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

Hanock Kwak

2016-11-22 (Tue.)

BI Lab

Seoul National University

Abstract

- Simple, elegant solution to translate between multiple languages.
- Introduces an artificial token at the beginning of the input sentence to specify the required target language.
  - The rest of the model is shared across all languages.
- Single multilingual model surpasses state-of-the-art results on WMT’14 and WMT’15 benchmarks.
- Transfer learning and zero-shot translation is possible.

Johnson, Melvin, et al. "Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation." *arXiv preprint arXiv:1611.04558*(2016).

Key Features

- Simplicity
  - The model is same for all languages.
  - Any new data is simply added.
- Low-resource language improvements
  - All parameters are implicitly shared by all the language pairs.
  - This forces the model to generalize across language boundaries during training.
- Zero-shot translation
  - The model implicitly learns to translate between language pairs it has never seen.
  - ex) Train Portuguese→English and English→Spanish
    - Then it can generate Portuguese→Spanish. ☺

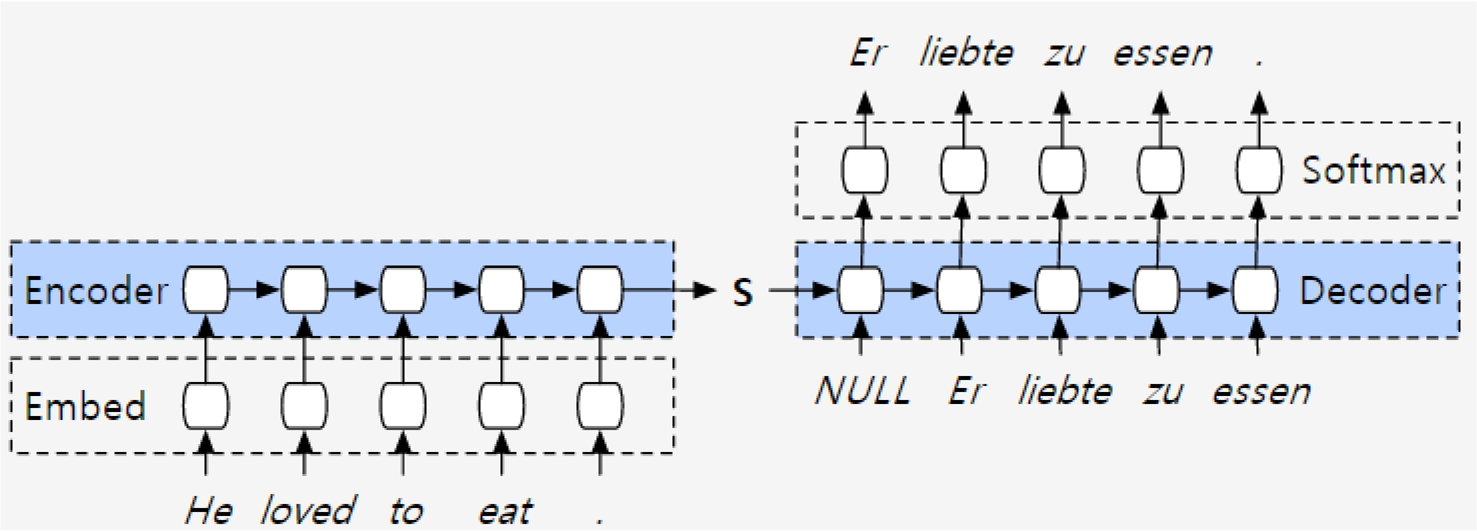
Evolution of Neural Translation Machine

- We'll start with a traditional encoder decoder machine translation model and keep evolving it until it matches GNMT

