with 9,978 steps corresponding to the same number of attention vectors. Since attention vectors are high-dimensional, we use the TensorFlow Embedding Projector[2] to map them into more accessible 3D space via t-SNE [15]. In the following diagrams, each point represents a single decoding step during the translation process. Points that represent steps for a given sentence are connected by line segments.

In Figure 2 is shown a global view of all 9,978 attention vectors. Points produced from the same original sentence triple are all given the same (random) color. Inspection of these clusters shows that each strand represents a single sentence, and clusters of strands generally represent a set of translations of the same underlying sentence, but with different source and target languages.

At right are two close-ups: one of an individual cluster, still coloring based on membership in the same triple, and one where we have colored by source language.

5.2 Partially Separated Representations

Not all models show such clean semantic clustering. Sometimes we observed joint embeddings in some regions of space coexisting with separate large clusters which contained many attention vectors from just one language pair.

For example, Figure 3a shows a t-SNE projection of attention vectors from a model that was trained on Portuguese→English (blue) and English→Spanish (yellow) and performing zero-shot translation from Portuguese→Spanish (red). This projection shows 153 semantically identical triples translated as described above, yielding 459 total translations. The large red region on the left primarily contains zero-shot Portuguese→Spanish translations. In other words, for a significant number of sentences, the zero-shot translation has a different embedding than the two trained translation directions. On the other hand, some zero-shot translation vectors do seem to fall near the embeddings found in other languages, as on the large region on the right.

It is natural to ask whether the large cluster of "separated" zero-shot translations has any significance. A definitive answer requires further investigation, but in this case zero-shot translations in the separated area do tend to have lower BLEU scores.

To measure the relationship between translation quality and distance between embeddings of the same semantic sentence, we first calculated BLEU scores for each translation. (This is possible since all triples of phrases were extracted from ground truth data.) Next, we needed to define a dissimilarity measure for embeddings of different sentences, accounting for the fact that two sentences might consist of different numbers of wordpieces. To do so, for a sentence of n wordpieces $w_0, w_1, \ldots, w_{n-1}$ where the ith wordpiece
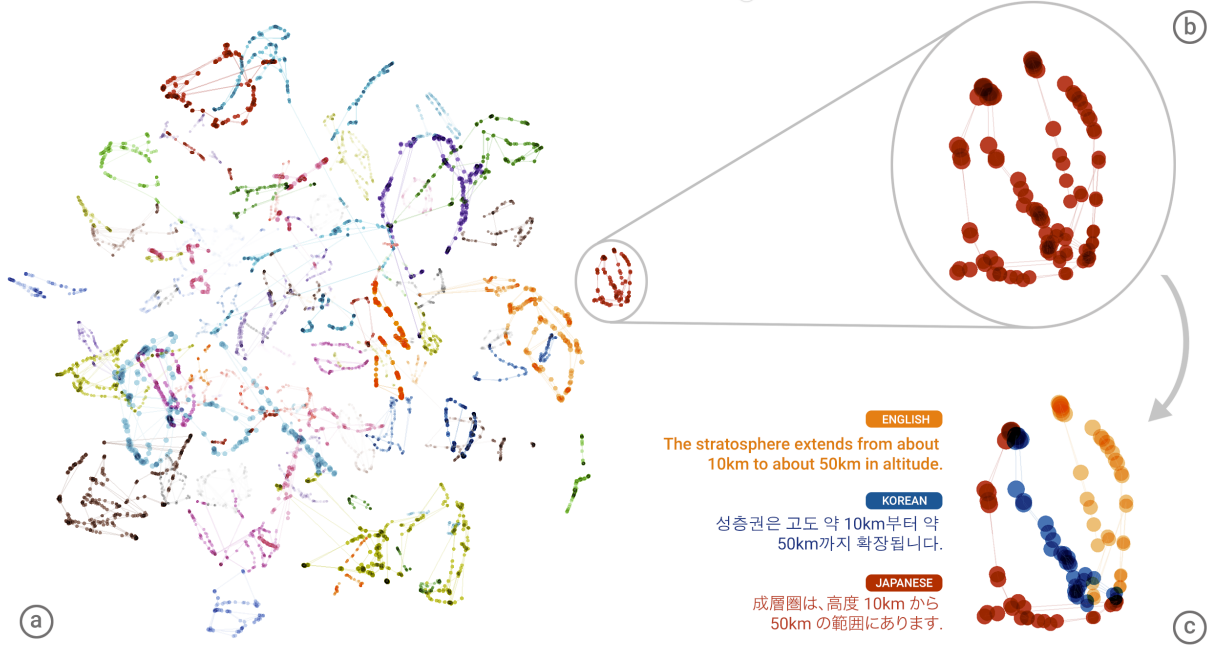

