Ofri Masad 201267226
Nir Friedman 305283178

# Project Report – Sequence2Sequence

## Title: Sequence to Sequence Learning with Neural Networks

## Reference:

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." *Advances in neural information processing systems*. 2014.

## Motivation:

This paper refers to the problem of machine translation, i.e. translating sentences from one language to another using machine learning, and in this case specifically using a multilayered LSTM neural network.

## Short description:

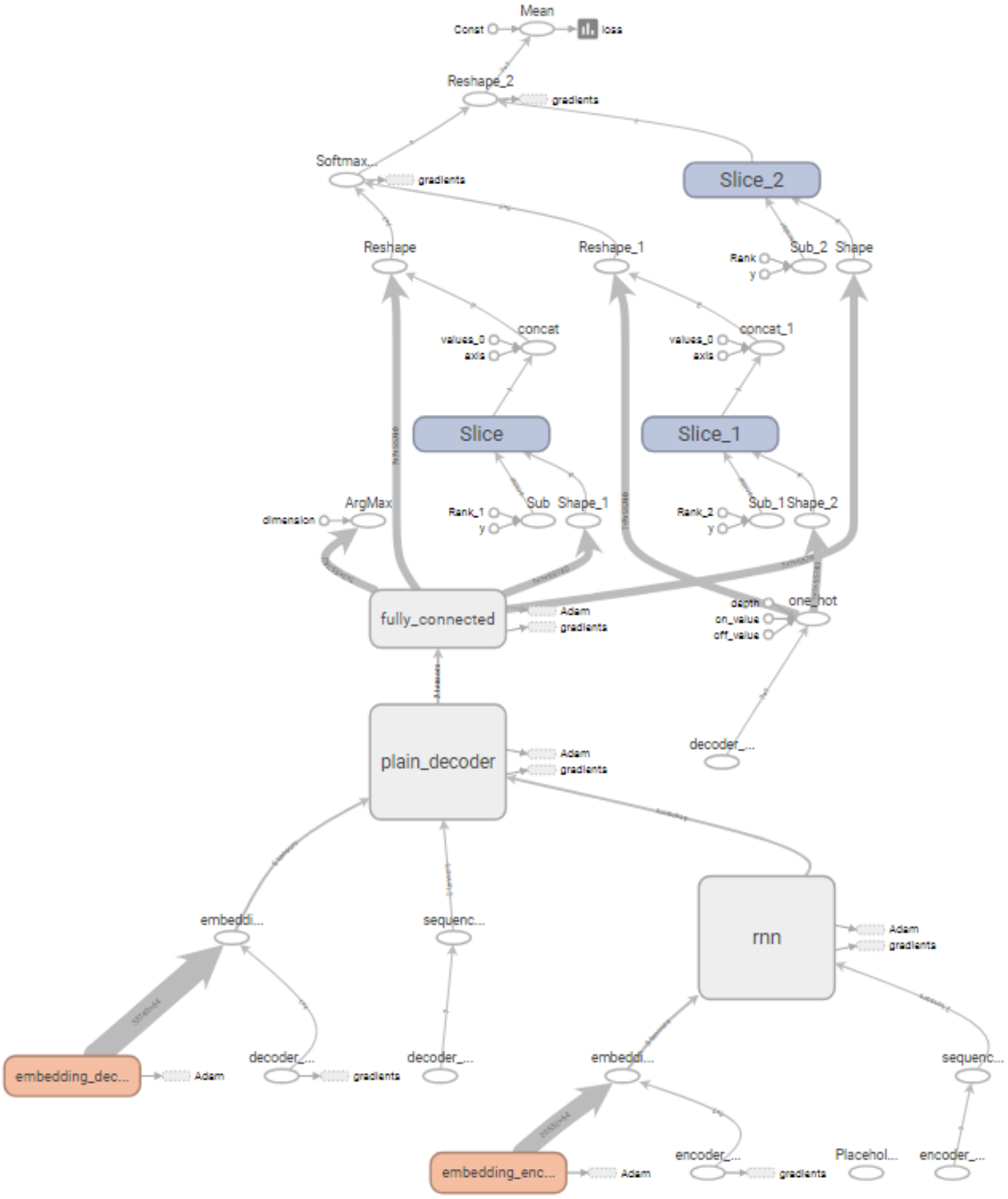
The algorithm was designed to receive a sequence of words in some language and output the translated sequence in a target language. Each word was embedded to a 1000-dimensional vector, before passing as an input to the network. The network consisted of 4 layers – 2 layers for encoding the input sequence, and 2 layers decoding the output sequence. Also, the authors stated that reversing the order of the input had an extremely positive influence on the results, so we did the same.