http://smerity.com/articles/2016/google_nmt_arch.html

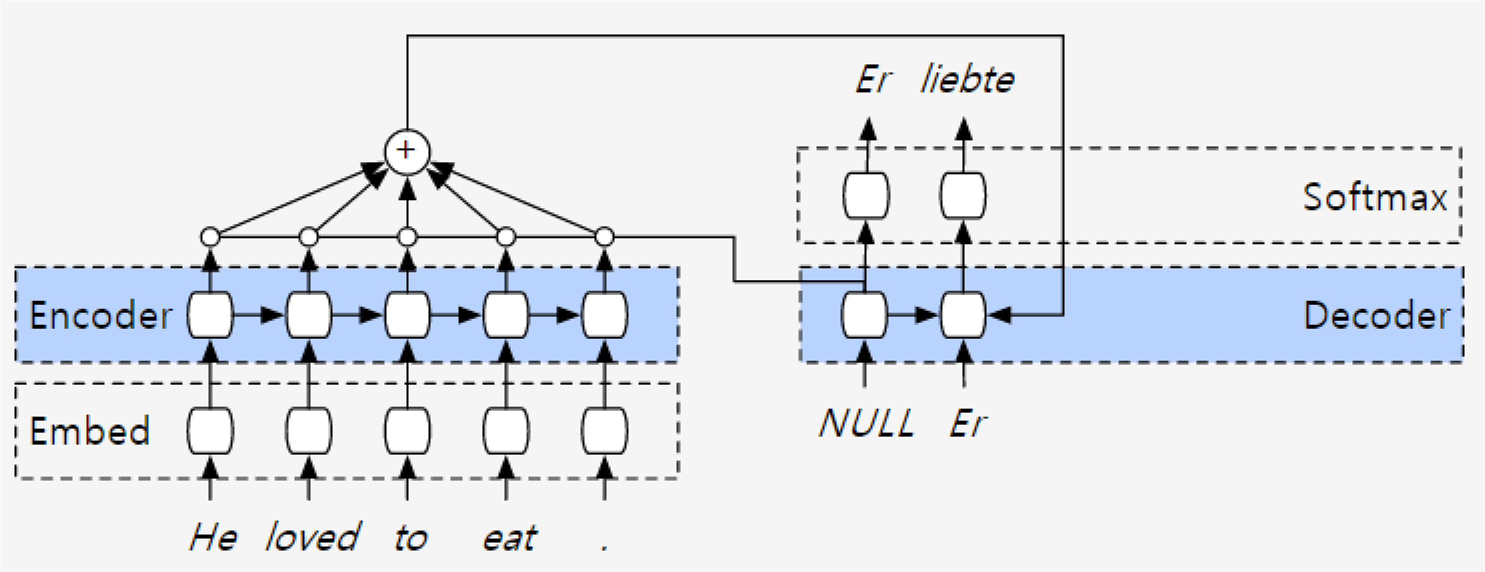
V1: Encoder-decoder

- The encoder spits out a hidden state.
- This hidden state is then supplied to the decoder, which generates the sentence in language B



