LSTMs Versus Transformers in Spanish-English Machine Translation

Tawa Suleman
The Graduate Center, CUNY
LING83600: Language Technology
Fall 2021
tsuleman@gradcenter.cuny.edu

The goal of this project was to explore the FairSeq sequence modeling toolkit for machine translation. My primary interest was to see how well it worked for translation when applied to sentences, as well as to compare the performance of LSTM models and Transformer models.

1 Introduction

Both LSTM and Transformer models have proven to be effective in machine translation, often compared to the performance of statistical machine translation (SMT) models (Ahmadnia and Dorr, 2019; Adebara et al., 2021). I look to investigate how the two models compare to one another in Spanish-English translation.

This paper is organized as follows: Section 2 follows with a description of the data used and data processing completed for this task. Section 3 details the experiment methods. In Section 4 I present the results of this experiment, including a brief error analysis. I include a few remarks in Section 5 before ending with my conclusion in Section 6. Relevant files and scripts can be found on GitHub at https://github.com/tsuleman/langtech-final/ .

2 Data

For my project, I used Spanish-English parallel data from the WMT 2006 translation task[1]. This task uses data from the Europarl Parallel Corpus, which contains sentences from European Parliament proceedings (Koehn, 2005). The data was case-folded with a simple python script[2]. Then, I combined the training and development data[3], and used a random 80-10-10 split to create my training, development, and test sets respectively. There are a total of 204,740 sentences in my data set.

LSTM

Hyperparameters	Values
encoder	bidirectional
dropout	0.2
encoder-embed-dim decoder-embed-dim decoder-out-embed-dim encoder-hidden-size decoder-hidden-size	128
criterion	label smoothed cross entropy
label-smoothing	0.1
learning rate	0.001
optimizer	adam
clip-norm	1
batch-size	50

Table 1: Hyperparameters and values for LSTM models.

3 Methods

I train two models–LSTM and Transformer–on the aforementioned Spanish-English parallel data using the Fairseq sequence modeling toolkit[4]. The two models are trained varying the number of updates twice: once at 4,000 updates and again at 8,000 updates. This is to investigate the improvements models exhibit as update number increases.

3.1 LSTM

I trained a bidirectional LSTM, using the Adam optimizer. More parameters for the LSTM model can be found in Table 1.

[1] https://www.statmt.org/wmt06/shared-task/
[2] see: fairseq_prepreprocessing.py
[3] see: data_compile.py in langtech-final repo
[4] https://github.com/facebookresearch/fairseq

Transformer	
Hyperparameters	**Values**
optimizer	adam
adam-betas	0.9,0.98
criterion	label-smoothed cross entropy
label smoothing	0.1
dropout	0.2
clip-norm	1
learning rate	0.001
learning rate scheduler	inverse square root
warmup-init-lr	1E-7
encoder layers	4
encoder attention heads	4
encoder-embed-dim	128
encoder normalize	before
decoder layers	4
decoder attention heads	4
decoder-embed-dim	128
decoder normalize	before
activation function	relu

Table 2: Hyperparameters and values for Transformer models.

3.2 Transformer

My Transformer model also uses the Adam optimizer, with 4 encoder and decoder layers each. Table 2 displays more parameters for the Transformer model.

4 Evaluation

The models' performance was evaluated by computing the achieved BLEU scores (Papineni et al., 2002). This was done using the BLEU Score module of the NLTK (Bird et al., 2009) Translate subpackage[5]. The BLEU scores were calculated for 1-grams, 2-grams, 3-grams, 4-grams, and BLEU-4.

4.1 Results

Overall, the models achieved the best BLEU scores when scored with respect to unigrams. The scores drop drastically–by more than half in each case–from unigram to bigram scores. We continue to see a decrease, though not as drastic, from bigram to trigram scores and trigram to 4-gram. The scores achieved by the models can be found in Table 3.