The article tested translation from English to French, and we decided to translate from English to Hebrew instead.

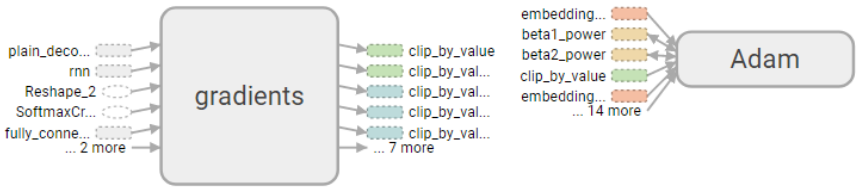
## Architecture:

As stated above, this paper used a 4 layers LSTM network, with a 1000-dimensional embedding to translate a sentence from one language to another. For our training process to end in feasible time, we used an embedding size of 64, 64 LSTM cells (hidden states) in each LSTM layer and calculated the validation score every 250 iterations. Furthermore, we used teacher forcing in the learning process, the article did not declare whether they used it or not. In addition, SGD did not get as good results at all, so we decided to use Adam instead. This architecture can be examined in the following diagram:

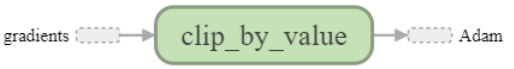
Ofri Masad 201267226
Nir Friedman 305283178

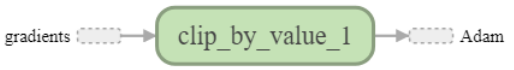


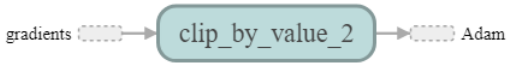
Ofri Masad 201267226
Nir Friedman 305283178

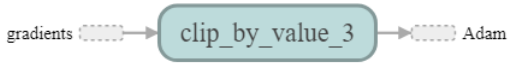




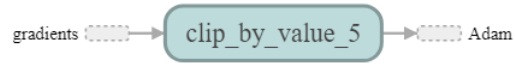



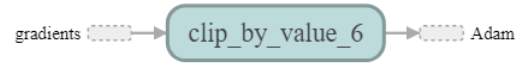



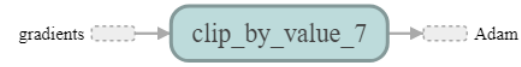



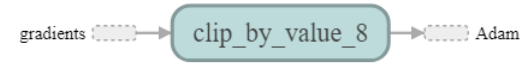



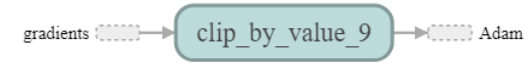


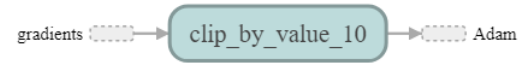


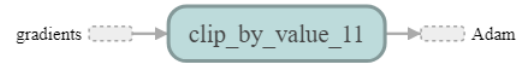



Ofri Masad 201267226
Nir Friedman 305283178

**Hyper-parameters:**

| | Paper Original Values | Our Implementation Values |
|---|---|---|
| **Initial learning rate** | 0.7 | 0.001 |
| **learning rate decline** | After 5 epochs, begin halving the learning rate every half epoch. | None as we start from low learning rate. |
| **Non-linear functions** | Did not state | ReLu |
| **Loss function** | Categorical cross entropy | Categorical cross entropy |
| **# of epochs** | 7.5 | 25 |
| **Batch size** | 128 | 64 |
| **Embedding size** | 1000 | 64 |
| **LSTM layers size (# hidden states)** | 1000 | 64 |
| **Parameters initialization** | [-0.08 , 0.08] | [-0.08 , 0.08] |
| **Validation while training** | None | 20% of the training data was used for validation |
| **Optimization Function** | SGD | Adam |