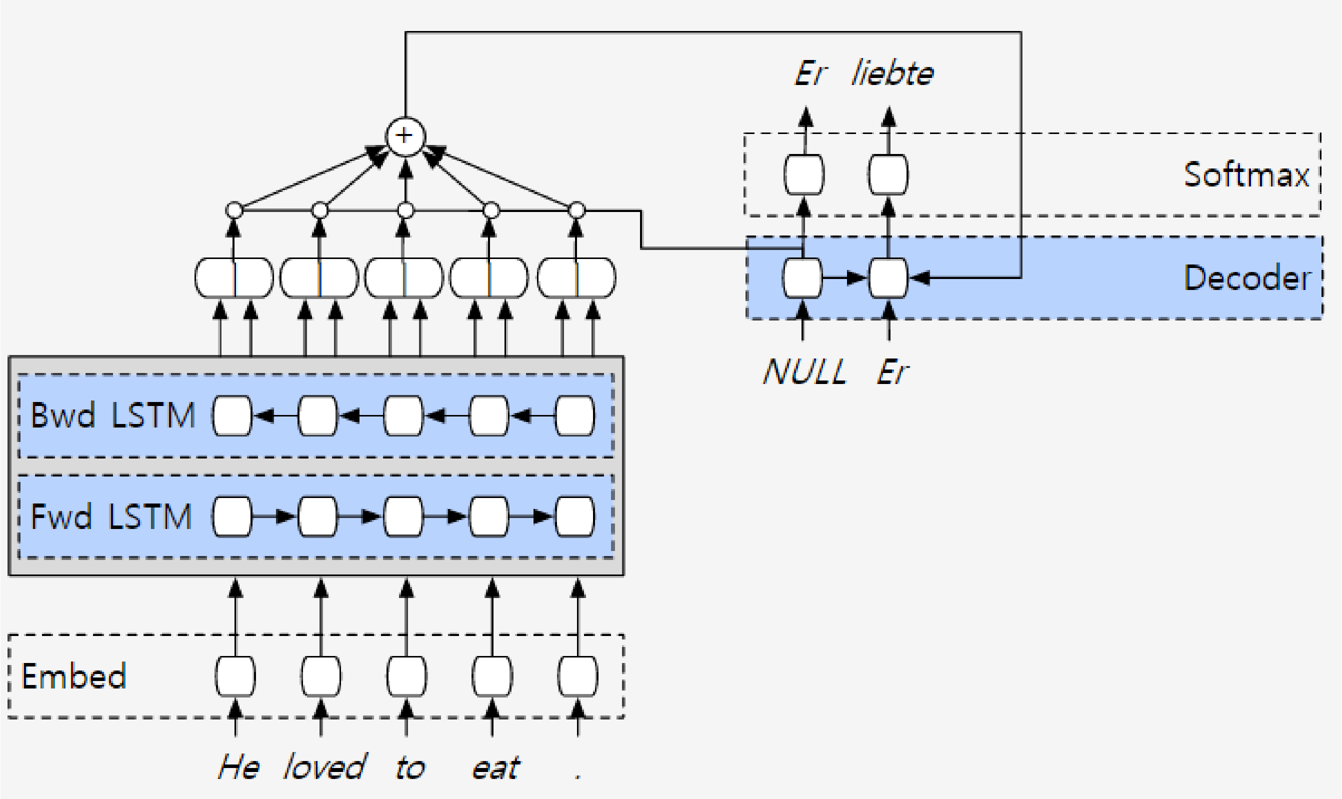
V2: Attention based encoder-decoder

- The encoder query each output asking how relevant they are to the current computation on the decoder side

