Shakespeare - English Sequence to Sequence Modeling

Billy Watson, Morris Kracier, Riley Scott

Problem

Data

- **8 aligned** plays
  - GIZA++ and Moses SMT systems
- **10,365** sentence pairs
- 9,004 source words
- 7,497 target words
- 233,282 total words
- Taken from **eNotes**

Play	Line Count
Hamlet	2,010
Julius Caesar	1,201
Macbeth	1,085
Merchant of Venice	831
Midsummer Night's Dream	833
Othello	1,893
Romeo and Juliet	1,743
Tempest	769
TOTAL	**10,365**

Preprocessing

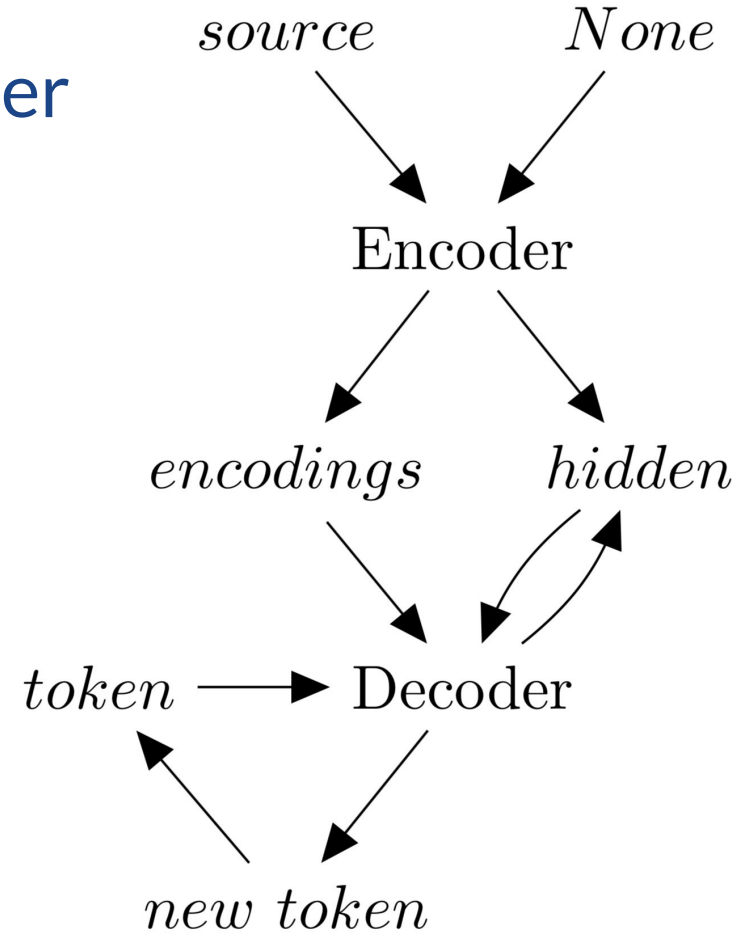
- **Replace proper nouns** with a special token
- Use NLTK **tokenization** to split sentences, contractions, etc
- Use Start of Sentence (SOS) Tokens to begin phrases
- Use End of Sentence (EOS) Tokens to signal end of sentence
- **Pad target sentences** with EOS for **batching** by source sentence length

Data Split

Split	Line Count	Percentage
Train	9,069	87.5 %
Dev	1,036	10.0 %
Test	260	2.5 %
TOTAL	**10,365**	**100 %**