Figure 2: A t-SNE projection of the embedding of 74 semantically identical sentences translated across all 6 possible directions, yielding a total of 9,978 steps (dots in the image), from the model trained on English↔Japanese and English↔Korean examples. (a) A bird's-eye view of the embedding, coloring by the index of the semantic sentence. Well-defined clusters each having a single color are apparent. (b) A zoomed in view of one of the clusters with the same coloring. All of the sentences within this cluster are translations of "The stratosphere extends from about 10km to about 50km in altitude." (c) The same cluster colored by source language. All three source languages can be seen within this cluster.
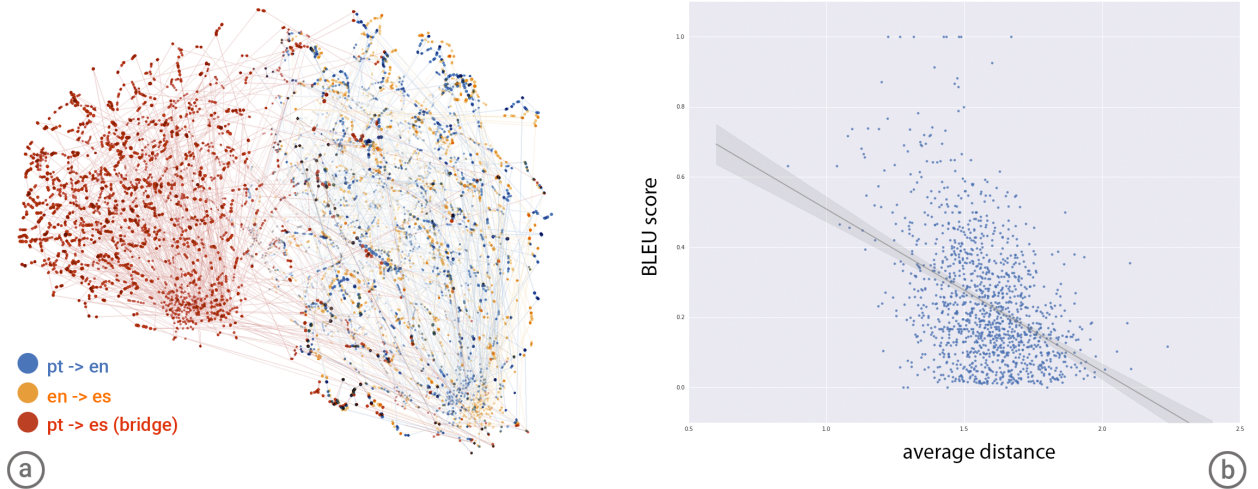

Figure 3: (a) A bird's-eye view of a t-SNE projection of an embedding of the model trained on Portuguese→English (blue) and English→Spanish (yellow) examples with a Portuguese→Spanish zero-shot bridge (red). The large red region on the left primarily contains the zero-shot Portuguese→Spanish translations. (b) A scatter plot of BLEU scores of zero-shot translations versus the average point-wise distance between the zero-shot translation and a non-bridged translation. The Pearson correlation coefficient is −0.42.

has been embedded at $y_i \in \mathbb{R}^{1024}$, we defined a curve $\gamma : [0,1] \to \mathbb{R}^{1024}$ at "control points" of the form $\frac{i}{n-1}$ by:

$$\gamma\Big(\frac{i}{n-1}\Big) = y_i$$

and use linear interpolation to define γ between these points. The dissimilarity between two curves γ_1 and γ_2, where m is the maximum number of wordpieces in both sentences, is defined by

$$dissimilarity(\gamma_1, \gamma_2) = \frac{1}{m} \sum_{i=0}^{m-1} d\Big(\gamma_1\Big(\frac{i}{m-1}\Big), \gamma_2\Big(\frac{i}{m-1}\Big)\Big)$$

From Figure 3b we can see that the value of this dissimilarity score is correlated with the quality of the zero-shot translation with a Pearson correlation coefficient of -0.42, indicating moderate correlation. An interesting area for future research is to find a more reliable correspondence between embedding geometry and model performance to predict the quality of a zero-shot translation during decoding by comparing it to the embedding of the translation through a trained language pair.