While both the LSTM and Trasformer models' translations improved when the number of updates increased, the Transformer model displayed less of an increase. The improvement shown by the LSTM model ranges from about 2.7 to 6. This contrasts the Transformer's range of improvement of about 1 to 2.55.

The Transformer models outperforms the LSTM for all respective BLEU scores, but the LSTM at 8,000 updates marginally outperforms the Transformer at 4,000 updates for 3-gram, 4-gram and BLEU-4 scores.

4.2 Error Analysis

The following is a brief error analysis including some of the types of errors I noticed in the model translations.

Pronouns & Gender proved to be a challenge for the models, especially with a lower number of updates and when a sentence did not contain any overt evidence of the pronoun's reference (See Example 1 in Table 4). With the exception of the Transformer at 4,000 updates, the models were able to find the appropriate pronoun when there was an overt reference such as *comisario* in the example below:

Original: agradezco al comisario su respuesta .
Gold: i am grateful to the commissioner for *his* answer .
LSTM (4K) output: i thank the commissioner for *his* answer .
LSTM (8K) output: i thank the commissioner for *his* answer .
Transformer (4K) output: i thank the commissioner .
Transformer (8K) output: i thank the commissioner for *his* answer .

However, the models did not always perform well with just any sort of overt evidence of gender, reference, etc. As shown below, the models did not translate the feminine direct object *la* 'her' of the present perfect phrase *la ha visto* 'has seen her'.

Original: nadie la ha visto desde entonces .
Gold: no one has seen her since .
LSTM (4K) output: nobody has been made in the case .
LSTM (8K) output: nobody has been seen since then .
Transformer (4K) output: no one has been done .

[5] https://www.nltk.org/_modules/nltk/translate/bleu_score.html

LSTM					
	1-gram	2-gram	3-gram	4-gram	BLEU-4
4K updates	40.77	17.95	9.09	4.70	13.27
8K updates	46.64	22.88	12.93	7.46	17.91
Transformer					
	1-gram	2-gram	3-gram	4-gram	BLEU-4
4K updates	48.18	23.03	12.73	7.19	17.85
8K updates	50.73	25.35	14.55	8.57	20.01

Table 3: BLEU scores achieved by LSTM and Transformer models at 4,000 updates and 8,000 updates.

Transformer (8K) output: nobody has seen since then .

Even with evidence of gender (e.g. *la presidenta*), the models sometimes use incorrect pronouns. Did someone say gender bias?

Original: agradezco a la presidenta en ejercicio su respuesta .
Gold: i thank the president-in-office for *her* reply .
LSTM (4K) output: i thank the president-in-office of *his* answer .
LSTM (8K) output: i would like to thank the president-in-office of *his* answer .
Transformer (4K) output: i thank the president-in-office for *his* reply .
Transformer (8K) output: i thank the president-in-office for *his* reply .

Longer Sentences. Oftentimes the models omitted information when translating longer sentences, as in Example 2 in Table 4.

Repetition. The translations also would have repeated phrases. This was most often noticed in the models at 4,000 updates (e.g. The LSTM translation of Example 2, the Transformer translation of Example 4).

Meaning. There were several instances of the translations not capturing the message of the original sentence. This was either by having omitted words, using words of opposite or different nuance, or phrases that were simply not equivalent.

Original: lo ha hecho con modestia , y eso le honra .
Gold: you have done so modestly , and that is to your credit .
LSTM (4K) output: it has been said , and that is a mistake .
LSTM (8K) output: it has been done with us , and that is a good thing .
Transformer (4K) output: that has been done , and that has happened .
Transformer (8K) output: he has done this , and it has done it .

5 Challenges, Future Work & Other Remarks

The initial models I trained had a much lower number of 400 updates. The produced translations held no substance as the model would only print the word 'the'. As I increased the update number I noticed how much more coherent the translations became. This made me curious to see how much the amount of updates can improve models. Given more time and resources, I would be interested in comparing the results of models with other amounts of updates.

