Sanyam Bhutani · Follow
Mar 21, 2019 · 4 min read · Listen

# “Sequence to Sequence Learning with Neural Networks”: Paper Discussion

The fifth blog post in the 5-minute Papers series.

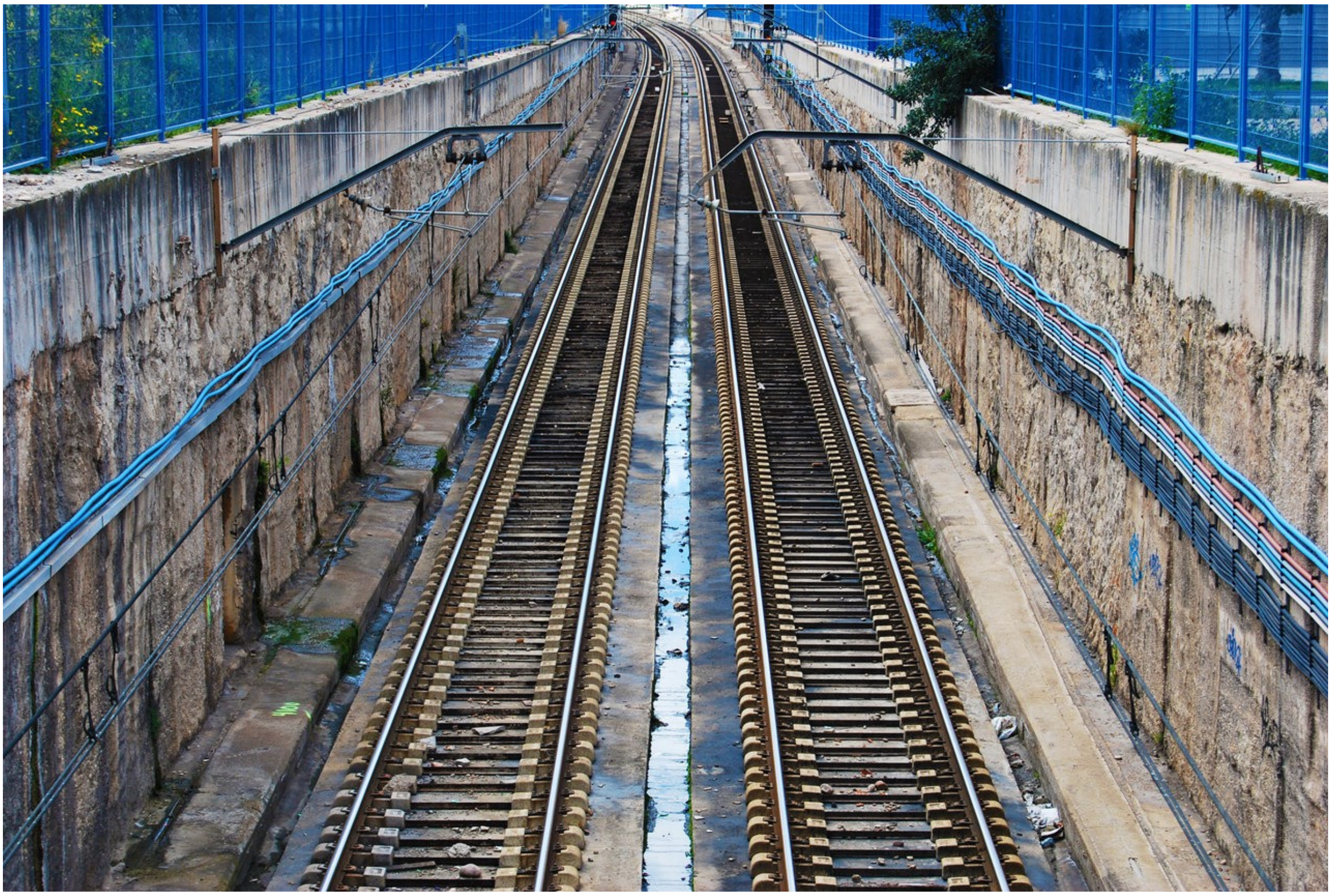


Photo by Francisco Casero on Unsplash

For today’s paper summary, I will be discussing one of the “classic”/pioneer papers for Language Translation, from 2014 (!): “Sequence to Sequence Learning with Neural Network” by Ilya Sutskever et al

