Multilingual Translators

Anurag, Sengupta
asengup2@uci.edu

Aditya, Agarwal
agarwal3@uci.edu

Abstract

In this project we build a multilingual translator which converts an english sentence into french and spanish sentences. The translators are built on top of Euro-Parliment dataset. We build and compare Encoder-Decoder models with Bahndau's Attention and Transformer Model. These models are further enhanced using self trained word-embedding matrix and we compare the speed up in model training as a result of this. We build an inference model and compare the performance of Greedy decoder approach. Finally the models are evaluated using BLEU score and we visualize some user entered sentences.

1 Introduction

The advent of World Wide Web (WWW) has led to a massive explosion in literature and social communication all around the world. According to the Linguistic Society of America [1], as of 2009 we have had 6909 languages across the world. These languages represent the culture and heritage of a particular civilization and they play a pivotal role in voicing socio-economic issues stunting the progress of that region. In major forums like UN, speakers often need to rely on human translators to put their point across to representatives of various countries. These translations if done with a malicious intent on the behalf of the translator can result in bitterness in relationships between countries.

Neural Machine Translation (NMT) has come a long way in providing state of the art translation models. These models are primarily based on Sequence to Sequence model [2]. NMT models have extended basic sequence to sequence models to build conditional language models called as encoder decoder models. Bahndau introduced the concept of attention model [3] to address the issue of long range sequences, where we do not need to consider the complete sequence as it is to the decoder. The words at the decoder decide which word in the encoder is an important parameter. Until recently, this was extended to build self attention model [4] called as transformers which led to translated results much better than any previous models known. In our project, we extend these models to build Parallel Decoder System (PDS) as shown in section 2.2.

Most of the times these models have a huge dependency on the computer resources. Pre-trained word embedding have generally shown a high speed up in training time in these architectures [5]. Glove [6] has been revelation in the field of Natural Language Processing (NLP) with vocabulary and trained vectors in multiple languages. We use this embedding concept to pre-train our sentences into vectors to our embedding layers in the above architecture.

Inference networks on decoder are based on Greedy Search. Bilingual Evaluation Understudy (BLEU) has been used by most language models as a performance parameter [8]. Human evaluation of model plays a big part in evaluating the strength of our model and we explore that for our model using visualizations in NLP.

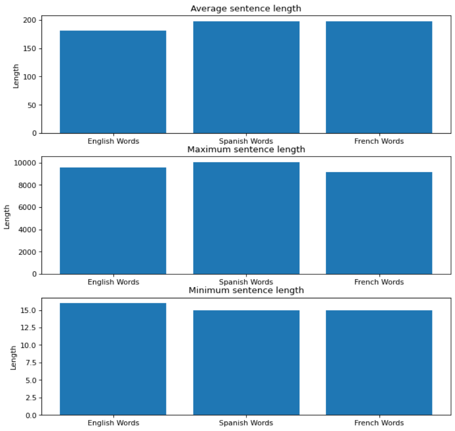


Figure 1: Length of sentence.

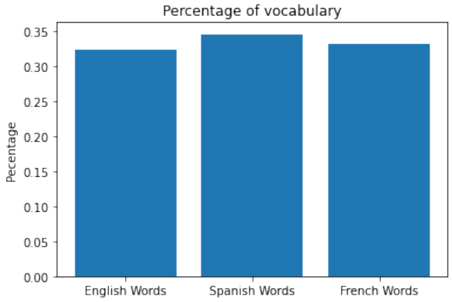


Figure 2: Percentage of vocabulary in the language.

2 Methodology

2.1 Dataset

We use Euro-Parliament dataset [9] and choose sentences in english, french and spanish. These are general conversations in the parliament between the head of states which happened over 15 years (1996-2011). We arrange the conversations in the datasets in such a way that english sentence line directly corresponds to the spanish and french translation.

The number of sentence-pairs in the dataset are 475,834. We represent the average length each sentence, maximum length sentence and minimum length sentence in each language in the figures below.

As shown Figure 1 the average length of a sentence in each of the language is between 175-200 with spanish and french having the highest (200). Although the maximum length sentence is in Spanish and the minimum length sentence is in french Figure 2.