For this task I also needed to figure out the most appropriate evaluation method. I originally intended to use both BLEU scores and word error rate (WER). However, after reviewing the translations produced, I did not feel WER would provide a good evaluation. Even when translations are accurate, they often are not phrased exactly as the gold translations. Variations of WER, such as Translation Edit Rate (TER) (Snover et al., 2006), may provide more insight than WER in this case, but for this task I chose to only report the BLEU scores.

After computing the BLEU scores using the NLTK package I noticed the scores provided by the Fairseq generation. These scores differ quite a bit (see: Table 5). I would be interested in learning what may cause discrepancies between the scoring methods.

For further investigation, I think it would be fruitful to explore more architectures, potentially those pretrained and built for multilingual translation. In addition, I think it could be helpful to implement BPE tokenization on the data. I would also be curious to see the results of English-Spanish translation, and to investigate the effects of varying other

Gold	Model	No. of Updates	Model Output
(1) we appreciate her commitment	LSTM	4K	we welcome its compromise
		8K	we welcome your compromise
	Transformer	4K	in his compromise
		8K	we appreciate your commitment
(2) it is true that the european union is often accused of promising aid to countries struck by disaster or catastrophe and then taking so long to pay it , so long to keep its financial promises that either it creates the impression that it regrets having promised it in the first place or , when it finally does pay , it is already too late .	LSTM	4K	it is true that it is true that there is a lot of the european union 's aid to the european union , but if it would be able to be able to be able to be able to be able to be able to be able to be able to be able to deal with the money , or if the money has already been too late .
		8K	it is true that the european union is often aware of the european union that although it calls for aid to countries affected by different problems or disasters , which has been taken too late .
	Transformer	4K	it is true that the european union is often interested in the european union , even though , or the various countries have already been held in these countries , and it is too late .
		8K	it is true that , in fact , the european union , even if it comes to aid to countries affected by different disasters or disasters , it is too late .
(3) i believe it is against the european convention on human rights .	LSTM	4K	i believe that this will be against the european convention against human rights .
		8K	i believe that this is against the european convention on human rights .
	Transformer	4K	i believe that this is against human rights .
		8K	i believe that this is against the european human rights convention .
(4) future activity must involve real cooperation and negotiation with the recipients of aid .	LSTM	4K	the future of the future is that the future of the european cooperation and a whole of aid .
		8K	the future activities must include a real cooperation and real negotiations with those who receive aid .
	Transformer	4K	the future needs to include cooperation and cooperation .
		8K	the future activity must include a real and real cooperation with which they receive aid .

Table 4: Examples of translations produced by the models with 4,000 and 8,000 updates.

LSTM			
	1-gram	2-gram	3-gram
4K updates	42.4	18.4	9.4
8K updates	55.7	29.9	18.1
	4-gram	BLEU-4	
4K updates	4.9	13.80	
8K updates	11.2	23.85	
Transformer			
	1-gram	2-gram	3-gram
4K updates	54.8	24.6	12.9
8K updates	61.7	34.3	21.4
	4-gram	BLEU-4	
4K updates	6.9	15.48	
8K updates	13.7	25.99	

Table 5: BLEU scores calculated by Fairseq.

parameters in the models.

6 Conclusion

With this task I gained insight into how the performance of machine translation models can be affected by their parameters, in particular the number of updates. While the result is not completely surprising, this has made me curious as to what determines the optimal number of updates for a model, and when increasing updates no longer improves a model.

References

Ife Adebara, Muhammad Abdul-Mageed, and Miikka Silfverberg. 2021. Translating the unseen? yorùbá → english MT in low-resource, morphologically-unmarked settings. *CoRR*, abs/2103.04225.

Benyamin Ahmadnia and Bonnie Dorr. 2019. Enhancing phrase-based statistical machine translation by learning phrase representations using long short-term memory network. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 25–32, Varna, Bulgaria. INCOMA Ltd.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.