How does it work? Google Translate

CSE 705

Baasit Sharief (baasitsh@buffalo.edu)

June 1st, 2022

University at Buffalo The State University of New York

Background and Motivation

Why do we need it?

- You’re at a foreign country, and you need help with directions - locals don't know the languages you speak
- Friends talking smack behind your back in a different language
- Talking about numbers, about 50% of information on the web is in English, but only 20% know English
- Human only translation approach is too expensive or too slow to translate for personal work i.e. you can’t always have a human translator
- A lot of times, a completely accurate translation isn’t necessary

Non-ML based approach

- Word-Phrase translation
- Also known as PBMT (Phrase Based Machine Translation)
- Involves usage of a look-up table (LUT)
- Word/Phrase in a language is mapped to another word/phrase in the resulting language
- That seems pretty easy, right?

English Greetings in French	
good morning	bonjour
good evening	bonsoir
good night	bonne nuit
goodbye	au revoir
hi / bye	salut
thank you	merci
thank you very much	merci beaucoup

Problems with Non-ML approach

- Any language has two components
  - **Tokens**
  - **Grammar**
- Grammar defines the arrangement of tokens such that a given sequence of tokens makes sense
- Word-Word and Phrase Translations do not incorporate grammar and are very naive approaches

Sentence:

"It is a beautiful day"

Word-Phase Translation (French):

"Ce-est-un-belle-le jour"

(It) (is) (a) (beautiful) (day)

Translation:

"C'est une belle journée"

University at Buffalo The State University of New York

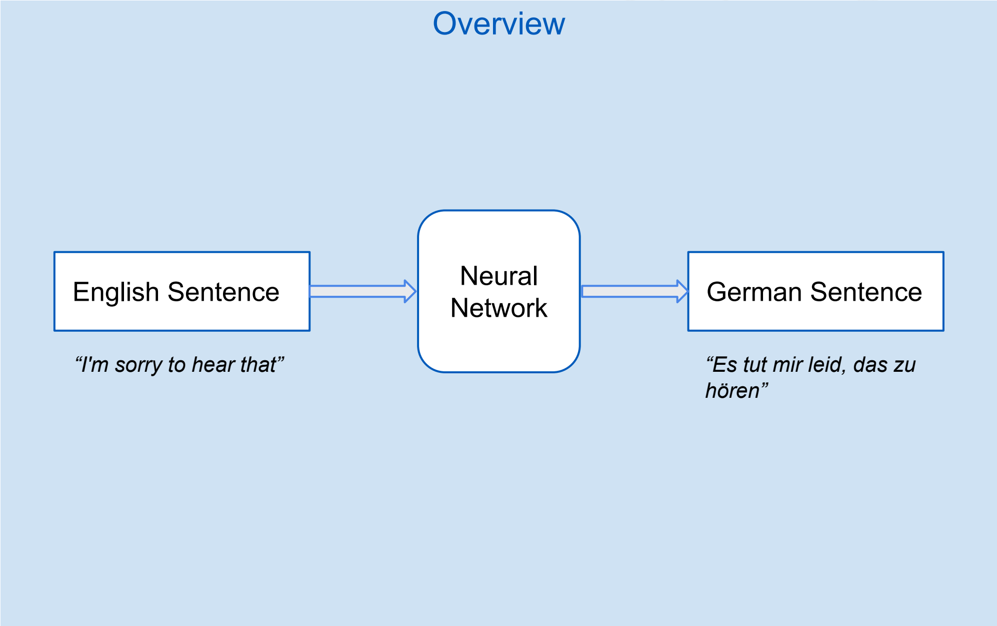
Neural Machine translation

Modeling the problem

- Sentence to Sentence translation over word-word/phrase
- Model sequences to another sequence
  - English sentence to German sentence
  - Input sequence → Translator → Output Sequence
- Input and output shapes might be different
  - Depends on language
- *Syntax* is important - correct grammatical structure
- *Semantics* i.e. the resultant sentence has some contextual meaning

