Georgia Tech

Neural Machine Translation with Attention

Presenter: Bin Xie

CREATING THE NEXT®

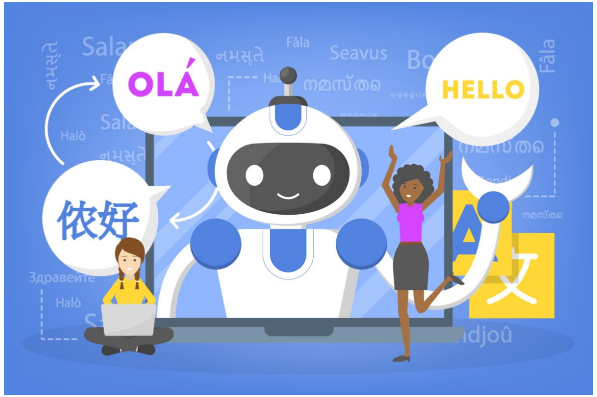
Background

Types of Machine Translation:

- Rule-Based Machine Translation (RBMT): Use a bilingual dictionary to map words
- Statistical Machine Translation (SMT): Use Bayes Rule to learn model based on parallel corpus
- Neural Machine Translation (NMT): Use a single neural network to translate

Motivation:

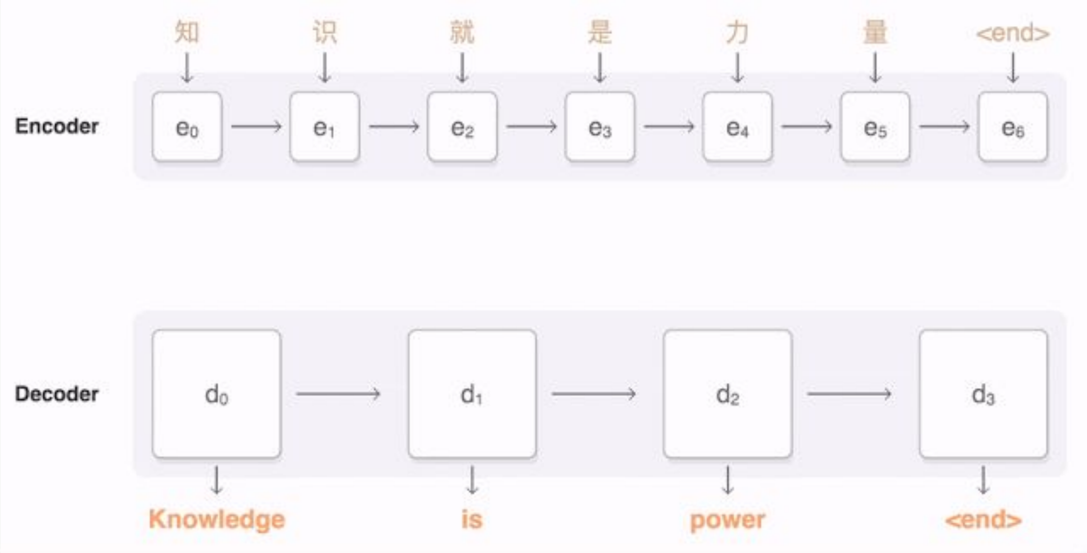
- We learned SMT in class and used RBMT in hw6
- Google Translator is based on SMT and NMT.
- Deep learning techniques like seq2seq model and attention mechanism are successful in NMT.



NMT Overview

Encoder-Decoder architecture

- Encoder: reads the source sentence build a "thought" vector
- Decoder: processes the sentence vector to emit a translation



Data Introduction

122936 Spanish to English sentence pairs from Tatoeba Corpus

Format: English + TAB + Spanish + TAB + Attribution

Examples:

I've seen them.	Los vi.	CC-BY 2.0 (France) Attribution: tatoeba.org #2248417 (CK)...
If only I knew!	¡Ojalá lo supiera!	CC-BY 2.0 (France) Attribution: tatoeba.org #276985 (CM)...
Is Tom at home?	¿Está Tom en casa?	CC-BY 2.0 (France) Attribution: tatoeba.org #2262070 (CK)...