6   Mixing Languages

Having a mechanism to translate from a random source language to a single chosen target language using an additional source token made us think about what happens when languages are mixed on the source or target side. In particular, we were interested in the following two experiments:

1. Can a multilingual model successfully handle multi-language input (code-switching), when it happens in the middle of the sentence?

2. What happens when a multilingual model is triggered not with a single but two target language tokens weighted such that their weight adds up to one (the equivalent of merging the weighted embeddings of these tokens)?

The following two section discuss these experiments.

6.1   Source Language Code-Switching

In this section we show how multilingual models deal with source language code-switching. Here we show an example from a multilingual model that was trained with Japanese,Korean→English data. Using this model, mixing Japanese and Korean in the source produces in many cases correct English translations, showing the code-switching can be handled by this model, although no such code-switching samples were present in the training data. Note that the model can effectively handle the different typographic scripts since the individual characters/wordpieces are present in our wordpiece vocabulary.

- **Japanese:** 私は東京大学の学生です。 → I am a student at Tokyo University.

- **Korean:** 나는 도쿄 대학의 학생입니다. → I am a student at Tokyo University.

- **Mixed Japanese/Korean:** 私は東京大学학생입니다. → I am a student of Tokyo University.

Interestingly, the translation for the mixed-language input differs slightly from both of the single source language translations. In practice, it is not too hard to find examples where code-switching in the input does not result in good outputs; in some cases the model will simply copy parts of the source sentence instead of translating it.

6.2   Weighted Target Language Selection

In this section we test what happens when we mix target languages. We take a multilingual model trained with multiple target languages, for example, English→{Japanese, Korean}. Then instead of feeding the embedding vector for "<2ja>" to the bottom layer of the encoder LSTM, we feed a linear combination

$(1 - w)<2\text{ja}> + w<2\text{ko}>$. Clearly, for $w = 0$ the model should produce Japanese, for $w = 1$ it should produce Korean, but what happens in between?

One expectation could be that the model will output some sort of intermediate language ("Japarean"), but the results turn out to be less surprising. Most of the time the output just switches from one language to another around $w = 0.5$. In some cases, for intermediate values of w the model switches languages mid-sentence.

A possible explanation for this behavior is that the target language model, implicitly learned by the decoder LSTM, may make it very hard to mix words from different languages, especially when these languages use different scripts. In addition, since the token which defines the requested target language is placed at the beginning of the sentence, the further the decoder progresses, the less likely it is to put attention on this token, and instead the choice of language is determined by previously generated target words.

Table 7 shows examples of mixed target language using three different multilingual models. It is interesting that in the first example (Russian/Belarusian) the model switches from Russian to Ukrainian (underlined) as target language first before finally switching to Belarusian. In the second example (Japanese/Korean), we observe an even more interesting transition from Japanese to Korean, where the model gradually changes the grammar from Japanese to Korean. At $w_{ko} = 0.58$, the model translates the source sentence into a mix of Japanese and Korean at the beginning of the target sentence. At $w_{ko} = 0.60$, the source sentence is translated into full Korean, where all of the source words are captured, however, the ordering of the words does not look natural. Interestingly, when the w_{ko} is increased up to 0.7, the model starts to translate the source sentence into a Korean sentence that sounds more natural.[3]

Table 7: Several examples of gradually mixing target languages in multilingual models.

Russian/Belarusian:	I wonder what they'll do next!
$w_{be} = 0.00$	Интересно, что они сделают дальше!
$w_{be} = 0.20$	Интересно, что они сделают дальше!
$w_{be} = 0.30$	Цікаво, что они будут делать дальше!
$w_{be} = 0.44$	Цікаво, що вони будуть робити далі!
$w_{be} = 0.46$	Цікаво, що вони будуть робити далі!
$w_{be} = 0.48$	Цікаво, што яны зробяць далей!
$w_{be} = 0.50$	Цікава, што яны будуць рабіць далей!
$w_{be} = 1.00$	Цікава, што яны будуць рабіць далей!