## TL;DR

The Seq2Seq with Neural Networks was one of the pioneer papers to show that Deep Neural Nets can be used to perform “End to End” Translation. The paper demonstrates that LSTM can be used with minimum assumptions, proposing a 2 LSTM (an “Encoder”-“Decoder”) architecture to do Langauge Translation from English To French, showing the promise of Neural Machine Translation (NMT) over Statistical Machine Translation (SMT)

## Context

To highlight again, please keep in mind that the paper is from 2014, when there were no widely open sourced Frameworks such as TF or PyTorch and DNN(s) were just starting to show promise so many ideas presented in the paper might seem very obvious to us today.

The task is to perform Translation of a “Sequence” of sentences/words from English to French.

The DNN techniques expected a fixed dimensionality which was a limitation for NLP, Speech.

- First one acts an Encoder:
  Takes your input and maps it into a fixed dimension vector
- The second acts as a Decoder:
  Takes the fixed vector and maps it to an output sequence.

**The Model**

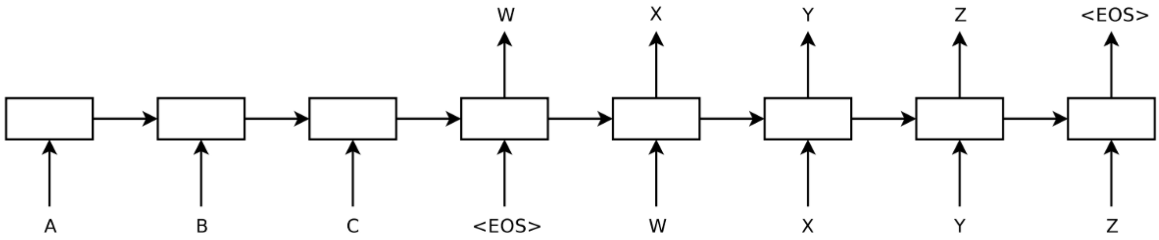


Figure 1: Our model reads an input sentence “ABC” and produces “WXYZ” as the output sentence. The model stops making predictions after outputting the end-of-sentence token. Note that the LSTM reads the input sentence in reverse, because doing so introduces many short term dependencies in the data that make the optimization problem much easier.

The LSTM is tasked to predict the conditional probability of a target sequence given an input sequence generated from the last layer. The generated sequence using this probability may have a length different from the source text.

$$p(y_1, \ldots, y_{T'} | x_1, \ldots, x_T) = \prod_{t=1}^{T'} p(y_t | v, y_1, \ldots, y_{t-1})$$

- **Two LSTM(s) (Encoder-Decoder):**
  This allows training the LSTM on multiple language pairs simultaneously.
- **“Deep LSTM(s)”:**
  The paper mentions Deep LSTM(s) of 4 layers perform better.
- **Reversing the order of Input:**
  The paper really highlights the trick of inverting the input sequence when mapping it to the output sequence which makes it “easier for SGD” to “establish communication” between input and output.
  It also enhances both short and long term predictions of the LSTM. The authors suggest that this might be due to “minimal time lag” where the distance between the generated and source words is minimized by reversing the order.

**Training details:**

- 160,000 of the most frequent words for the source language and 80,000 of the most frequent words for the target language. Every out-of-vocabulary word was replaced with a special “UNK” token.
- The training was done to maximize the log probability.