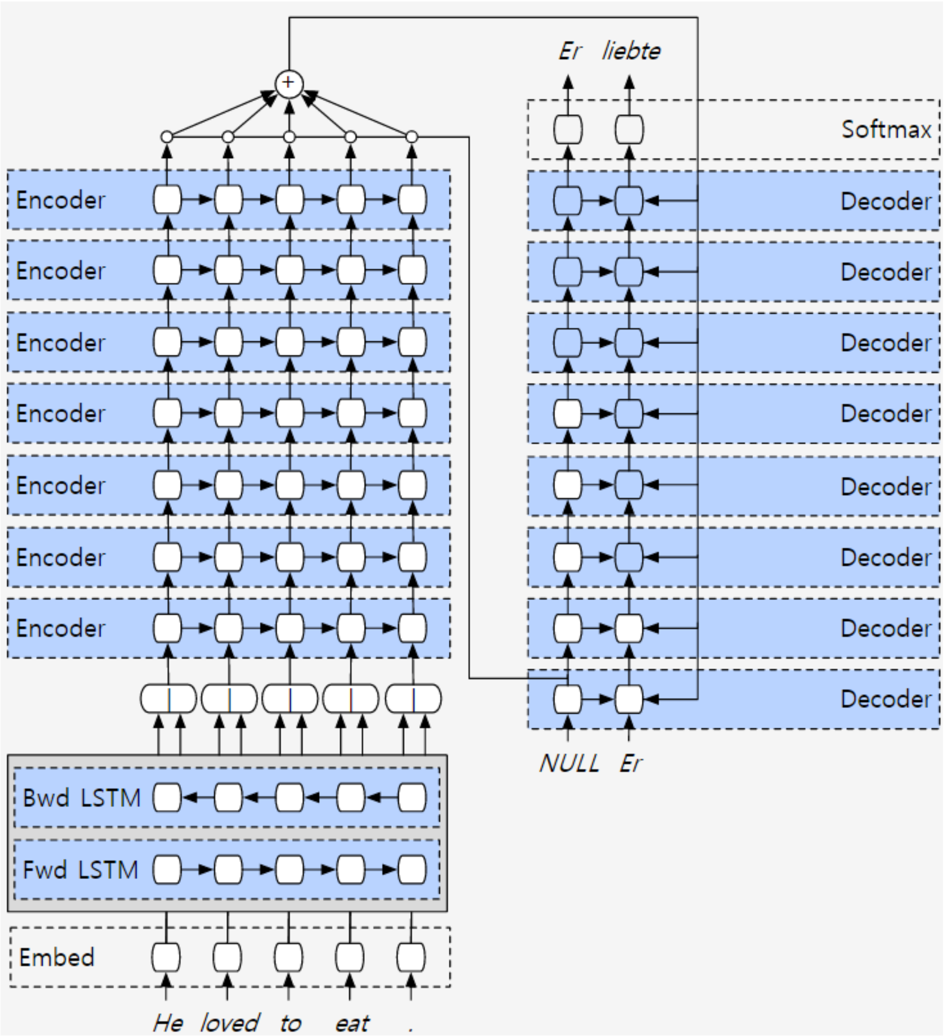


V3: Bi-directional encoder layer

- We would like the annotation of each word to summarize not only the preceding words, but also the following words

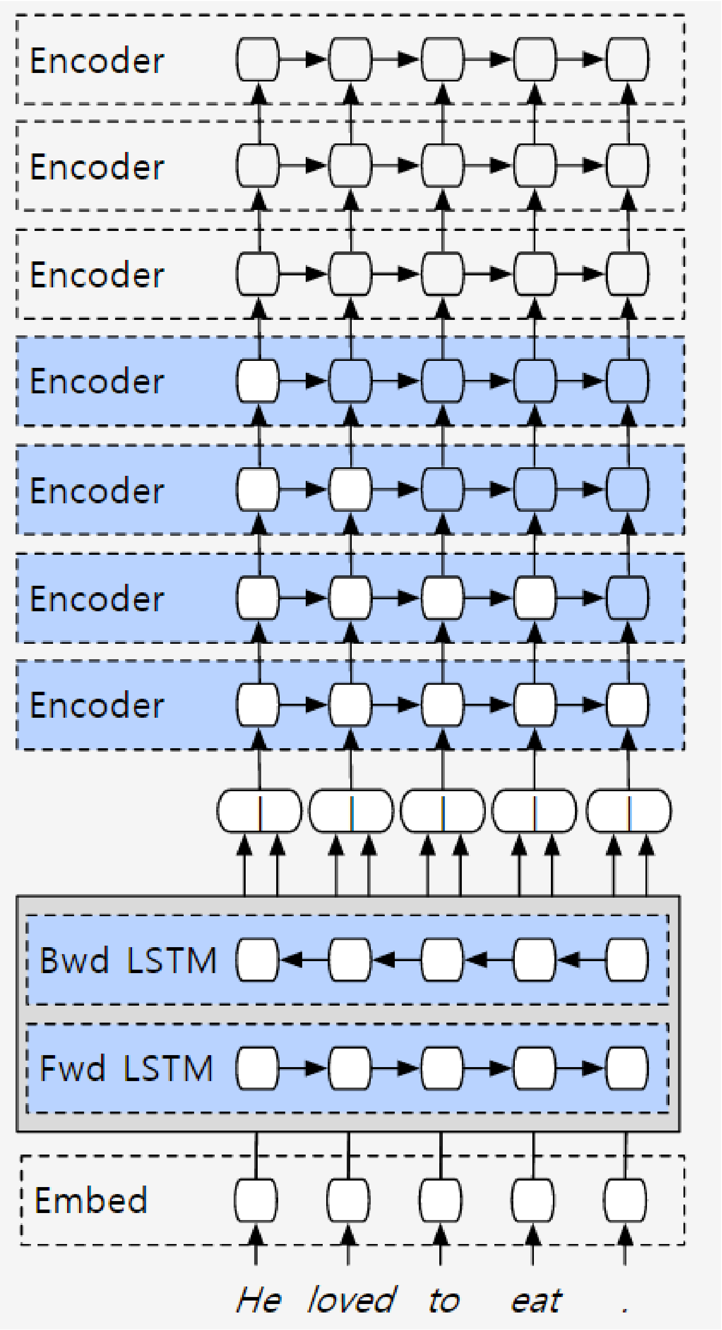


V4: "The deep is for deep learning"