Japanese/Korean:	I must be getting somewhere near the centre of the earth.
$w_{ko} = 0.00$	私は地球の中心の近くにどこかに行っているに違いない。
$w_{ko} = 0.40$	私は地球の中心近くのどこかに着いているに違いない。
$w_{ko} = 0.56$	私は地球の中心の近くのどこかになっているに違いない。
$w_{ko} = 0.58$	私は지구の中心의가까이에어딘가에도착하고있어야한다。
$w_{ko} = 0.60$	나는지구의센터의가까이에어딘가에도착하고있어야한다。
$w_{ko} = 0.70$	나는지구의중심근처어딘가에도착해야합니다。
$w_{ko} = 0.90$	나는어딘가지구의중심근처에도착해야합니다。
$w_{ko} = 1.00$	나는어딘가지구의중심근처에도착해야합니다。

Spanish/Portuguese:	Here the other guinea-pig cheered, and was suppressed.
$w_{pt} = 0.00$	Aquí el otro conejillo de indias animó, y fue suprimido.
$w_{pt} = 0.30$	Aquí el otro conejillo de indias animó, y fue suprimido.
$w_{pt} = 0.40$	Aquí, o outro porquinho-da-índia alegrou, e foi suprimido.
$w_{pt} = 0.42$	Aqui o outro porquinho-da-índia alegrou, e foi suprimido.
$w_{pt} = 0.70$	Aqui o outro porquinho-da-índia alegrou, e foi suprimido.
$w_{pt} = 0.80$	Aqui a outra cobaia animou, e foi suprimida.
$w_{pt} = 1.00$	Aqui a outra cobaia animou, e foi suprimida.

---

[3]The Korean translation does not contain spaces and uses '。' as punctuation symbol, and these are all artifacts of applying a Japanese postprocessor.

7   Conclusion

We present a simple and elegant solution to multilingual NMT. We show that we can train multilingual NMT models that can be used to translate between a number of different languages using a single model where all parameters are shared, which as a positive side-effect also improves the translation quality of low-resource languages in the mix. We also show that zero-shot translation without explicit bridging is possible using these models, which is the first time to our knowledge that a form of true transfer learning has been shown to work for machine translation. To explicitly improve the zero-shot translation quality, we explore two different ways of adding available parallel data and find that small additional amounts are sufficient to reach satisfactory results. In our largest experiment we merge 12 language pairs into a single model of the same size as each single language pair model, and achieve only slightly lower translation quality as for the single language pair baselines despite the drastically reduced amount of modeling capacity per language in the multilingual model. Visual interpretation of the results shows that these models learn a form of interlingua representation for the multilingual model between all involved language pairs. The simple architecture makes it possible to mix languages on the source or target side to yield some interesting translation examples. Our approach has been shown to work reliably in a Google-scale production setting and enables us to scale to a large number of languages quickly.

Acknowledgements

References

[1] ABADI, M., BARHAM, P., CHEN, J., CHEN, Z., DAVIS, A., DEAN, J., DEVIN, M., GHEMAWAT, S., IRVING, G., ISARD, M., KUDLUR, M., LEVENBERG, J., MONGA, R., MOORE, S., MURRAY, D. G., STEINER, B., TUCKER, P., VASUDEVAN, V., WARDEN, P., WICKE, M., YU, Y., AND ZHENG, X. Tensorflow: A system for large-scale machine learning. Tech. rep., Google Brain, 2016. arXiv preprint.

[2] BAHDANAU, D., CHO, K., AND BENGIO, Y. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations* (2015).

[3] CAGLAYAN, O., ARANSA, W., WANG, Y., MASANA, M., GARCÍA-MARTÍNEZ, M., BOUGARES, F., BARRAULT, L., AND VAN DE WEIJER, J. Does multimodality help human and machine for translation and image captioning? In *Proceedings of the First Conference on Machine Translation* (Berlin, Germany, August 2016), Association for Computational Linguistics, pp. 627–633.

[4] CARUANA, R. Multitask learning. In *Learning to learn.* Springer, 1998, pp. 95–133.