Feature Construction

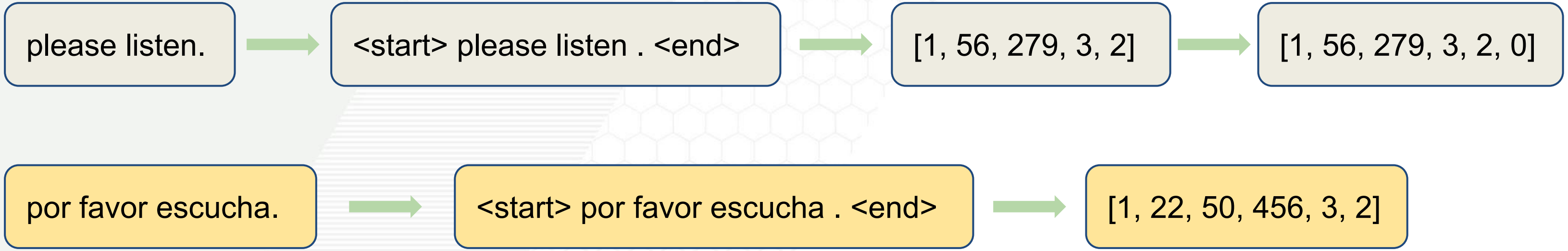
Prepare the data:

1. Add a start and end token to each sentence.

2. Remove special characters.

3. Create a word index and reverse word index

4. Pad each sentence to a maximum length.



please listen. → <start> please listen . <end> → [1, 56, 279, 3, 2] → [1, 56, 279, 3, 2, 0]

por favor escucha. → <start> por favor escucha . <end> → [1, 22, 50, 456, 3, 2]

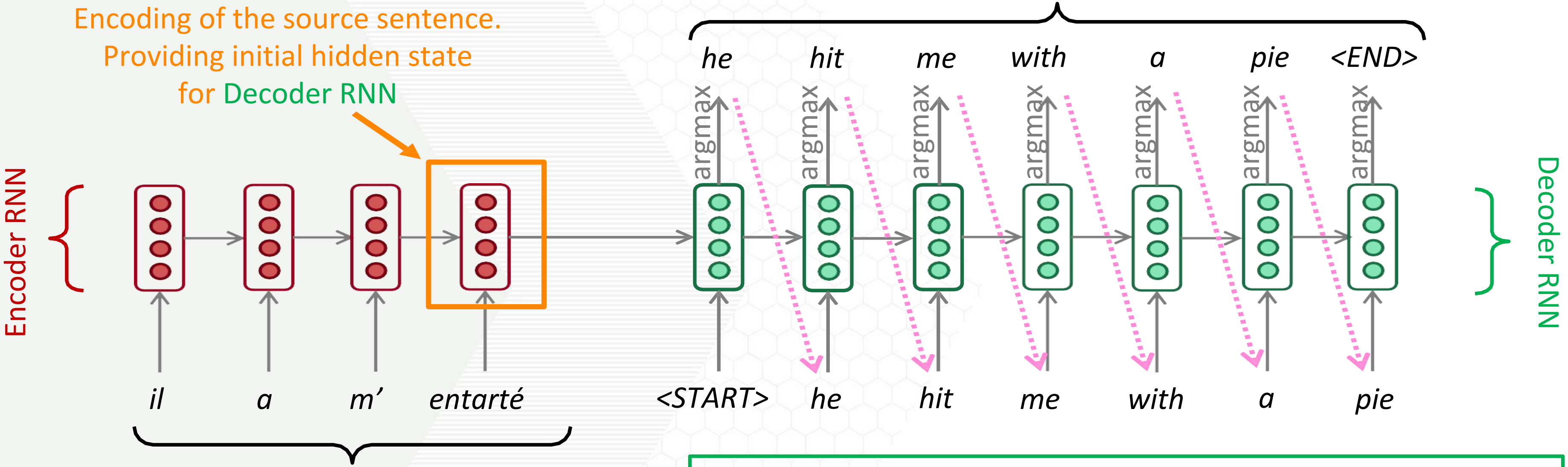
Encoder and Decoder Model

The sequence-to-sequence model

Encoding of the source sentence.