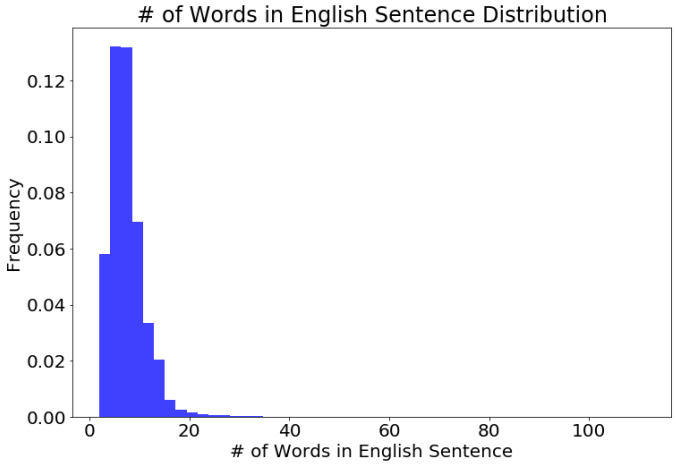
## Dataset:

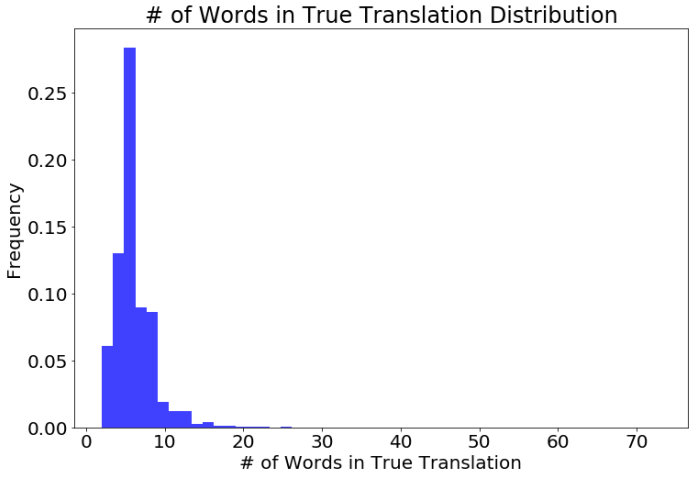
As stated above, the paper translated sentences from English to French, but we decided to translate from English to Hebrew instead, so there are many differences between the data that was used:

| | Paper Original Values | Our Implementation Values |
|---|---|---|
| **Data Set** | WMT'14 English to French | English to Hebrew from https://tatoeba.org/eng |
| **Size** | 12 million sentences | 160,951 sentences |
| **Source Vocabulary size** | 160,000 of the most frequent words | 20,550 words |
| **Target Vocabulary size** | 160,000 of the most frequent words | 55,736 words |

This dataset contains multiple Hebrew translations to the same English sentence, and we included all of them. This action might add difficulty in learning the model and its performance.

As seen in the figure below, the sentences lengths differ substantially from English to Hebrew.