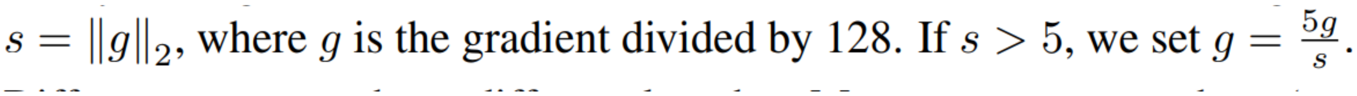
$$1/|\mathcal{S}| \sum_{(T,S)\in\mathcal{S}} \log p(T|S)$$

- Translations are produced using the most probable outcomes:

- The “Hypothesis” is created by doing a beam search which is stopped when it reaches an “<EOS>” (End Of String character)

> *Hypothesis are the pairs of sentences that are generated*

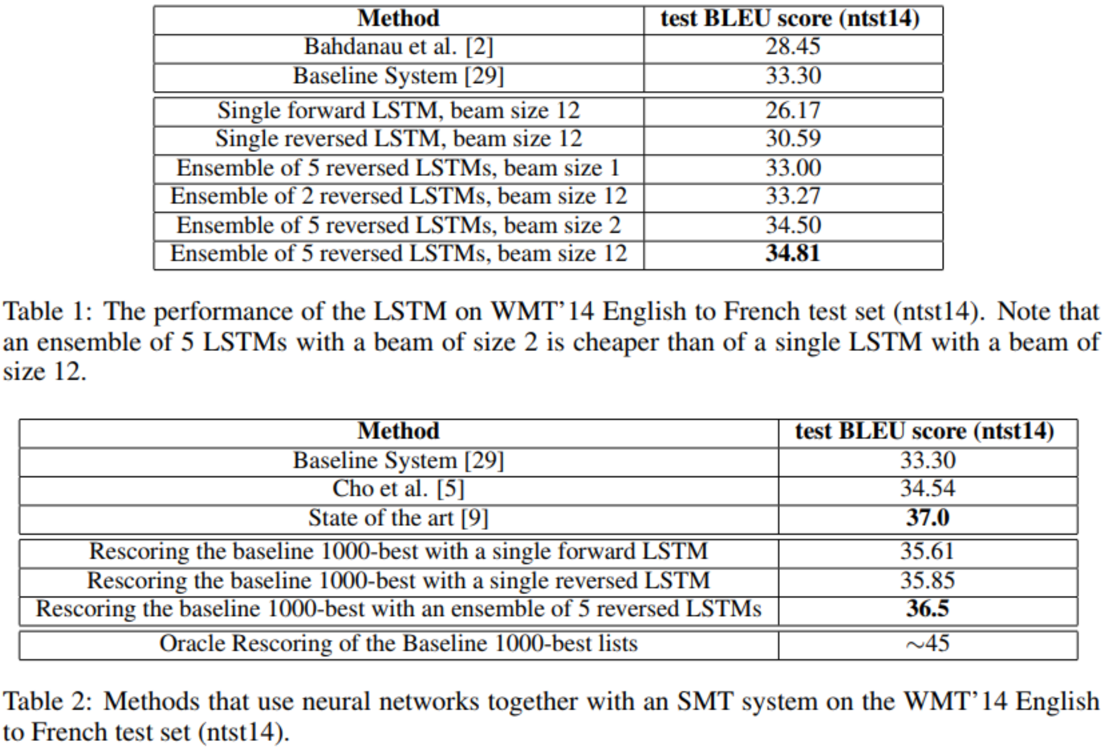
- All of the LSTM’s parameters are initialized with the uniform distribution between -0.08 and 0.08.
- To deal with exploding gradients, the authors scale the gradients for each batch as follows:



$s = \|g\|_2$, where $g$ is the gradient divided by 128. If $s > 5$, we set $g = \frac{5g}{s}$.

### Results

- The best results are obtained with an ensemble of LSTMs that differ in their random initializations and in the random order of mini-batches.