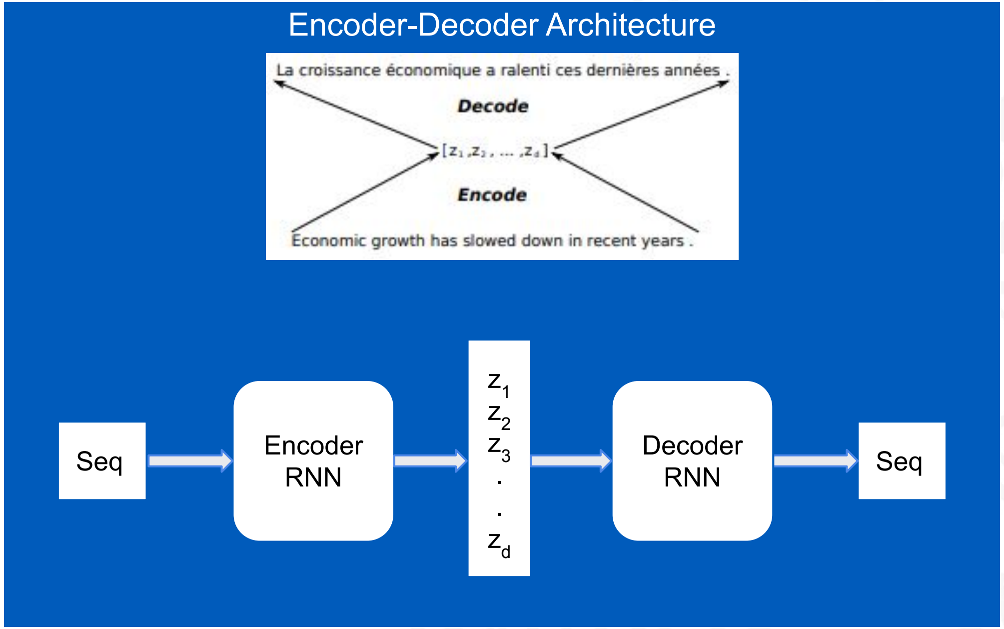
Using a Neural Network

- The advancements in Deep Learning has allowed us to model sequences from sequences with help of sequence modeling
- Seq2Seq model architecture which involves Encoder and Decoder Recurrent Neural Network (RNN) cells
- Can model different sized inputs to different sized outputs as well



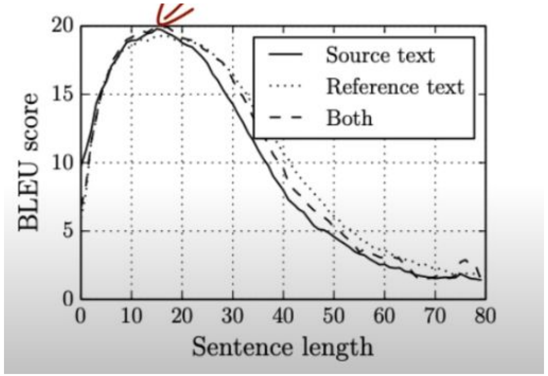
Encoder-Decoder Architecture (Seq2Seq)

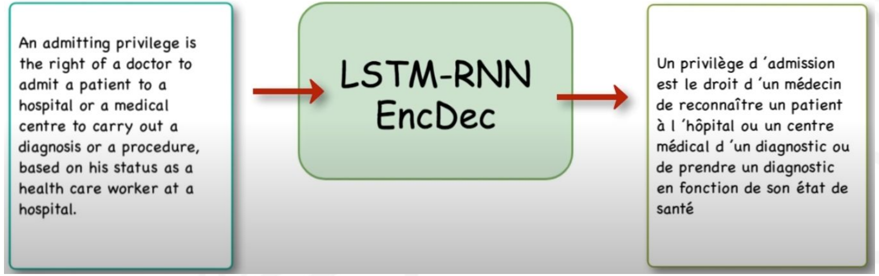
- Encoder inputs a sequence and outputs a vector
- Decoder takes the vector and converts it into a sequence
- Encoder+Decoder inputs a sequence and outputs a sequence, hence Seq2Seq
- Also used in time-series prediction