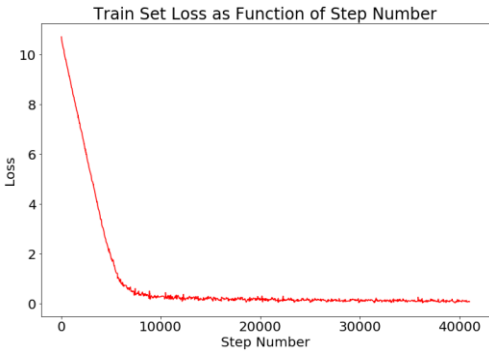


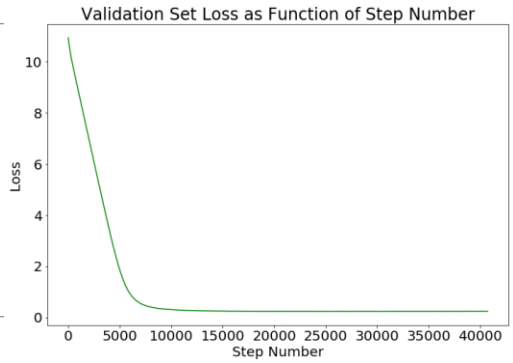



Ofri Masad 201267226
Nir Friedman 305283178

## Results:

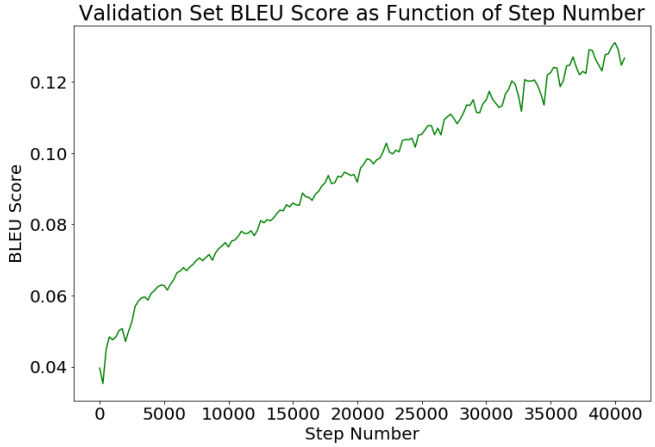
- **BLEU score-** The main metric that was used in the article. This score was created to evaluate the quality of text which has been machine-translated[1]. It uses a modified form of precision that takes into consideration the fact that machine translation tends to generate more words than in a reference text. In the paper they used a specific implementation of BLEU, written in pearl which made the integration within the training more difficult, so we used a version provided by the NLTK package.
- **Final results on test set-** The metric that was used to evaluate the results was BLEU. The average value on the test set was *0.13* compared to *0.127* in the last validation set.
- **Training validation and test set loss:**

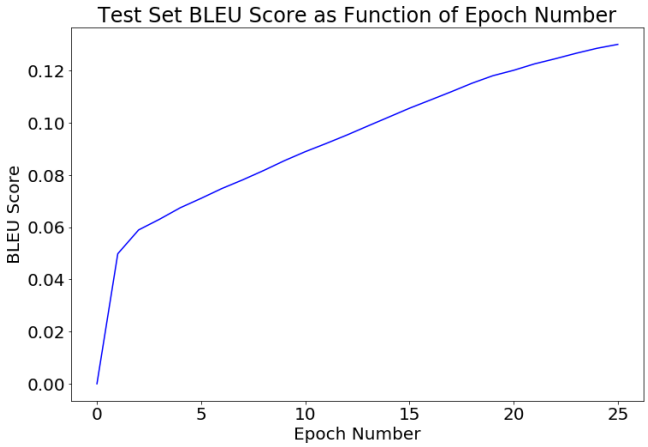




- **Validation and test set BLEU score:**



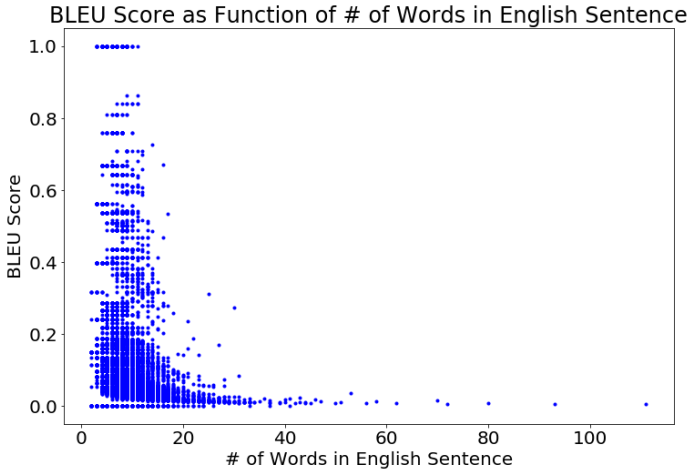

As seen in the above figure, we get the expected results – loss decreases rapidly at first, and then slowly reaches a plateau, while the BLEU score increases with training.

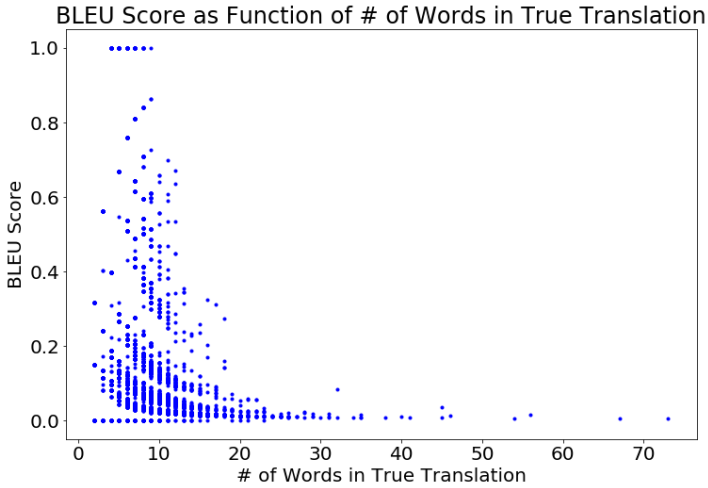
Ofri Masad 201267226
Nir Friedman 305283178