Providing initial hidden state

for Decoder RNN



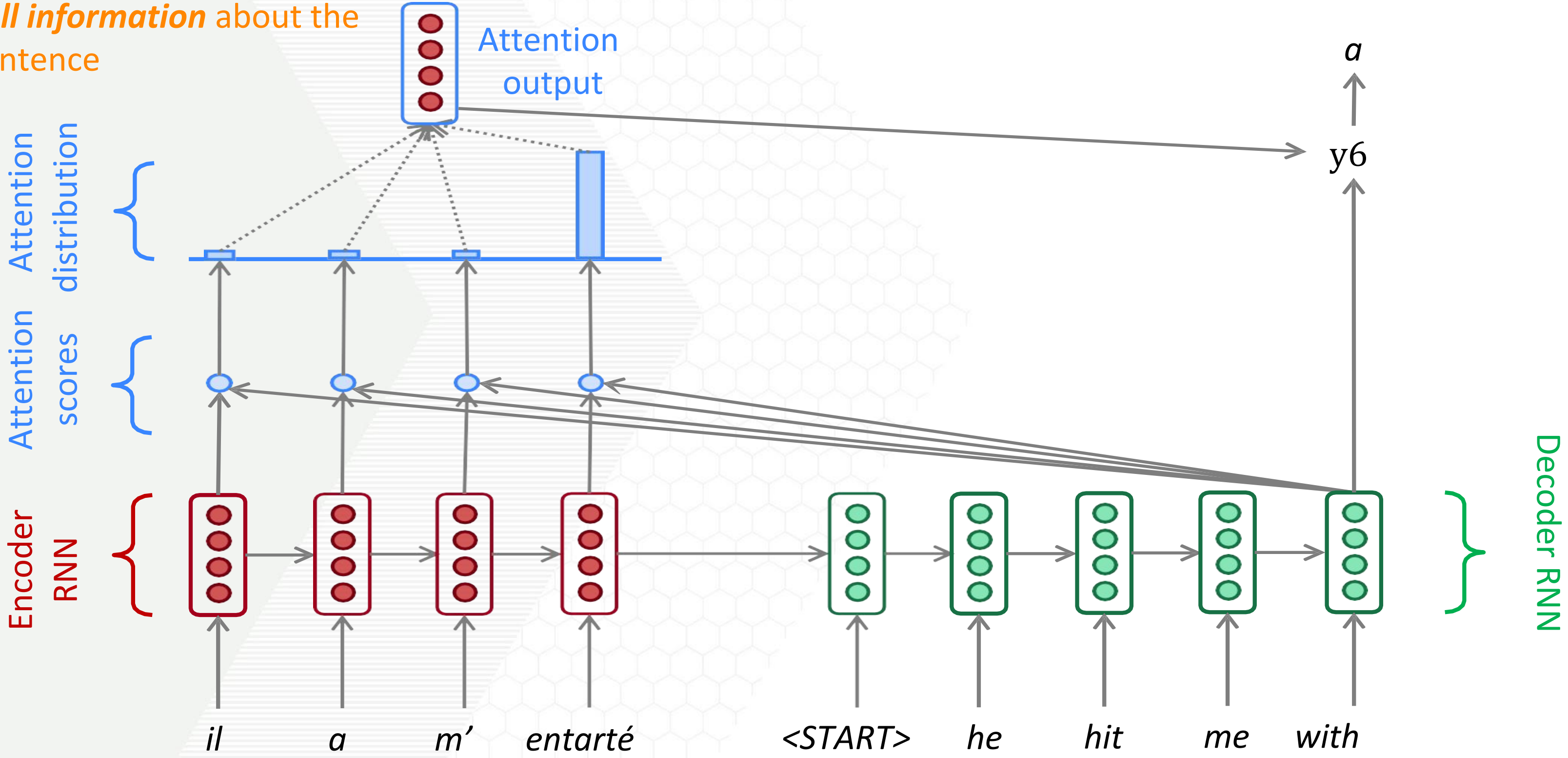
Encoder RNN produces an encoding of the source sentence.

Decoder RNN is a Language Model that generates target sentence, *conditioned on* encoding.