[5] CHO, K., VAN MERRIENBOER, B., GÜLÇEHRE, Ç., BAHDANAU, D., BOUGARES, F., SCHWENK, H., AND BENGIO, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR abs/1406.1078* (2014).

[6] DONG, D., WU, H., HE, W., YU, D., AND WANG, H. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers* (2015), pp. 1723–1732.

[7] FIRAT, O., CHO, K., AND BENGIO, Y. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016* (2016), pp. 866–875.

[8] FIRAT, O., SANKARAN, B., AL-ONAIZAN, Y., VURAL, F. T. Y., AND CHO, K. Zero-resource translation with multi-lingual neural machine translation. *arXiv preprint arXiv:1606.04164* (2016).

[9] GILLICK, D., BRUNK, C., VINYALS, O., AND SUBRAMANYA, A. Multilingual language processing from bytes. *CoRR abs/1512.00103* (2015).

[10] HUTCHINS, W. J., AND SOMERS, H. L. *An introduction to machine translation*, vol. 362. Academic Press London, 1992.

[11] JASON, L., CHO, K., AND HOFMANN, T. Fully character-level neural machine translation without explicit segmentation. *arXiv preprint arXiv:1610.03017* (2016).

[12] LUONG, M.-T., LE, Q. V., SUTSKEVER, I., VINYALS, O., AND KAISER, L. Multi-task sequence to sequence learning. In *International Conference on Learning Representations* (2015).

[13] LUONG, M.-T., PHAM, H., AND MANNING, C. D. Effective approaches to attention-based neural machine translation. In *Conference on Empirical Methods in Natural Language Processing* (2015).

[14] LUONG, M.-T., SUTSKEVER, I., LE, Q. V., VINYALS, O., AND ZAREMBA, W. Addressing the rare word problem in neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (2015).

[15] MAATEN, L. V. D., AND HINTON, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research 9* (2008).

[16] RICHENS, R. H. Interlingual machine translation. *The Computer Journal 1*, 3 (1958), 144–147.

[17] SCHULTZ, T., AND KIRCHHOFF, K. *Multilingual speech processing*. Elsevier Academic Press, Amsterdam, Boston, Paris, 2006.

[18] SCHUSTER, M., AND NAKAJIMA, K. Japanese and Korean voice search. *2012 IEEE International Conference on Acoustics, Speech and Signal Processing* (2012).

[19] SÉBASTIEN, J., KYUNGHYUN, C., MEMISEVIC, R., AND BENGIO, Y. On using very large target vocabulary for neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (2015).

[20] SENNRICH, R., HADDOW, B., AND BIRCH, A. Controlling politeness in neural machine translation via side constraints. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016* (2016), pp. 35–40.

[21] SENNRICH, R., HADDOW, B., AND BIRCH, A. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (2016).

[22] SUTSKEVER, I., VINYALS, O., AND LE, Q. V. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems* (2014), pp. 3104–3112.

[23] TSVETKOV, Y., SITARAM, S., FARUQUI, M., LAMPLE, G., LITTELL, P., MORTENSEN, D. R., BLACK, A. W., LEVIN, L. S., AND DYER, C. Polyglot neural language models: A case study in cross-lingual phonetic representation learning. *CoRR abs/1605.03832* (2016).

[24] WU, Y., SCHUSTER, M., CHEN, Z., LE, Q. V., NOROUZI, M., MACHEREY, W., KRIKUN, M., CAO, Y., GAO, Q., MACHEREY, K., KLINGNER, J., SHAH, A., JOHNSON, M., LIU, X., ŁUKASZ KAISER, GOUWS, S., KATO, Y., KUDO, T., KAZAWA, H., STEVENS, K., KURIAN, G., PATIL, N., WANG, W., YOUNG, C., SMITH, J., RIESA, J., RUDNICK, A., VINYALS, O., CORRADO, G., HUGHES, M., AND DEAN, J. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144* (2016).

[25] Zhou, J., Cao, Y., Wang, X., Li, P., and Xu, W. Deep recurrent models with fast-forward connections for neural machine translation. *CoRR abs/1606.04199* (2016).

[26] Zoph, B., and Knight, K. Multi-source neural translation. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016* (2016), pp. 30–34.