Models

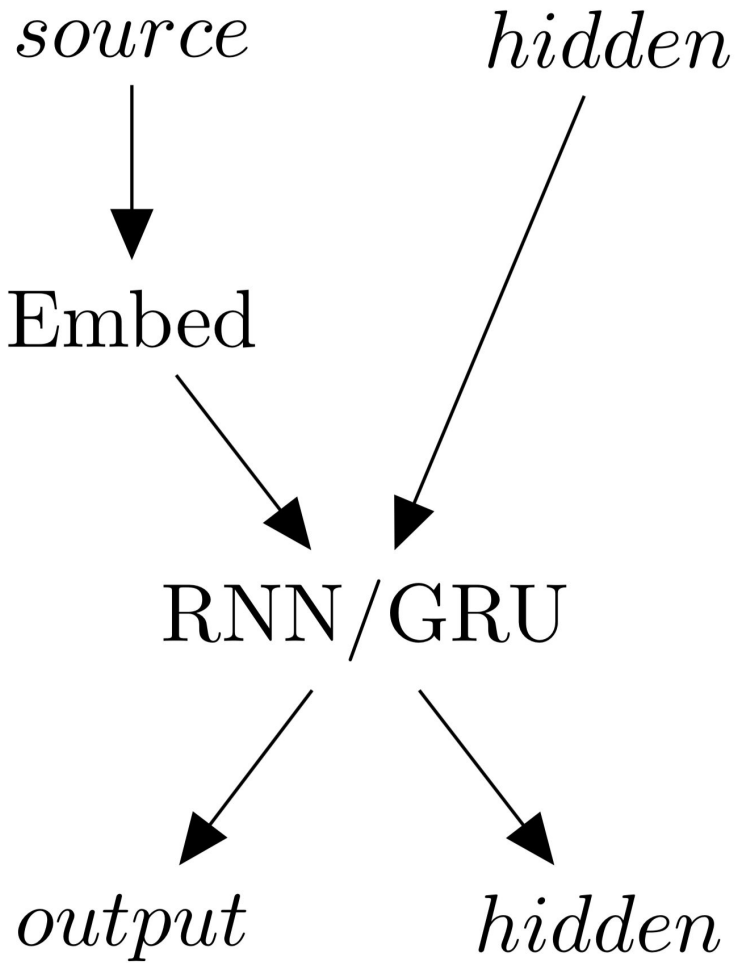
Encoder-Decoder



Encoder

- **Inference Network**

$$\vec{h} = \begin{cases} \mathrm{RNN}(W_e[\vec{x}]) \\ \mathrm{GRU}(W_e[\vec{x}]) \\ \mathrm{BiGRU}(W_e[\vec{x}]) \end{cases}$$



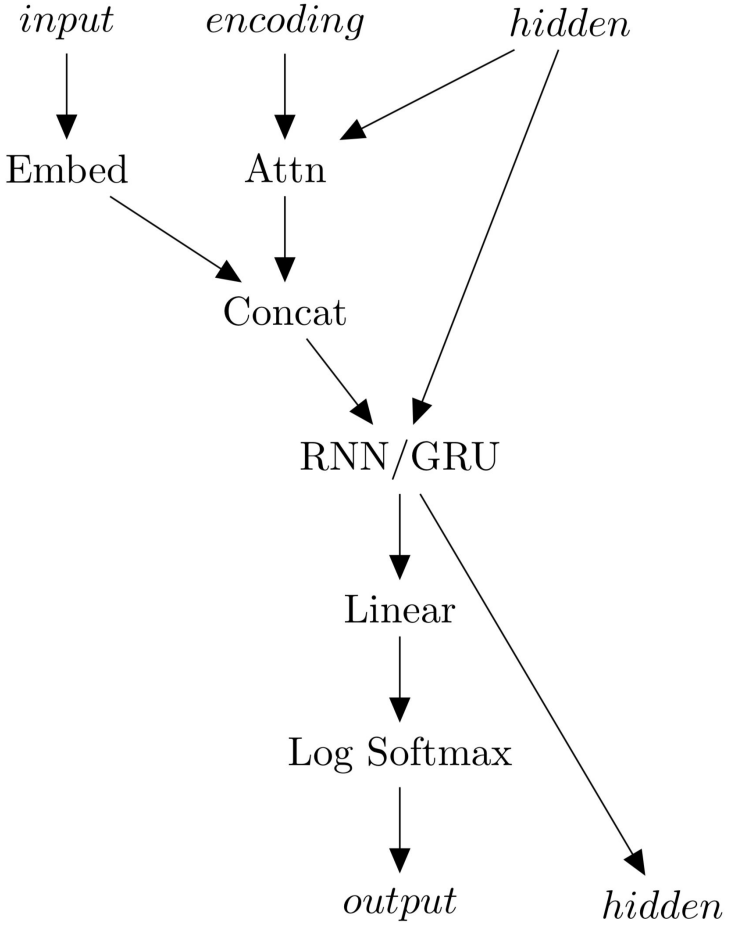
Decoder

- **Generative Network**

$$\tau_i = \begin{cases} W_e[t_i] & \textit{No Attention} \\ W_e[t_i] \parallel c_i & \textit{With Attention} \end{cases}$$