- **Test Set Translation Examples:**

| Original Sentence | Our Model Translation | True Translation |
|---|---|---|
| what a surprise to see you here ! | איזה הפתעה לראות אותך כאן ! | איזו הפתעה לפגוש אותך פה! |
| where is your cap ? | איפה הכובע שלך ? | איפה הכובע שלך ? |
| i have no objection to against your plan . | אין לי התנגדות לתכנית שלך . | אין לי התנגדות לתכנית שלך. |
| stop bothering me ! | תפסיק להטריד אותי ! | תפסיק להטריד אותי! |
| tom helped mary wash the dishes . | תום עזר למרי לשטוף את הכלים . | תום עזר למרי לשטוף את הכלים . |
| the property was almost completely overgrown with wild blackberry bushes . | החדר זה הזה היה הוא היא במשאבי בש | האתר כוסה כמעט כולו בשיחי אסנת בר. |
| rome is an italian city . | קנדה היא גדלה של | רומא עיר איטלקית. |
| tom has a perfect record . | לתום יש רקורד מושלם . | לתום היה רקורד מושלם. |

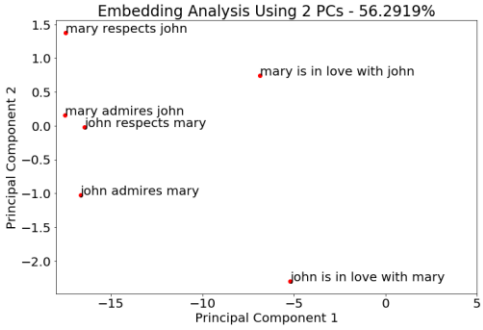
- **Sentence length analysis:**

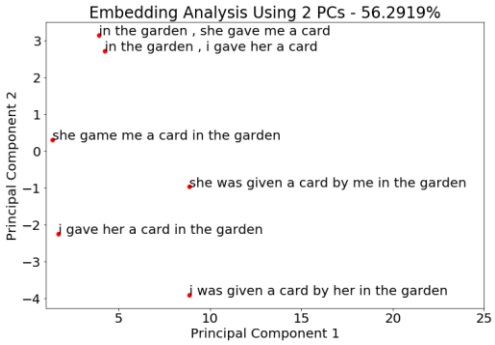



As seen in the above figures, the model obtained the high BLEU scores for the shorter sentences (whether original or translated), as compared to the longer ones as one would expect.

- **Embedding analysis:**




As seen in the figures, even though the embedded distances are not as good as in the original paper, the general relations are kept.

Ofri Masad 201267226
Nir Friedman 305283178

## Results analysis:

In general, our model did not perform very well compared to the original article is very low. This is relatively expected, as we had significantly less data. Moreover, the differences between Hebrew and English makes the translation harder. Differences such as the sentences words orders when adjectives are in the sentence, gender of verbs, and one too many translations etc. Another thing that might've contributed to the poor performance was the use of multiple Hebrew translations to the same English sentence, as seen in the last translation example.

We thought that our model might have better results for shorter sentences, but that wasn't the case. This might be because of differences between the languages – a short sentence in English can be long in Hebrew and vice versa. Secondly, even though English's origin is German, and French is a Latin language, it is known that English was highly influenced by Latin [2] and thus have some basic similarities that English and Hebrew don't.

## Conclusions:

Despite the relatively low results, we believe that this model was able to capture some of the connections between the languages, as seen in the translation examples above, and simply didn't have enough data to learn from.

As stated above, there were 2 main aspects that we hypothesize that gave as these results – the size of the data set and the use of Hebrew as a target language instead of French. We believe that training on more data and tweaking the model (perhaps don't reverse the input) will retain better results. We also believe that choosing a better dataset, that is gender consistent and without multiple translations to the same sentence.

## References:

.1 Papineni, K., et al. *BLEU: a method for automatic evaluation of machine translation*. in *Proceedings of the 40th annual meeting on association for computational linguistics*. 2002. Association for Computational Linguistics.

.2 Baugh, A.C. and T. Cable, *A history of the English language*. 1993: Routledge.