

Shakespeare - English

Sequence to Sequence Modeling

Billy Watson, Morris Kracier, Riley Scott

Problem

Data

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

Sample Data Pairs

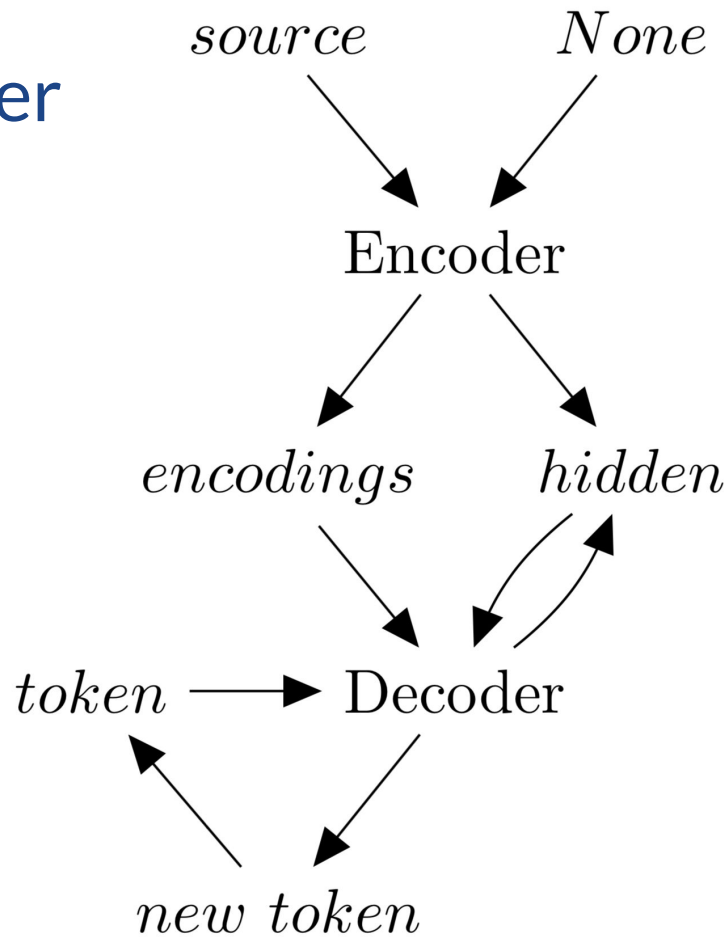
Shakespeare Sentence	Modern Sentence
take me with you , take me with you , wife .	catch me , catch me , wife .
that 's fouler .	that 's even more evil .
then weep no more .	so stop crying .
of propn , a n't please your mastership .	of propn , if it pleases you , sir .
how ill this taper burns !	how badly this candle burns !
the charm 's wound up .	the charm s going to bring things to a head .
before the time be out ?	before the debt is paid ?