$$\mathbf{y} = W_d \cdot \begin{cases} \mathrm{RNN}(\tau_i,\ s_{i-1}) \\ \mathrm{GRU}(\tau_i,\ s_{i-1}) \end{cases}$$

$$\log \sigma(\mathbf{y}) = \log \frac{\exp(y_i)}{\sum_j \exp(y_j)}$$

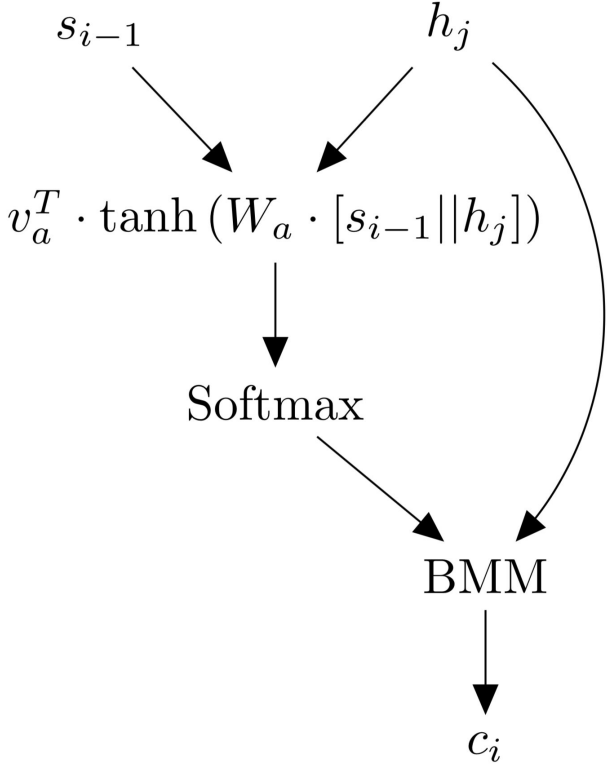


Attention Mechanisms - Concatenation

$$score(s_{i-1}, h_j) = v_a^T \cdot \tanh(W_a \cdot [s_{i-1} \| h_j])$$

$$a(s_{i-1}, h_j) = \frac{\exp(score(s_{i-1}, h_j))}{\sum_{j'} \exp(score(s_{i-1}, h_{j'}))}$$

