DeepBLEU: A Neural Network Approach to Machine Translation Evaluation

Austin Narcomey
aon2@stanford.edu

Andrew Narcomey
aon1@stanford.edu

Khalid Ahmad
kahmad@stanford.edu

Objective

- BLEU has been a key tool in evaluation of machine translation, as it has allowed for rapid development of new neural network models, but comes at the expense of lacking true language understanding, and alternative methods pool complex external resources to make computationally expensive evaluations
- We aim to utilize neural network architectures and test attention, given its success in Machine Translation and Natural Language Inference, and in Transformer models, to encode inputs

Our Goal:

Create an efficient, scalable evaluation system using LSTM neural networks and attention that optimizes correlation to human judgment, the gold standard for rating machine translation [2]

Data

WMT data: a collection of human-evaluated machine translations, scored by ranking

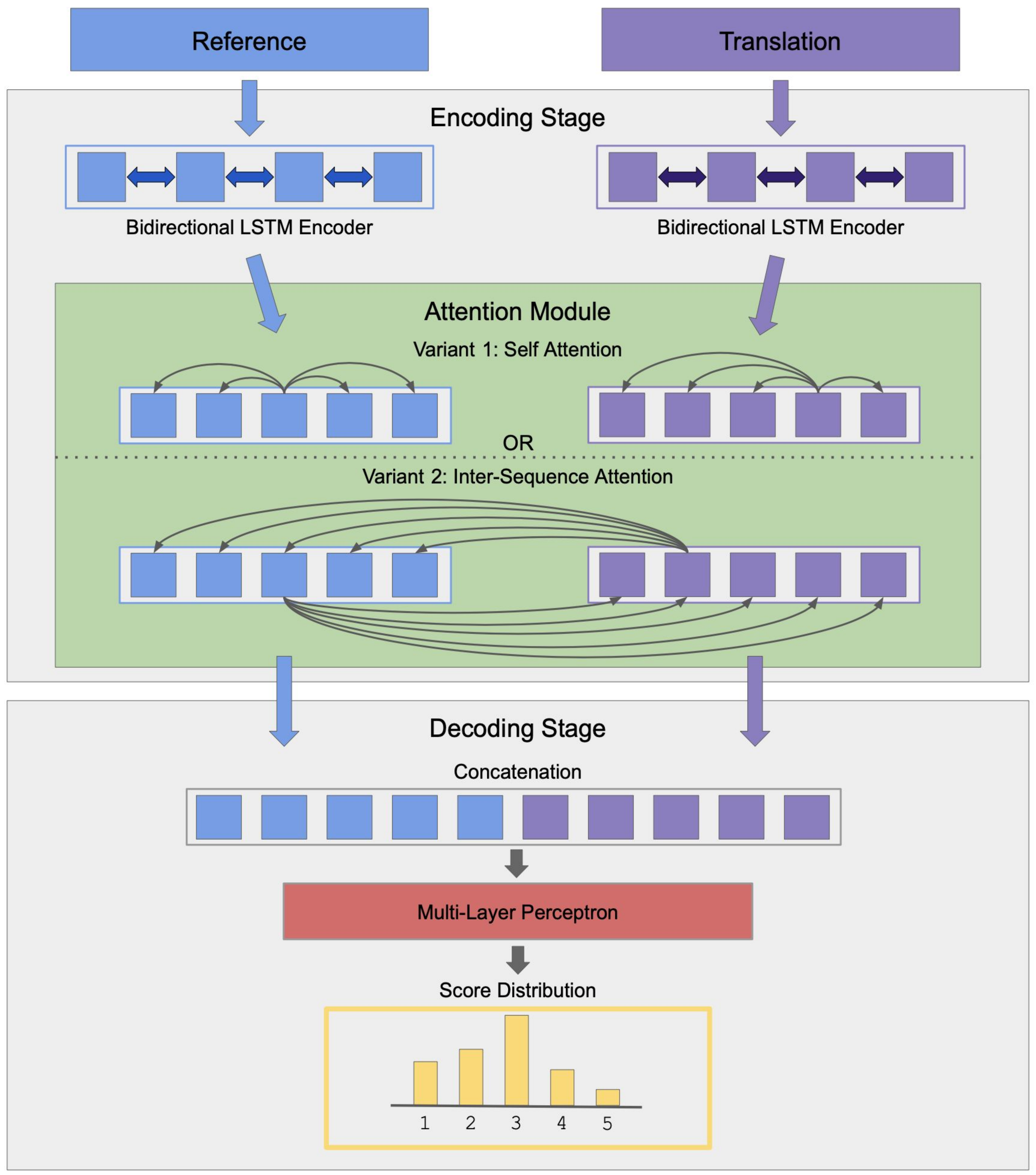
- Each example includes reference translation, MT translation, and a rating from 1 (worst) to 5 (best), averaged over multiple reviewers
- ~ 50,500 training examples in WMT-LARGE