Note: This diagram shows **test time** behavior: decoder output is fed in ------> as next step’s input

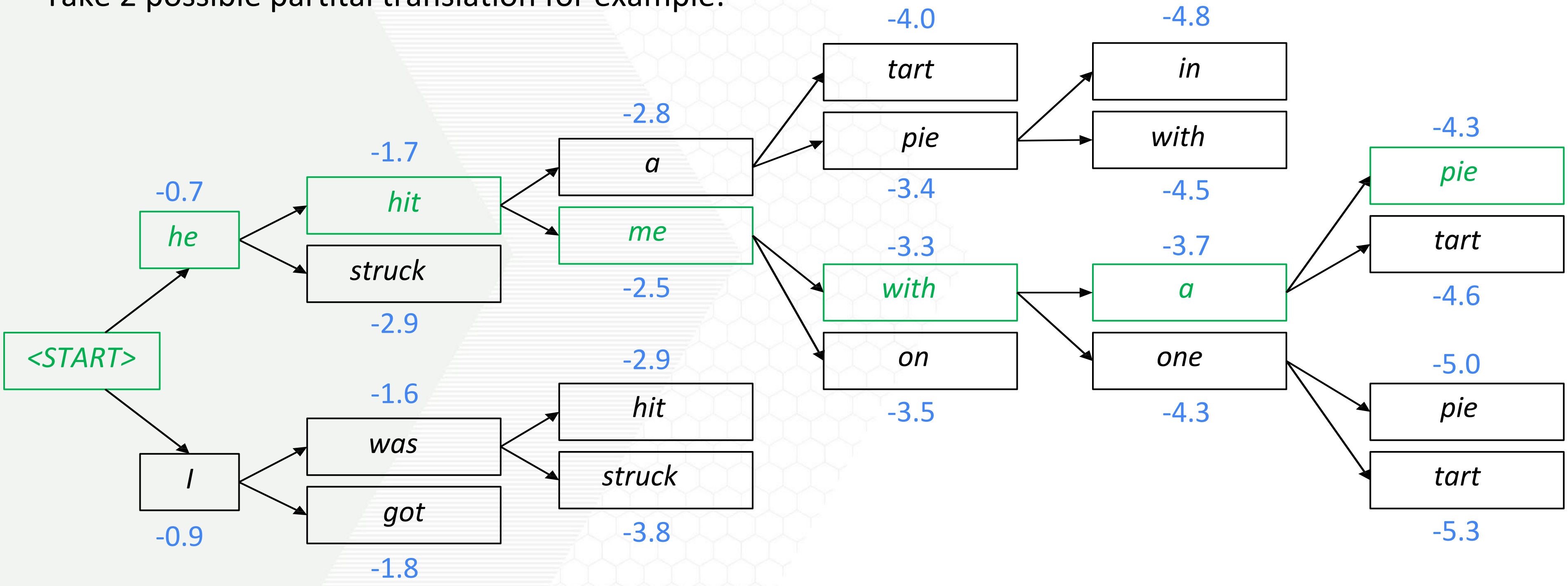
Attention Improvement

Encoding of the source sentence capture ***all information*** about the source sentence