$$c_i = \sum_{j'} a(s_{i-1}, h_{j'}) \cdot h_{j'}$$

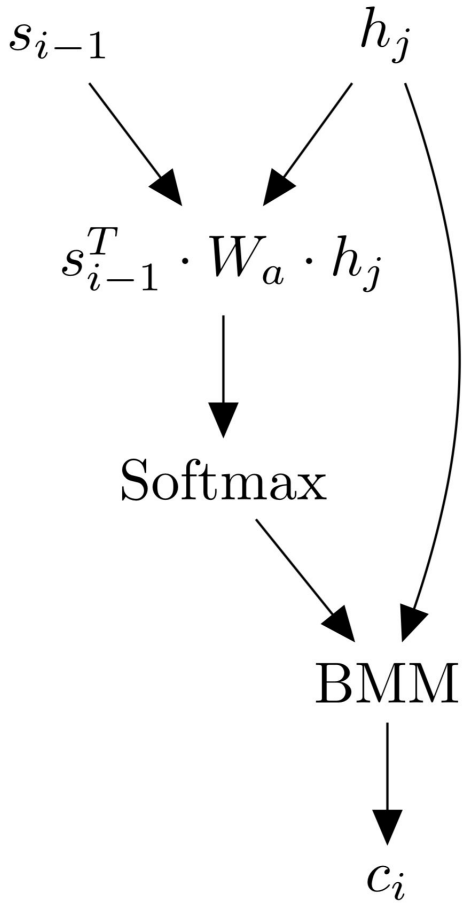


Attention Mechanisms - General

$$score(s_{i-1}, h_j) = s_{i-1}^T \cdot W_a \cdot h_j$$

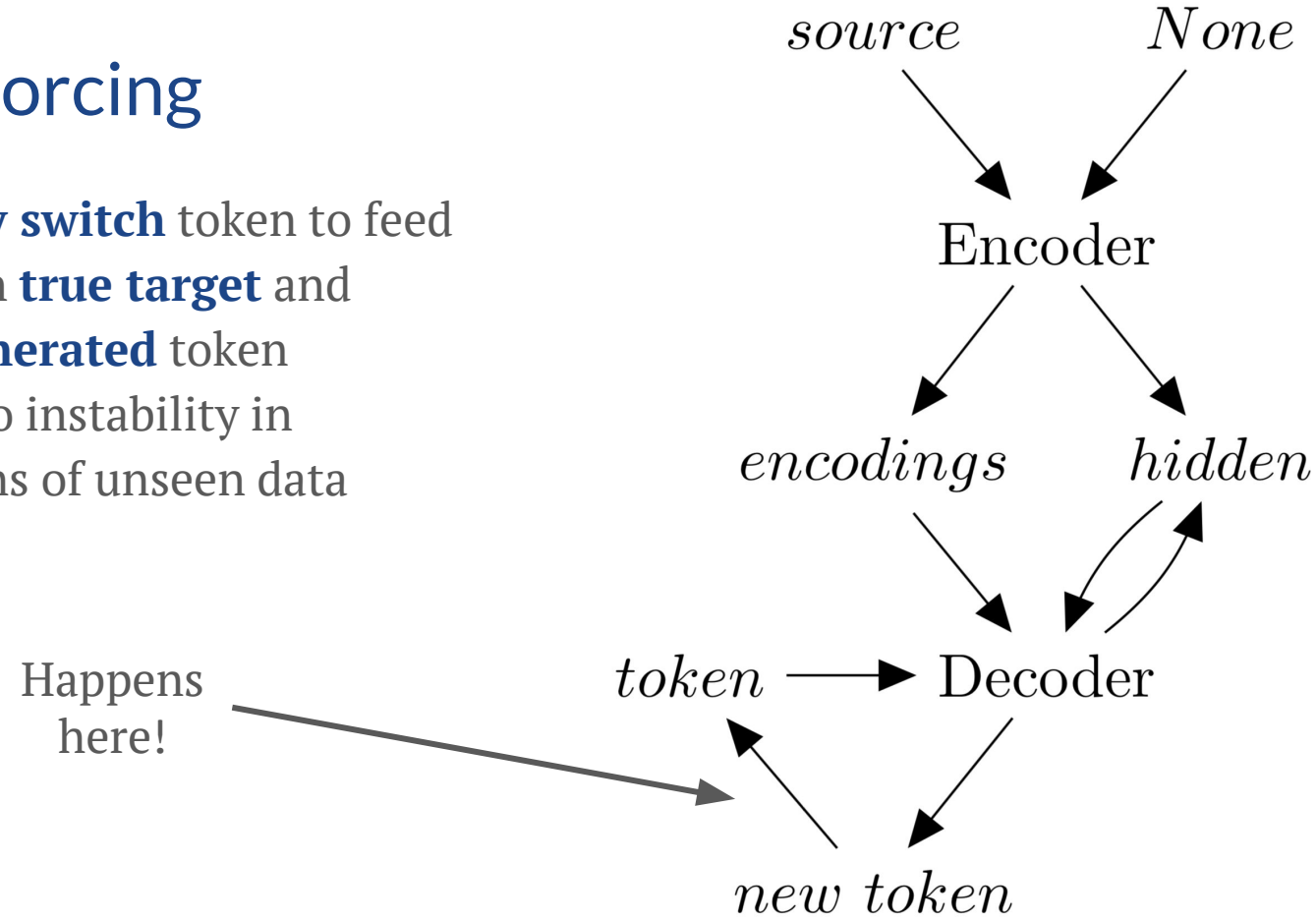
$$a(s_{i-1}, h_j) = \frac{\exp(score(s_{i-1}, h_j))}{\sum_{j'} \exp(score(s_{i-1}, h_{j'}))}$$

$$c_i = \sum_{j'} a(s_{i-1}, h_{j'}) \cdot h_{j'}$$



Teacher Forcing

- **Randomly switch** token to feed in between **true target** and **model generated** token
- Can lead to instability in translations of unseen data