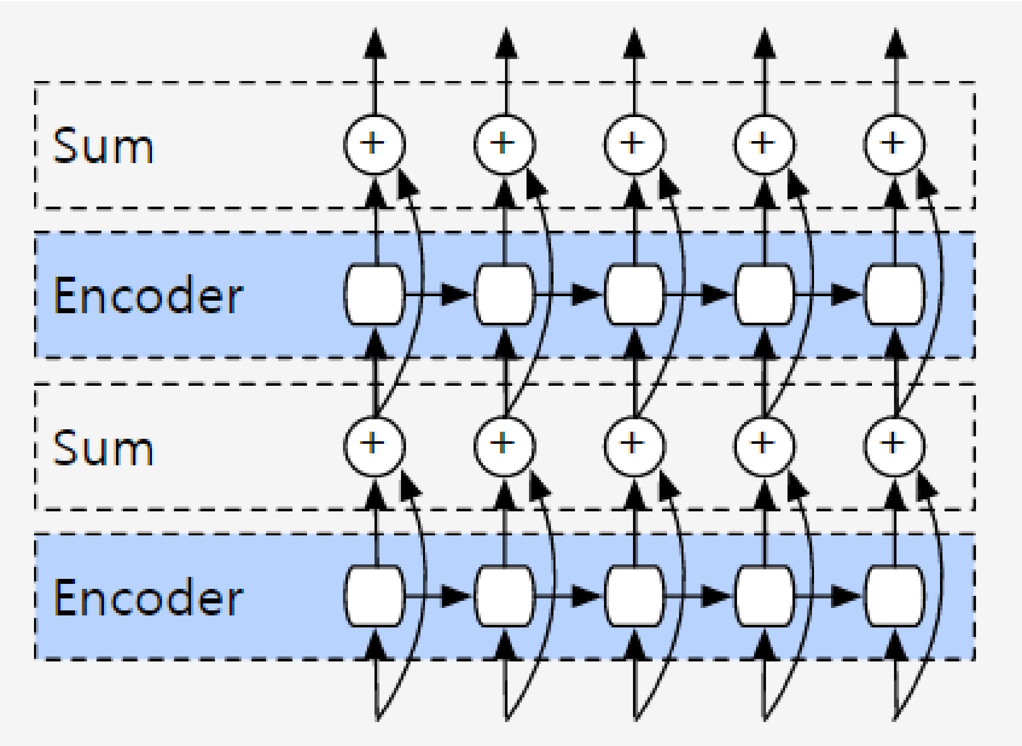
V5: Parallelization

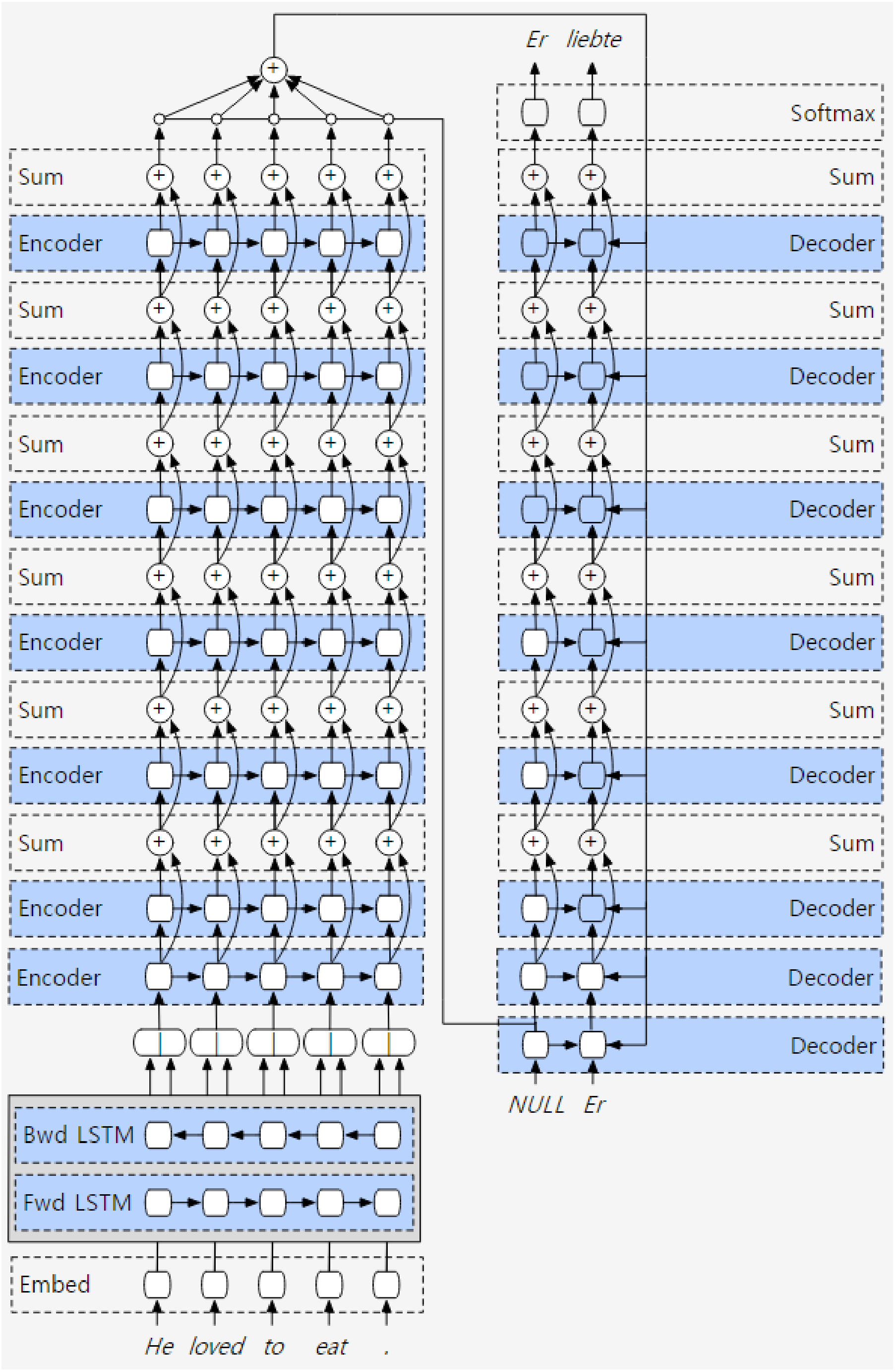
- To begin computation at one of the nodes, all of the nodes pointing toward you must already have been computed.
- A layer $i + 1$ can start its computation before layer i is fully finished.



V6: Residuals are the new hotness

- One solution for vanishing gradients is residual networks.
- The idea of a layer computing an identity function