Furthermore, when we explore the vocabulary of each language as a percentage of the total number of words in the vocabulary. We see that spanish and french both have a higher percentage share but on a comparative level all the words are almost equally distributed amongst the languages.

The word clouds for each of the languages also show higher number of occurrences for words which are present in all the languages for instance President in english, Presidente in spanish and Prèsident in french.

For simplicity of our models we consider, 50000 as the size of our complete dataset. We do a train-test split of around 80-20.

2.2 Model

Our model builds a multilingual translator which translates an english sentence into french and spanish sentences. We use parallel decoder architecture enhanced on an encoder-decoder LSTM based attention model with each decoder output corresponding to a particular language. A parallel decoder model would take the same encoded vector from the encoder for each of the decoder. In such a case the encoder circuit which consists of embedding layer and conditional weighted LSTMs would work in synchronously to provide the decoder with the encoded vector.

Figure 3: Word Cloud English.

Figure 4: Word Cloud Spanish.

Figure 5: Word Cloud French.

Although we would have different attention weights corresponding to the language.
We extend the architecture to transformers with multi-headed attention which replaces traditional LSTMs with blocks called as Transformers. We employ teacher forcing [10] as an extra parameter to each of the decoders for both the models.
Since training these models without using pre-trained embedding weight is a resourceful task, we train our own embedding weights using Glove. This model is used to assign a vector to each of our word which is then provided as a pre-trained parameter to the embedding layer. The performance and speed up in training is measured. For the Parallel Seq2Seq model we consider a glove embedding of 50 dimension and for Parallel Transformer we consider a glove embedding of 300 dimension. These different dimensions and hyper-parameters are interchanged so that we get the best training scores.
Finally, we evaluate our model using BLEU score. BLEU is a statistical approach to evaluate sentence translations between languages. Although, very powerful it fails to capture the sentence structure. We evaluate the translated outputs at the decoders using a heatmap for the corresponding words in the input to the corresponding translated word. This provides us with a score which takes care of the structure as well as statistical score for our models.
We train these models for 20 epochs and their performance based on time and computational usage is mentioned in the next section.

3 Results and Analysis

We build four models for multi-lingual translator, and calculate the BLEU scores on test dataset of size 9,000. The below table shows the BLEU scores for the respective Models.

Model	Spanish Score	French Score
Parallel Seq2Seq	27.25	36.97
Parallel Seq2Seq Embedded	37.94	45.39
Parallel Transformer	31.11	28.80
Parallel Transformer Embedded	29.52	28.02

Table 1: BLEU Scores of the models

As we can see from 1, the BLEU score is the highest for the Parallel Seq2Seq Embedded model and the least for the transformer models. This might give us a vague idea about the ability for out models to translate sentences, but doesn't provide a complete picture of the

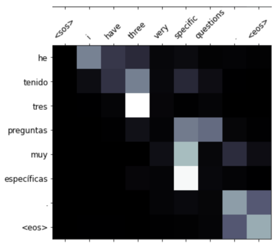


Figure 6: English to Spanish.

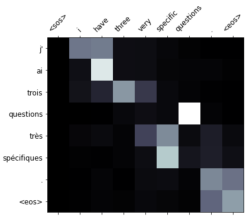
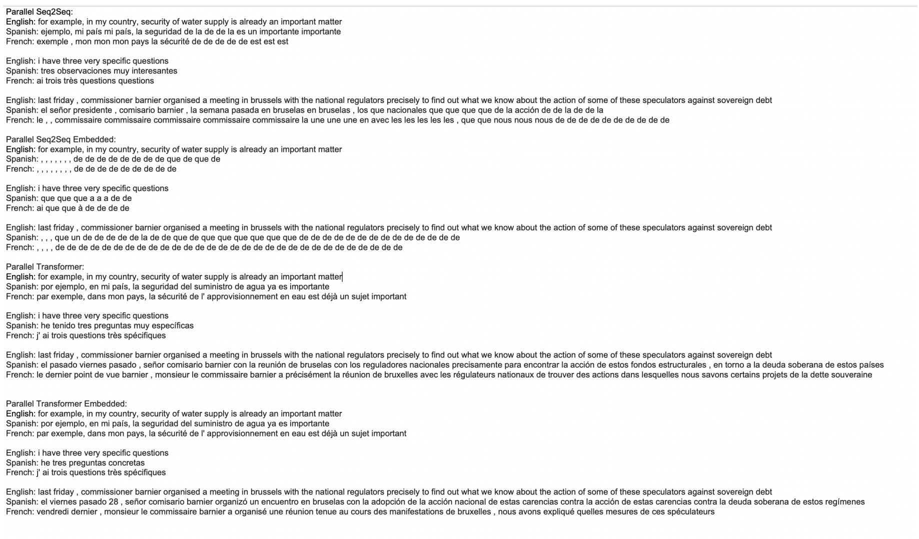