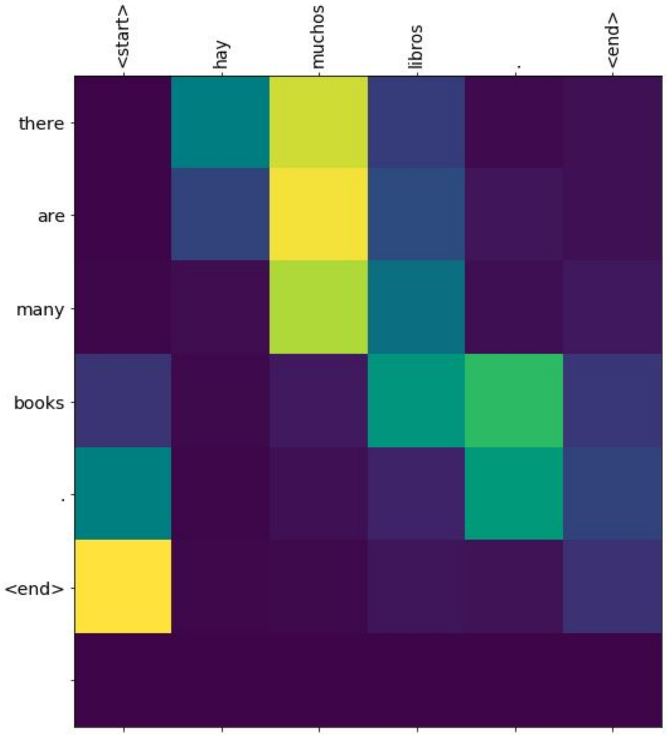
Translation Process

Take 2 possible partital translation for example:



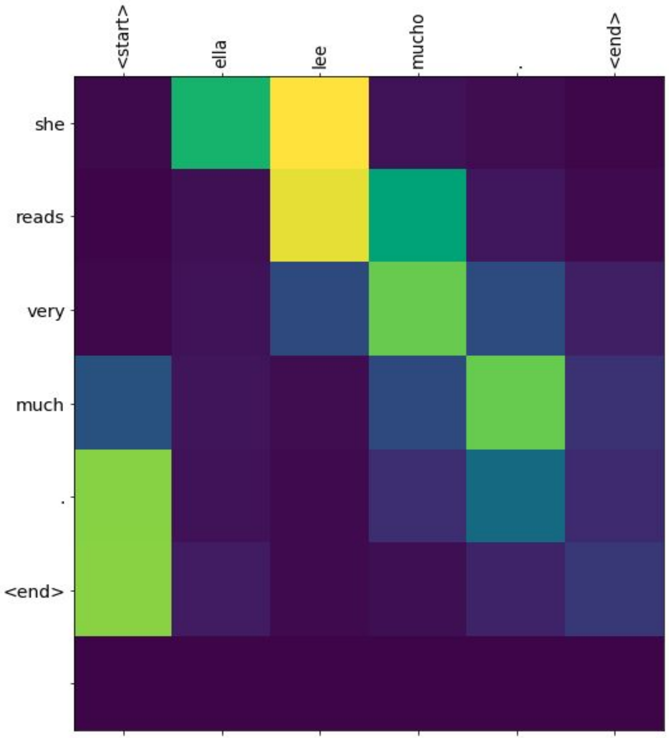
Good Cases

```
Input: <start> hay muchos libros . <end>
Predicted translation: there are many books . <end>
```



```
Gold: There are many books.
```

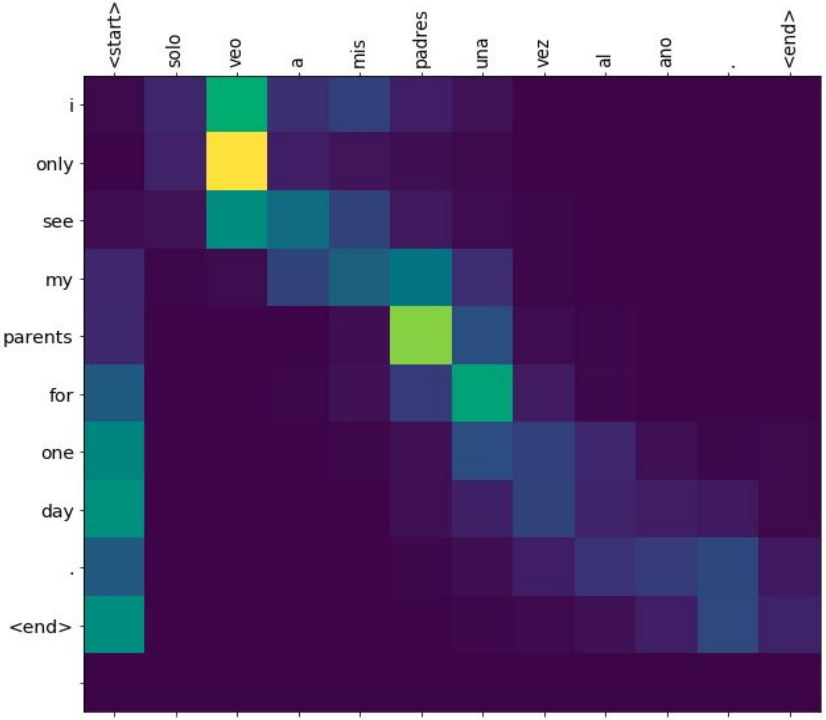
```
Input: <start> ella lee mucho . <end>
Predicted translation: she reads very much . <end>
```



```
Gold: She reads a lot
```