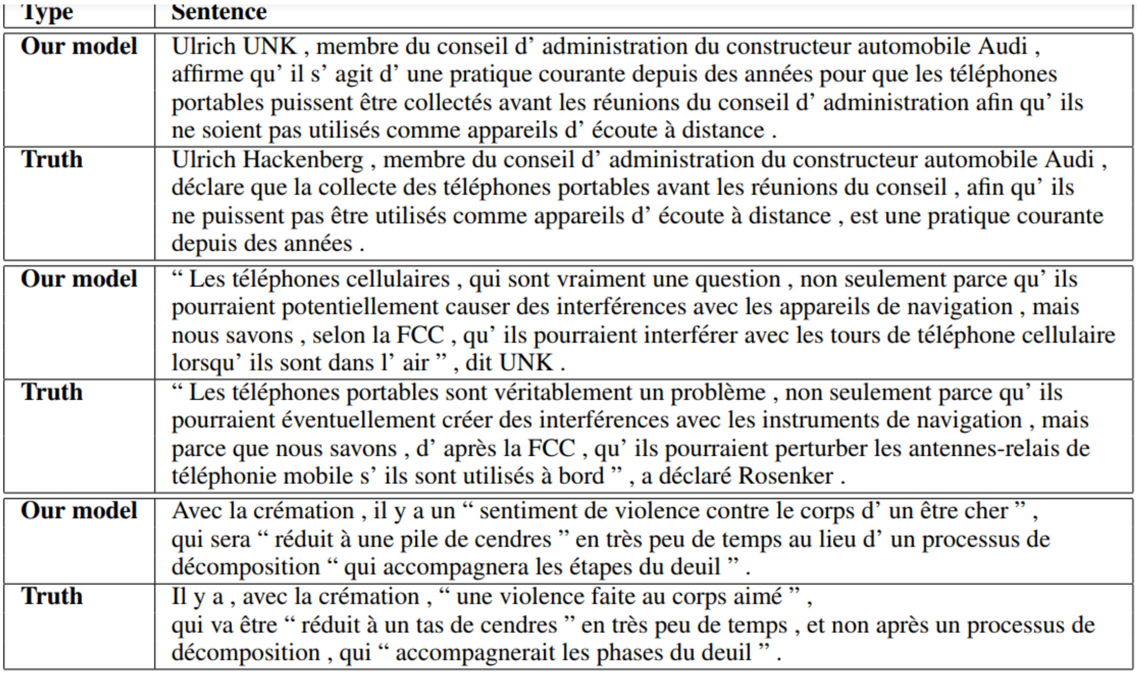
| Method | test BLEU score (ntst14) |
|---|---|
| Bahdanau et al. [2] | 28.45 |
| Baseline System [29] | 33.30 |
| Single forward LSTM, beam size 12 | 26.17 |
| Single reversed LSTM, beam size 12 | 30.59 |
| Ensemble of 5 reversed LSTMs, beam size 1 | 33.00 |
| Ensemble of 2 reversed LSTMs, beam size 12 | 33.27 |
| Ensemble of 5 reversed LSTMs, beam size 2 | 34.50 |
| Ensemble of 5 reversed LSTMs, beam size 12 | **34.81** |

Table 1: The performance of the LSTM on WMT’14 English to French test set (ntst14). Note that an ensemble of 5 LSTMs with a beam of size 2 is cheaper than of a single LSTM with a beam of size 12.

| Method | test BLEU score (ntst14) |
|---|---|
| Baseline System [29] | 33.30 |
| Cho et al. [5] | 34.54 |
| State of the art [9] | **37.0** |
| Rescoring the baseline 1000-best with a single forward LSTM | 35.61 |
| Rescoring the baseline 1000-best with a single reversed LSTM | 35.85 |
| Rescoring the baseline 1000-best with an ensemble of 5 reversed LSTMs | **36.5** |
| Oracle Rescoring of the Baseline 1000-best lists | $\sim$45 |

Table 2: Methods that use neural networks together with an SMT system on the WMT’14 English to French test set (ntst14).

- Long Sentences: The translation showed some surprising long length results:



| Type | Sentence |
|---|---|
| **Our model** | Ulrich UNK , membre du conseil d’ administration du constructeur automobile Audi , affirme qu’ il s’ agit d’ une pratique courante depuis des années pour que les téléphones portables puissent être collectés avant les réunions du conseil d’ administration afin qu’ ils ne soient pas utilisés comme appareils d’ écoute à distance . |
| **Truth** | Ulrich Hackenberg , membre du conseil d’ administration du constructeur automobile Audi , déclare que la collecte des téléphones portables avant les réunions du conseil , afin qu’ ils ne puissent pas être utilisés comme appareils d’ écoute à distance , est une pratique courante depuis des années . |
| **Our model** | “ Les téléphones cellulaires , qui sont vraiment une question , non seulement parce qu’ ils pourraient potentiellement causer des interférences avec les appareils de navigation , mais nous savons , selon la FCC , qu’ ils pourraient interférer avec les tours de téléphone cellulaire lorsqu’ ils sont dans l’ air ” , dit UNK . |
| **Truth** | “ Les téléphones portables sont véritablement un problème , non seulement parce qu’ ils pourraient éventuellement créer des interférences avec les instruments de navigation , mais parce que nous savons , d’ après la FCC , qu’ ils pourraient perturber les antennes-relais de téléphonie mobile s’ ils sont utilisés à bord ” , a déclaré Rosenker . |
| **Our model** | Avec la crémation , il y a un “ sentiment de violence contre le corps d’ un être cher ” , qui sera “ réduit à une pile de cendres ” en très peu de temps au lieu d’ un processus de décomposition “ qui accompagnera les étapes du deuil ” . |
| **Truth** | Il y a , avec la crémation , “ une violence faite au corps aimé ” , qui va être “ réduit à un tas de cendres ” en très peu de temps , et non après un processus de décomposition , qui “ accompagnerait les phases du deuil ” . |

Table 3: A few examples of long translations produced by the LSTM alongside the ground truth translations. The reader can verify that the translations are sensible using Google translate.

- The figure shows that the representations are sensitive to the order of words while being fairly insensitive to the replacement of an active voice with a passive voice.

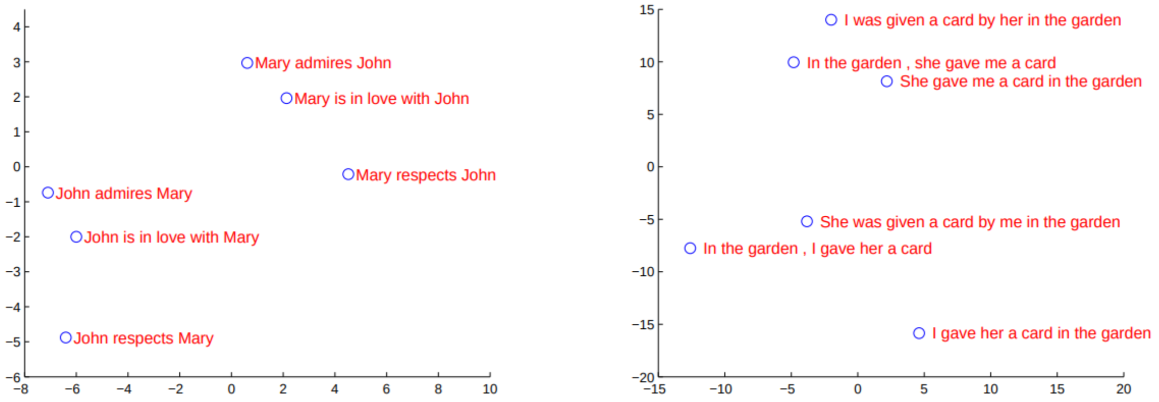


Figure 2: The figure shows a 2-dimensional PCA projection of the LSTM hidden states that are obtained after processing the phrases in the figures. The phrases are clustered by meaning, which in these examples is primarily a function of word order, which would be difficult to capture with a bag-of-words model. Notice that both clusters have similar internal structure.

### Conclusion and Thoughts

- The paper was one of the first to show that DNN(s) or specifically, LSTM(s) show much promise for “Seq2Seq Learning”
- The paper also mentions the use of 2 LSTM(s), first is used to map a varying length input to a fixed length vector which then gets mapped to the target.
- LSTM(s) were shown to be surprisingly good on long sentences.

*If you found this interesting and would like to be a part of My Learning Path, you can find me on Twitter here.*

*If you’re interested in reading about Deep Learning and Computer Vision news, you can check out my newsletter here.*

*If you’re interested in reading a few best advice from Machine Learning Heroes: Practitioners, Researchers, and Kagglers. Please click here*