Figure 7: English to French.



```
Parallel Seq2Seq:
English: for example, in my country, security of water supply is already an important matter
Spanish: ejemplo, mi país mi país, la seguridad de la de de la es un importante importante
French: exemple , mon mon mon pays la sécurité de de de de de est est est

English: i have three very specific questions
Spanish: tres observaciones muy interesantes
French: ai trois très questions questions

English: last friday , commissioner barnier organised a meeting in brussels with the national regulators precisely to find out what we know about the action of some of these speculators against sovereign debt
Spanish: el señor presidente , comisario barnier , la semana pasada en bruselas en bruselas , los que nacionales que que que que de la acción de de la de de la
French: le , , commissaire commissaire commissaire commissaire commissaire la une une une en avec les les les les les , que que nous nous nous de de de de de de de de de de

Parallel Seq2Seq Embedded:
English: for example, in my country, security of water supply is already an important matter
Spanish: , , , , , , , de de de de de de de de que de que de
French: , , , , , , , , de de de de de de de de de

English: i have three very specific questions
Spanish: que que que a a a de de
French: ai que que à de de de de

English: last friday , commissioner barnier organised a meeting in brussels with the national regulators precisely to find out what we know about the action of some of these speculators against sovereign debt
Spanish: , , , que un de de de de de la de de que de que que que que que que de de de de de de de de de de de de de de
French: , , , , de de de de de de de de de de de de de de de de de de de de de de de de de de de de de de de de

Parallel Transformer:
English: for example, in my country, security of water supply is already an important matter
Spanish: por ejemplo, en mi país, la seguridad del suministro de agua ya es importante
French: par exemple, dans mon pays, la sécurité de l' approvisionnement en eau est déjà un sujet important

English: i have three very specific questions
Spanish: he tenido tres preguntas muy específicas
French: j' ai trois questions très spécifiques

English: last friday , commissioner barnier organised a meeting in brussels with the national regulators precisely to find out what we know about the action of some of these speculators against sovereign debt
Spanish: el pasado viernes pasado , señor comisario barnier con la reunión de bruselas con los reguladores nacionales precisamente para encontrar la acción de estos fondos estructurales , en torno a la deuda soberana de estos países
French: le dernier point de vue barnier , monsieur le commissaire barnier a précisément la réunion de bruxelles avec les régulateurs nationaux de trouver des actions dans lesquelles nous savons certains projets de la dette souveraine


Parallel Transformer Embedded:
English: for example, in my country, security of water supply is already an important matter
Spanish: por ejemplo, en mi país, la seguridad del suministro de agua ya es importante
French: par exemple, dans mon pays, la sécurité de l' approvisionnement en eau est déjà un sujet important

English: i have three very specific questions
Spanish: he tres preguntas concretas
French: j' ai trois questions très spécifiques

English: last friday , commissioner barnier organised a meeting in brussels with the national regulators precisely to find out what we know about the action of some of these speculators against sovereign debt
Spanish: el viernes pasado 28 , señor comisario barnier organizó un encuentro en bruselas con la adopción de la acción nacional de estas carencias contra la acción de estas carencias contra la deuda soberana de estos regímenes
French: vendredi dernier , monsieur le commissaire barnier a organisé une réunion tenue au cours des manifestations de bruxelles , nous avons expliqué quelles mesures de ces spéculateurs
```