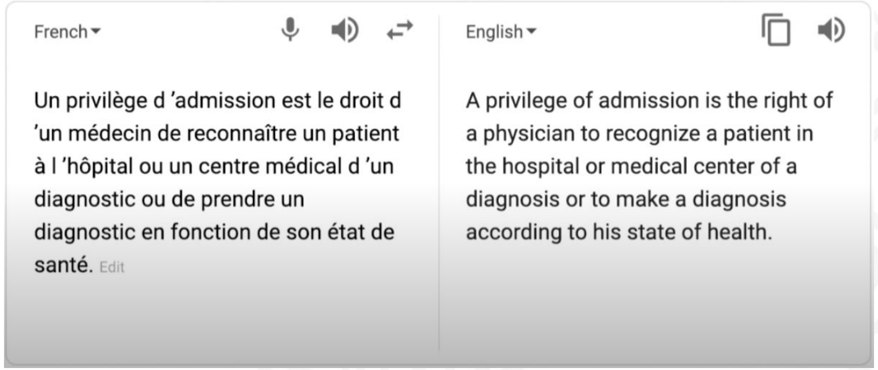


Problems with Vanilla Seq2Seq

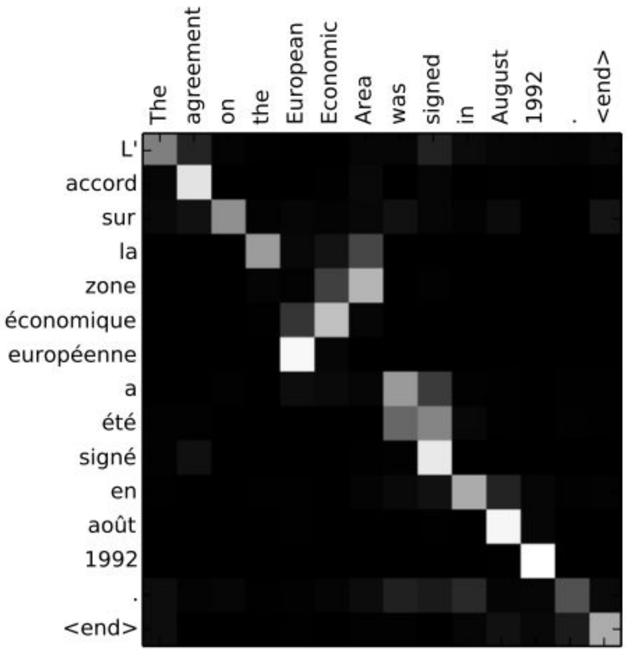
- Bad performance on longer sentences
- Even with LSTM cells, the best performance achieved is around 15 words



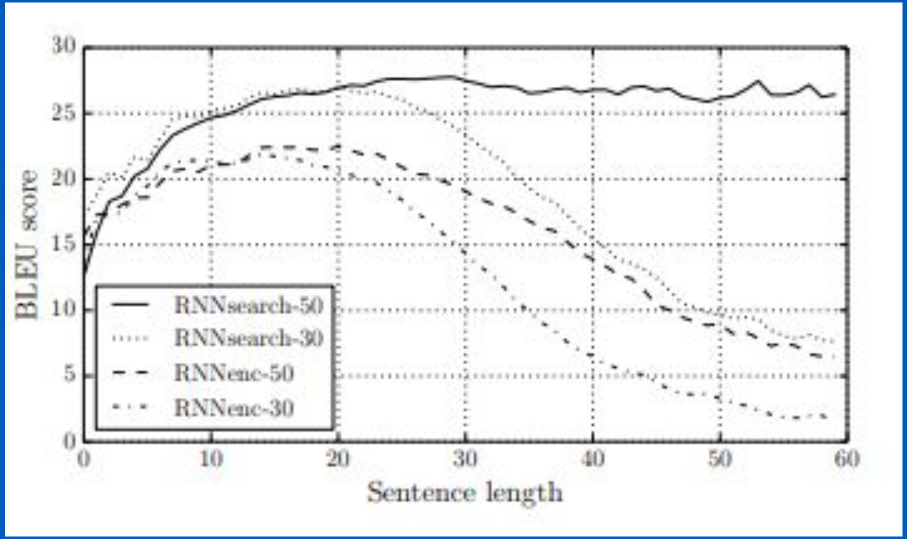




What if we try to learn to align and translate?