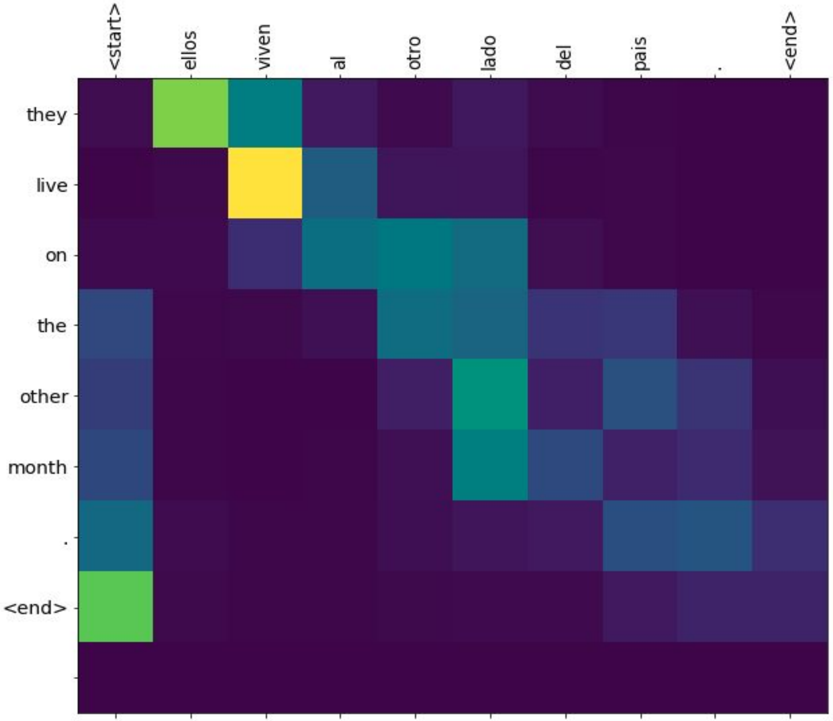
Bad Cases

```
Input: <start> solo veo a mis padres una vez al ano . <end>
Predicted translation: i only see my parents for one day . <end>
```



```
Gold: I only see my parents once a year.
```

```
Input: <start> ellos viven al otro lado del pais . <end>
Predicted translation: they live on the other month . <end>
```



```
Gold: They live on the other side of the country.
```

NMT Discussion

Advantages of NMT:

- Better performance, more fluent
- A single neural network to be optimized end-to-end and easy to train
- Requires much less human engineering effort. ***No rules! No dictionaries!***
- One method for ***all languages!***

Disadvantages of NMT:

- Hard to debug
- Difficult to control
- Depends on the data quality and model parameters
- Picks up the **bias** in data, like gender discrimination and so on

My Thoughts

- For the future work, I can train my model with larger dataset and more epochs to improve the performance.
- Google provides detailed tutorial and it's not so difficult to build your own neural machine translation tool on any pair of languages.
- Deep learning really reduce much more human work in machine translation than traditional methods. Neural machine translation is the **TRUE** machine translation.

References

[1] Sequence to sequence learning with neural networks, Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. NIPS, 2014.

[2] Learning phrase representations using RNN encoder-decoder for statistical machine translation, Kyunghyun Cho, Bart Van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. EMNLP 2014.

[3] Neural machine translation by jointly learning to align and translate, Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. ICLR, 2015.

[4] Effective approaches to attention-based neural machine translation, Minh-Thang Luong, Hieu Pham, and Christopher D Manning. EMNLP, 2015.

[5] Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation, Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, Jeffrey Dean. Technical Report, 2016.

[6] Stanford CS224 lecture8 Slide https://web.stanford.edu/class/cs224n/slides/cs224n-2019-lecture08-nmt.pdf