Details of the paper

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

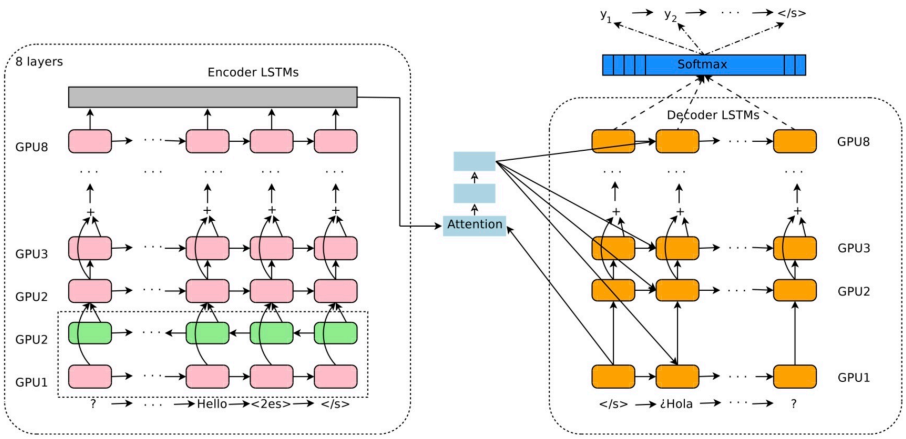

Figure 1: The model architecture of the Multilingual GNMT system. In addition to what is described in [29], our input has an artificial token to indicate the required target language. In this example, the token "<2es>" indicates that the target sentence is in Spanish, and the source sentence is reversed 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 [29].

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 sections 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 $(1-w)$<2ja>$+w$<2ko>. 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 8: 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는지구의中心의가까이에어딘가에도착하고있어야한다。
$w_{ko}=0.60$	나는지구의센터의가까이에어딘가에도착하고있어야한다。
$w_{ko}=0.70$	나는지구의중심근처어딘가에도착해야합니다。
$w_{ko}=0.90$	나는어딘가지구의중심근처에도착해야합니다。
$w_{ko}=1.00$	나는어딘가지구의중심근처에도착해야합니다。

Table 8 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]