Figure 8: Translation.

models ability to understand and translate sentences in a more comprehensible format. As we can see from the below heat maps, Parallel Transformers perform much better for understanding, capturing the idea of the source sentence and translating it into a better readable sentences. BLEU fails to capture the positional significance and tends to give a very high score for sentences having more number of repetitive words.
On the other hand the heat maps 6 and 7 show a relationship which shows a high correlation between similar words across languages. A brighter color can shows a high correlation in the heat maps. We can observe that in Parallel Transformer the translation in both Spanish and French, but Parallel Seq2Seq translations are incomprehensible as per 8. Although, for smaller length sequence, seq2seq models are able to performance is comparable.
Therefore, relying on BLEU scores can prove detrimental to the task of evaluating the performance of your model, especially when we are looking to build conversational Natural Language Generation models.

The same can be seen in the figure for the loss function figure 9. For Seq2Seq models the loss is very high and might well be under-fitting whereas for Transformers the loss is much lesser and as we can see from the test output, not overfitting either.

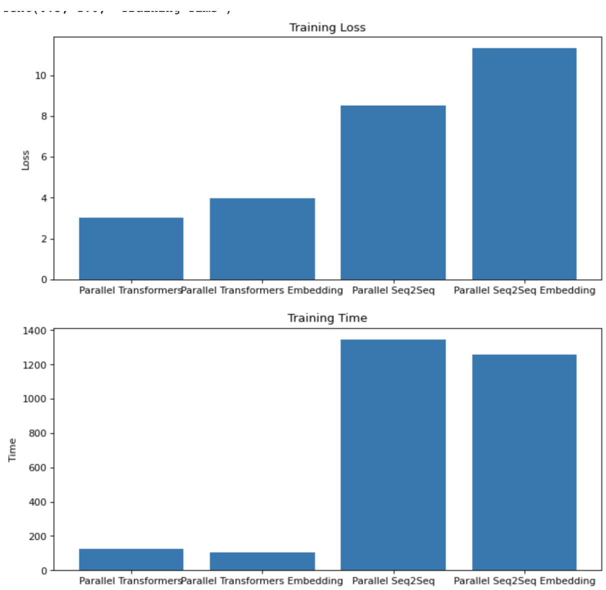


Figure 9: Training Metrics

Furthermore, the Parallel Transformers are much faster than the Parallel Seq2Seq models as seen in the figure 9. Even the training size of batch in the case of Parallel Transformers is 50 as opposed to Parallel Seq2Seq being 20 which can be run on 12GB of Google Colab RAM. In both the cases however, embedding reduces the training time.

4 Conclusion

In this paper, we present 2 different models based on Parallel Transformers and Parallel Seq2Seq models. We compare their relative performance based on pre-trained embedding models. A solution to multi-language decoder using a single encoded source is provided which saves the training computation and cost for different languages. Additionally, we comment on the performance of BLEU and how it can provide false results even insufficient models. A human evaluation for machine translated sentences can provide a better approach. As a further enhancement, we can try Beam Search approach along with various hyper-tuning measures and evaluate the performance of our model.

5 Resources

Code Link
Model Link
Source Data Link
Transformed Data Link

References

[1] S.R. Anderson and S. Anderson. *Languages: A Very Short Introduction.* Very Short Introductions. OUP Oxford, 2012.

[2] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[5] Ye Qi, Devendra Singh Sachan, Matthieu Felix, Sarguna Janani Padmanabhan, and Graham Neubig. When and why are pre-trained word embeddings useful for neural machine translation? *arXiv preprint arXiv:1804.06323*, 2018.

[6] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[7] Cong Duy Vu Hoang, Gholamreza Haffari, and Trevor Cohn. Towards decoding as continuous optimization in neural machine translation. *arXiv preprint arXiv:1701.02854*, 2017.

[8] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[9] Shared task: Machine translation, 2013.

[10] Ronald J Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280, 1989.