Results

Experiments

Encoder	Decoder	Attention	Teacher Forcing	Target BLEU
RNN	RNN	-	-	0.0000
GRU	GRU	-	-	0.0238
BiGRU	GRU	-	-	0.0336
BiGRU	GRU	General	-	0.0333
BiGRU	GRU	General	50 %	0.0330
BiGRU	GRU	General	100 %	0.0395
BiGRU	GRU	Concat	-	0.0842
BiGRU	GRU	Concat	50 %	0.1349
BiGRU	GRU	Concat	100 %	0.1042
BiGRU	**GRU**	**Concat**	**100 %**	**0.1450**

Sample Translation 1

Model	Sentence
Source	**Double, double, toil and trouble; fire burn and cauldron bubble.**
RNN	Damned as you are, you have cast the spell, the , (x490)
GRU	Double, double, toil and;; and burn burn burn burn caldron bubble.
BiGRU	Double, double, toil, and;,, burn,, bubble bubble.
BiGRU + General + 100% TF	Double, double, toil and trouble; fire, burn; and caldron, bubble
BiGRU + Concat + 100% TF	Double, double, toil and trouble; fire, burn; and caldron, bubble.
Best Model	**Double, double, toil and trouble; fire, burn; and caldron, bubble.**
Target	**Double, double, toil and trouble; fire, burn; and caldron, bubble.**

Sample Translation 2

Model	Sentence
Source	**What say you, propn?**
RNN	What, you,
GRU	What do you, propn propn?
BiGRU	What do you say, propn?
BiGRU + General + 100% TF	What do you say, propn?
BiGRU + Concat + 100% TF	What do you say, propn propn?
Best Model	**What do you say, propn?**
Target	**What do you say, propn?**

Sample Translation 3

Model	Sentence
Source	**Propn thou not see her paddle with the palm of his hand?**
RNN	don’t
GRU	Will n’t storms; his his his the??
BiGRU	Propn n’t with you like the the the the hand hand?
BiGRU + General + 100% TF	Might don’t see the his dying hand, might see about the?
BiGRU + Concat + 100% TF	Don’t you see her for the scripture of his body?
Best Model	**Didn’t go see her close with the same of his hand?**
Target	**Didn’t you see her play with the palm of his hand?**

Future Work

- **Batch Annealing**
  - **Adaptive batch sizes**, starting small and gradually increasing in size
- **English to Shakespeare Modeling**
  - Reverse training to turn English sentences with a Shakespeare twist
- **Local Attention with Window**
  - Implement local attention to focus model on **local** instead of the **global** context
- **Data Preprocessing**
  - Using more SMT toolkits to perform better preprocessing
  - Using **BPE tokenization** to reduce vocabulary size
  - Sample train/dev/test to **reduce unknown tokens** in test

Questions?

References

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[2] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. *CoRR*, abs/1506.03099, 2015.

[3] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[4] A. M. Lamb, A. G. A. P. GOYAL, Y. Zhang, S. Zhang, A. C. Courville, and Y. Bengio. Professor forcing: A new algorithm for training recurrent networks. In *Advances In Neural Information Processing Systems*, pages 4601–4609, 2016.

References

[5] M.-T. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[6] M. Morishita, Y. Oda, G. Neubig, K. Yoshino, K. Sudoh, and S. Nakamura. An empirical study of mini-batch creation strategies for neural machine translation. *arXiv preprint arXiv:1706.05765*, 2017.

[7] K. Papineni, S. Roukos, T. Ward, and W.-J. 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.

[8] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[9] W. Xu, A. Ritter, B. Dolan, R. Grishman, and C. Cherry. Paraphrasing for style. In *COLING*, pages 2899–2914, 2012.