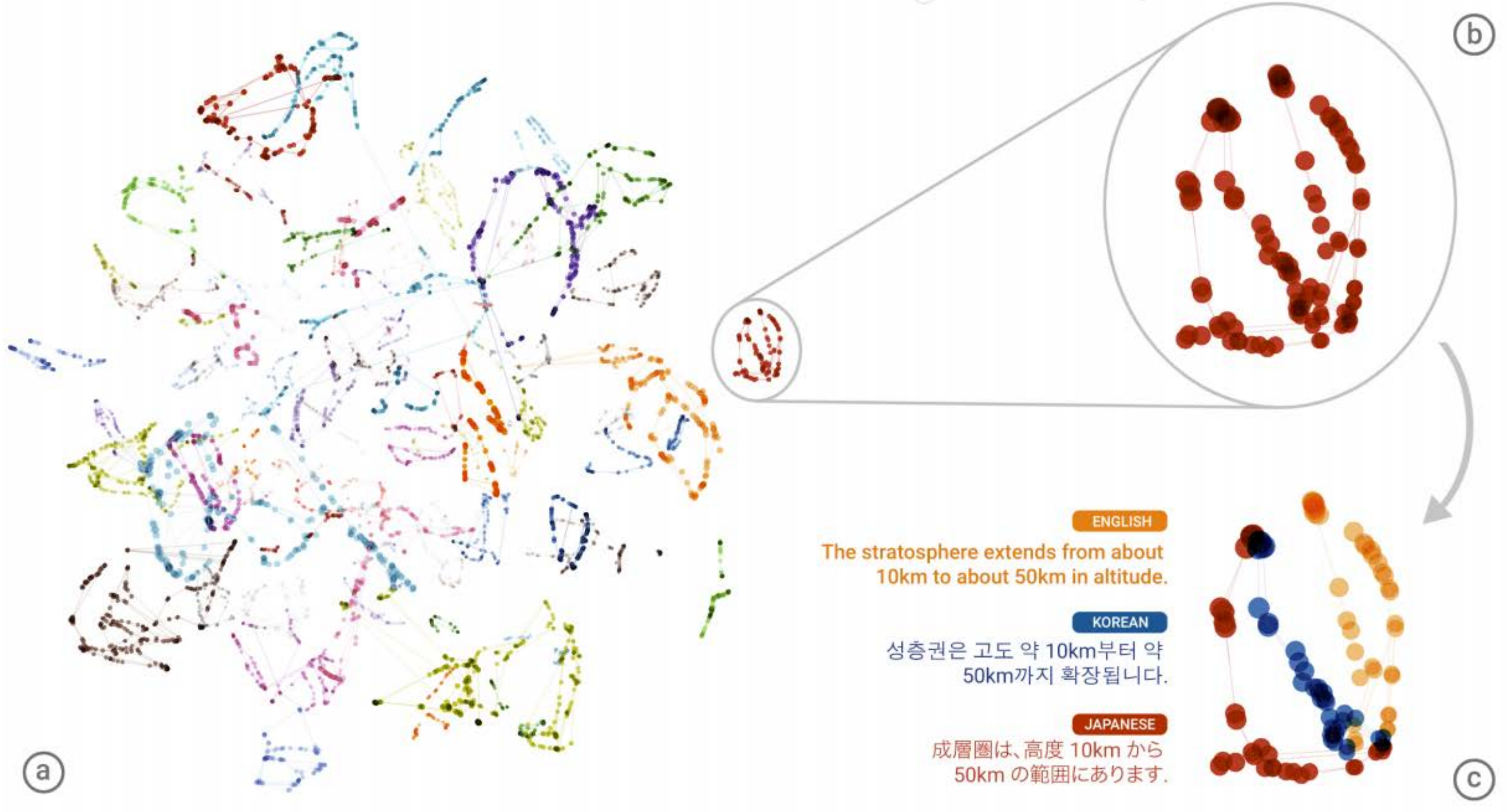




Er
liebte
Softmax
Sum
Encoder
Decoder
Bwd LSTM
Fwd LSTM
Embed
He
loved
to
eat
.
NULL
Er

Visualization

- A t-SNE projection of the embedding of 74 semantically identical sentences translated across all 6 possible directions



Source Language Code-Switching

- Mixing Japanese and Korean in the source produces in many cases correct English translations