Example With Rating 1 (Reference → Translation):
According to an ancient legend, pagan gods used to live on the mountain → According to the ancient legend reportedly lived on a mountain top pohanští gods

Evaluation Metrics

Pearson & Spearman Correlations - Measures linear and monotonic relationship, respectively, between distributions. Used to compare scores of our model and scores provided by human evaluators in WMT data

Models



Reference
Translation
Encoding Stage
Bidirectional LSTM Encoder
Bidirectional LSTM Encoder
Attention Module
Variant 1: Self Attention
OR
Variant 2: Inter-Sequence Attention
Decoding Stage
Concatenation
Multi-Layer Perceptron
Score Distribution

- Layers were combined to approximate capabilities of a Transformer that decodes concatenated encoded translation and reference translations into a representation of their similarity
- **Inter-Sequence Attention -** Captures how states in one sequence relate to the states in another, and vice versa
- **Self Attention -** Captures how the meaning of words in a given sequence are influenced by other words in the sequence
- **MLP -** Used as a feed-forward network to take encodings and extract from them a representation of their semantic similarity

Results & Discussion

Table 1: Models Trained On WMT+SICK Data

		NIST	BLEU	Vanilla	Stacked2	MLP2	Inter-Seq + MLP2	Self + MLP2	Inter-Seq + MLP3	Self + MLP3
Train	Pearson	31.11	37.32	50.16	**53.67**	49.19	50.10	51.8	48.66	48.92
	Spearman	30.36	38.02	48.77	**53.08**	50.33	50.99	52.47	49.73	50.11
Dev	Pearson	18.89	28.94	34.60	34.50	34.24	34.62	34.56	34.10	**35.14**
	Spearman	17.69	27.30	33.76	33.27	33.46	33.80	33.38	33.49	**34.33**
Test	Pearson	19.14	26.76	38.66	38.28	39.31	39.47	39.78	38.75	**40.22**
	Spearman	18.48	25.40	38.02	37.45	38.33	38.51	38.23	37.65	**39.19**

Table 2: Models Trained On WMT-LARGE Data

		NIST	BLEU	Vanilla	Stacked2	MLP2	Inter-Seq + MLP2	Self + MLP2	Inter-Seq + MLP3	Self + MLP3
Train	Pearson	16.37	28.87	**49.39**	49.25	39.15	41.40	37.93	43.58	43.48
	Spearman	15.97	27.25	**48.55**	48.33	38.14	40.60	36.93	42.82	42.54
Dev	Pearson	18.90	28.94	35.12	**37.19**	37.07	35.30	36.73	36.60	36.71
	Spearman	17.69	27.30	33.64	**35.48**	35.82	33.70	35.40	34.99	35.32
Test	Pearson	19.14	26.76	37.92	36.60	**38.03**	37.96	37.85	36.77	36.82
	Spearman	18.48	25.40	36.86	35.88	**37.11**	36.50	36.18	35.63	35.82

BLEU & NIST - Closely predict human ratings when reference and translation sentences have long segments which are near identical

Vanilla & Stacked2 - Simplest models with high train, low test correlation

MLP2 - Outperformed all other models on WMT-LARGE, showing a simpler model could better fit a less noisy dataset (without addition of SICK)

MLP2 vs. MLP3 - Mixed results due to limited dataset size

Self Attention - Scored higher than Inter-Sequence, producing better encodings with shorter training than Inter-Sequence

Conclusion

- The use of a MLP and Self Attention provided highest correlations
- Our models outperform NIST and BLEU correlations, showing the promise of neural network architectures for evaluation that measure closer to human judgment without any heavy external resources
- Future work would be to collect more data to be able to train models longer and deeper, and to be able to test inter-sequence attention more thoroughly

References

[1] Marelli, Marco, et al. "Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment." *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*. 2014.

[2] Papineni, Kishore, et al. "BLEU: a method for automatic evaluation of machine translation." *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002.