Results

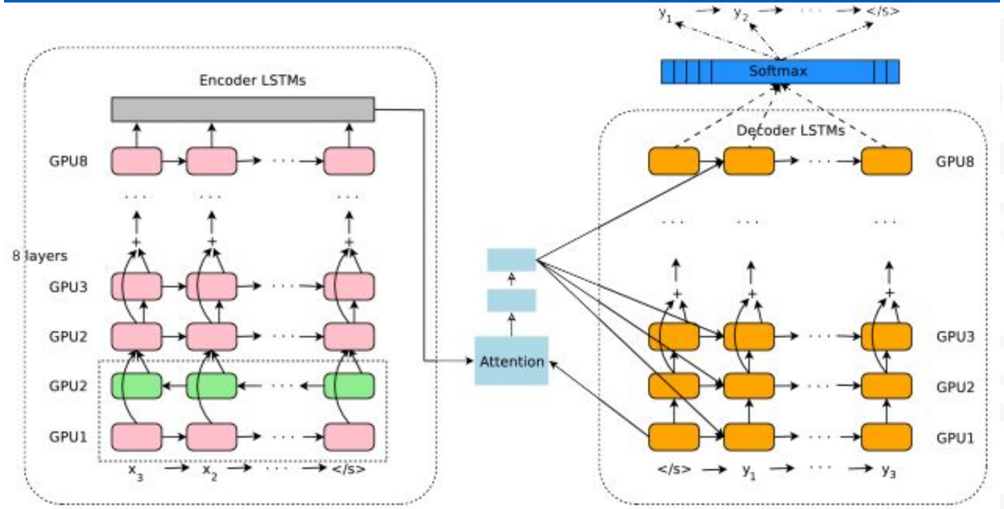


- Also known as the attention mechanism

Google's Neural Translation Model

- LSTM cells over vanilla RNN cells
- Use of residual connections for better gradient flow
- Attention function is a simple 1-layered MLP
- First layer of the encoder is bi-directional
- Uses a modified beam search on decoder outputs to get resultant sequence
- Coverage penalty and length normalization

Google NMT architecture (2016)



Training Criteria

- Maximum log-Likelihood objective - maximizing the sum of log probabilities of the ground-truth outputs given the corresponding inputs
  - does not reflect the task reward function as measured by BLEU score
  - does not explicitly encourage a ranking among *incorrect* output sequences
- In comes Reinforcement Learning
  - model refinement using the expected reward objective
  - Use GLEU score for sentence pairs
- First train model using the maximum likelihood objective until convergence
- Refine this model using a mixed maximum likelihood and expected reward objective, until BLEU score on a development set is no longer improving

$$\mathcal{O}_{\mathrm{ML}}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log P_{\theta}(Y^{*(i)} \mid X^{(i)})$$

Max log-likelihood objective

$$\mathcal{O}_{\mathrm{RL}}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \sum_{Y \in \mathcal{Y}} P_{\theta}(Y \mid X^{(i)}) \; r(Y, Y^{*(i)}).$$

Expected Reward Objective

$$\mathcal{O}_{\mathrm{Mixed}}(\boldsymbol{\theta}) = \alpha * \mathcal{O}_{\mathrm{ML}}(\boldsymbol{\theta}) + \mathcal{O}_{\mathrm{RL}}(\boldsymbol{\theta})$$

Fine-tuning Objective Function

Experimental Results

Without RL training

Table 4: Single model results on WMT En→Fr (newstest2014)

Model	BLEU	CPU decoding time per sentence (s)
Word	37.90	0.2226
Character	38.01	1.0530
WPM-8K	38.27	0.1919
WPM-16K	37.60	0.1874
WPM-32K	38.95	0.2118
Mixed Word/Character	38.39	0.2774
PBMT [15]	37.0	
LSTM (6 layers) [31]	31.5	
LSTM (6 layers + PosUnk) [31]	33.1	
Deep-Att [45]	37.7	
Deep-Att + PosUnk [45]	39.2	