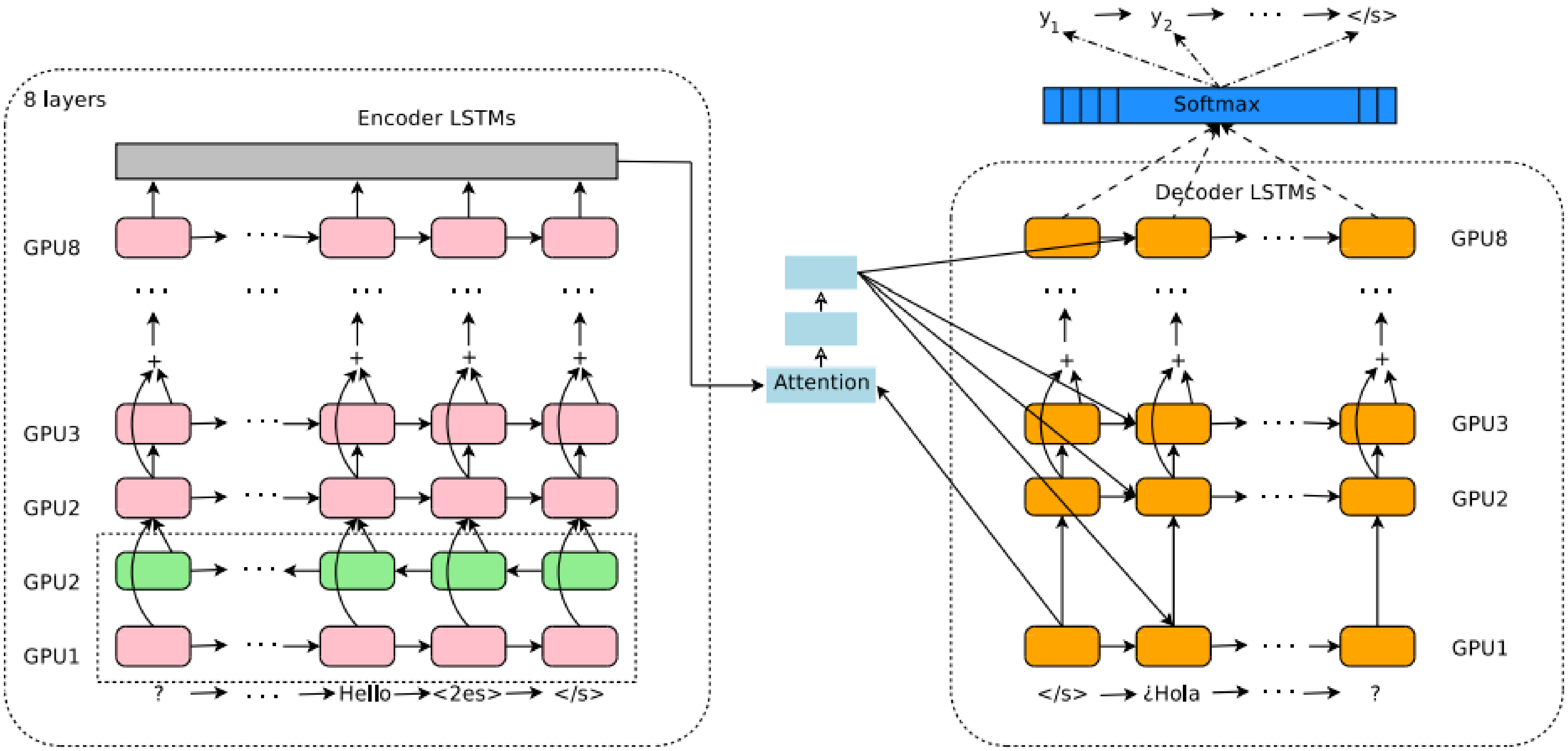
- **Japanese:** 私は東京大学の学生です。 → I am a student at Tokyo University.
- **Korean:** 나는 도쿄 대학의 학생입니다. → I am a student at Tokyo University.
- **Mixed Japanese/Korean:** 私は東京大学학생입니다. → I am a student of Tokyo University.

Weighted Target Language Selection

- We test what happens when we mix target languages.

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。

Big Picture



8 layers
Encoder LSTMs
GPU8
GPU3
GPU2
GPU2
GPU1
? → . . . → Hello → <2es> → </s>
Attention
y_1 → y_2 → . . . → </s>
Softmax
Decoder LSTMs
GPU8
GPU3
GPU2
GPU1
</s> → ¿Hola → . . . → ?