Table 5: Single model results on WMT En→De (newstest2014)

Model	BLEU	CPU decoding time per sentence (s)
Word	23.12	0.2972
Character (512 nodes)	22.62	0.8011
WPM-8K	23.50	0.2079
WPM-16K	24.36	0.1931
WPM-32K	24.61	0.1882
Mixed Word/Character	24.17	0.3268
PBMT [6]	20.7	
RNNSearch [37]	16.5	
RNNSearch-LV [37]	16.9	
RNNSearch-LV [37]	16.9	
Deep-Att [45]	20.6	

Model Ensemble and RL training

Table 7: Model ensemble results on WMT En→Fr (newstest2014)

Model	BLEU
WPM-32K (8 models)	40.35
RL-refined WPM-32K (8 models)	41.16
LSTM (6 layers) [31]	35.6
LSTM (6 layers + PosUnk) [31]	37.5
Deep-Att + PosUnk (8 models) [45]	40.4

Table 8: Model ensemble results on WMT En→De (newstest2014). See Table 5 for a comparison against non-ensemble models.

Model	BLEU
WPM-32K (8 models)	26.20
RL-refined WPM-32K (8 models)	26.30

Table 9: Human side-by-side evaluation scores of WMT En→Fr models.

Model	BLEU	Side-by-side averaged score
PBMT [15]	37.0	3.87
NMT before RL	40.35	4.46
NMT after RL	41.16	4.44
Human		4.82

Table 10: Mean of side-by-side scores on production data

	PBMT	GNMT	Human	Relative Improvement
English → Spanish	4.885	5.428	5.504	87%
English → French	4.932	5.295	5.496	64%
English → Chinese	4.035	4.594	4.987	58%
Spanish → English	4.872	5.187	5.372	63%
French → English	5.046	5.343	5.404	83%
Chinese → English	3.694	4.263	4.636	60%

Further developments

Transformer Architecture

- Completely replaces usage of recurrent neural network cells which are slow in computations
- Based solely on attention mechanisms
- Only drawback is longer training time
- Achieves 28.4 BLEU on the WMT 2014 English-to-German and BLEU score of 41.8 on the WMT 2014 English-to-French translation task, establishing a new single-model state-of-the-art BLEU score of 41.8
- Generalizes well to other tasks

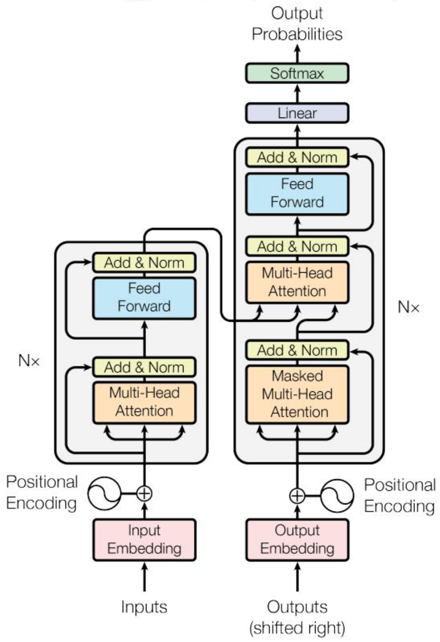


Figure 1: The Transformer - model architecture.

References

1. How Google Translate Works - The Machine Learning Algorithm Explained! - YouTube
2. Long Short-Term Memory (Hochreiter et al., 1997)
3. Recurrent Convolutional Neural Networks for Discourse Compositionality (Kalchbrenner et al., 2013)
4. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation (Cho et al., 2014)
5. Sequence to Sequence Learning with Neural Networks (Sutskever et al., 2014)
6. Bidirectional Recurrent Neural Networks (Schuster et al., 1997)
7. On the Properties of Neural Machine Translation: Encoder–Decoder Approaches (Cho et al., 2014)
8. Neural Machine Translation by jointly learning to align & translate (Bahdanau et al., 2016)
9. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation (Wu et al., 2016)
10. Attention Is All You Need (Vaswani et al., 2017)

Thank you!