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

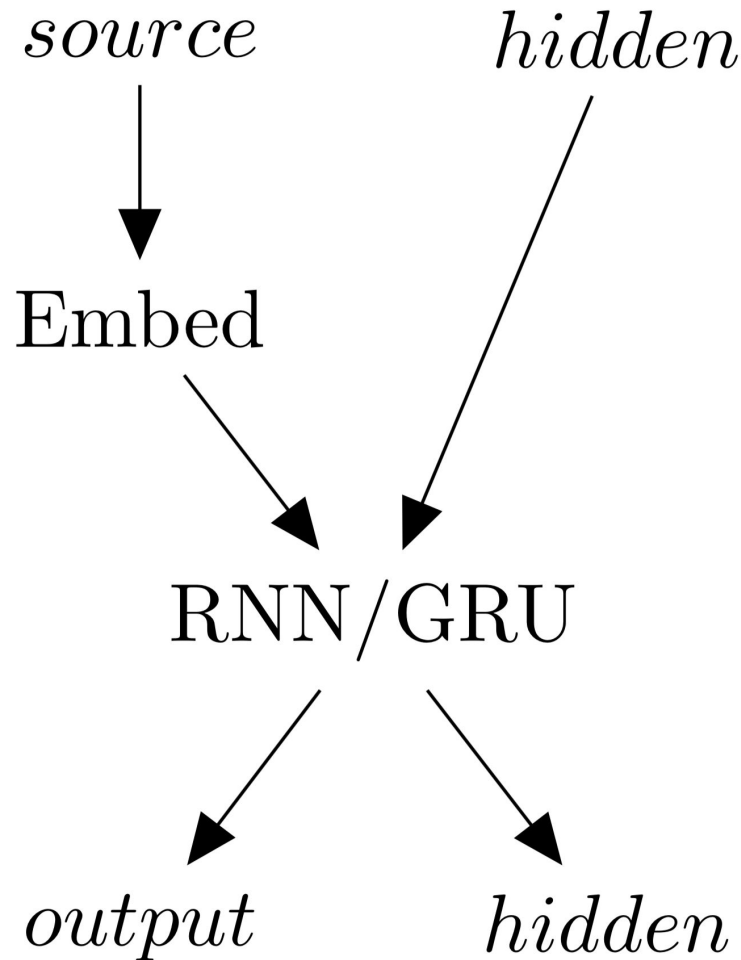
Encoder-Decoder



Encoder

- **Inference Network**

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

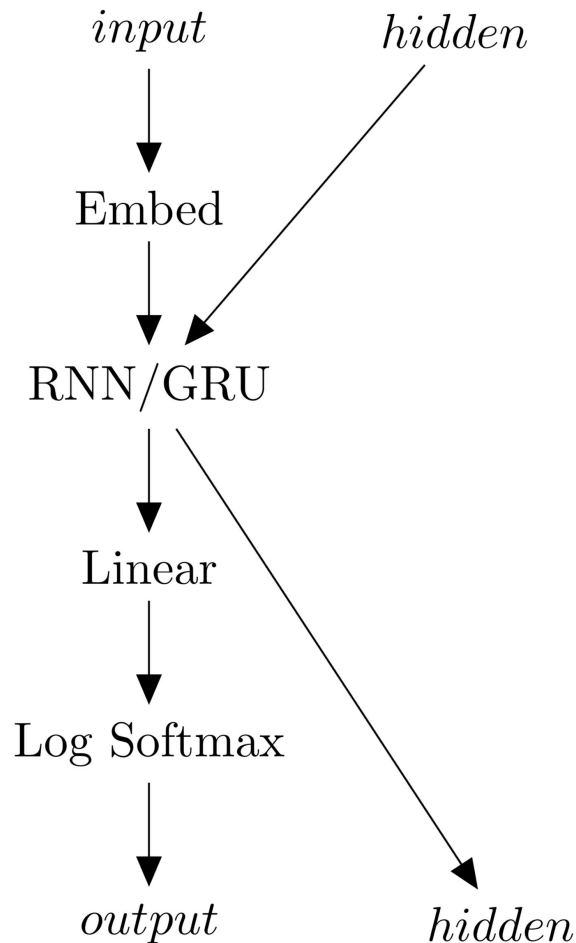


Decoder - Without Attention

- **Generative Network**

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

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

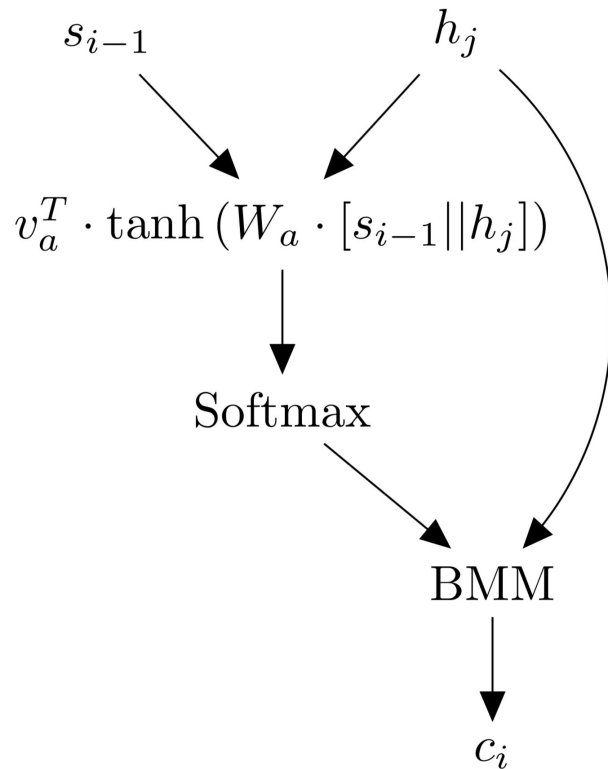


Attention Mechanisms - Concatenation

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

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

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

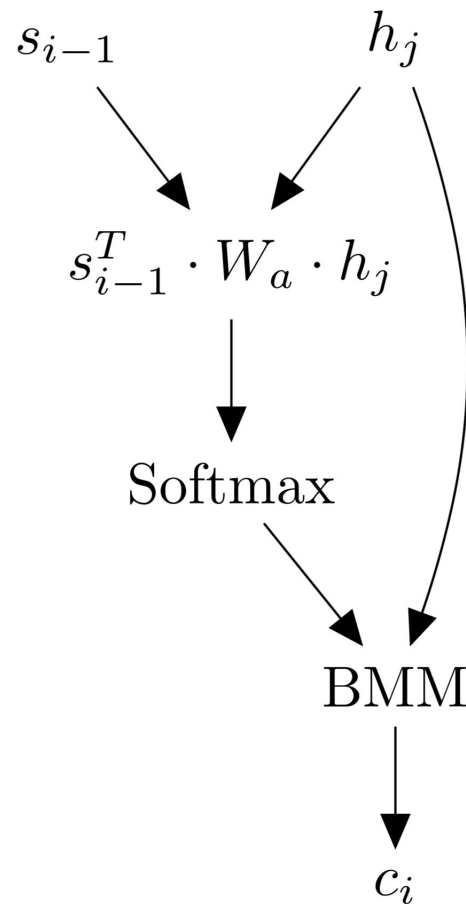


Attention Mechanisms - General

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

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

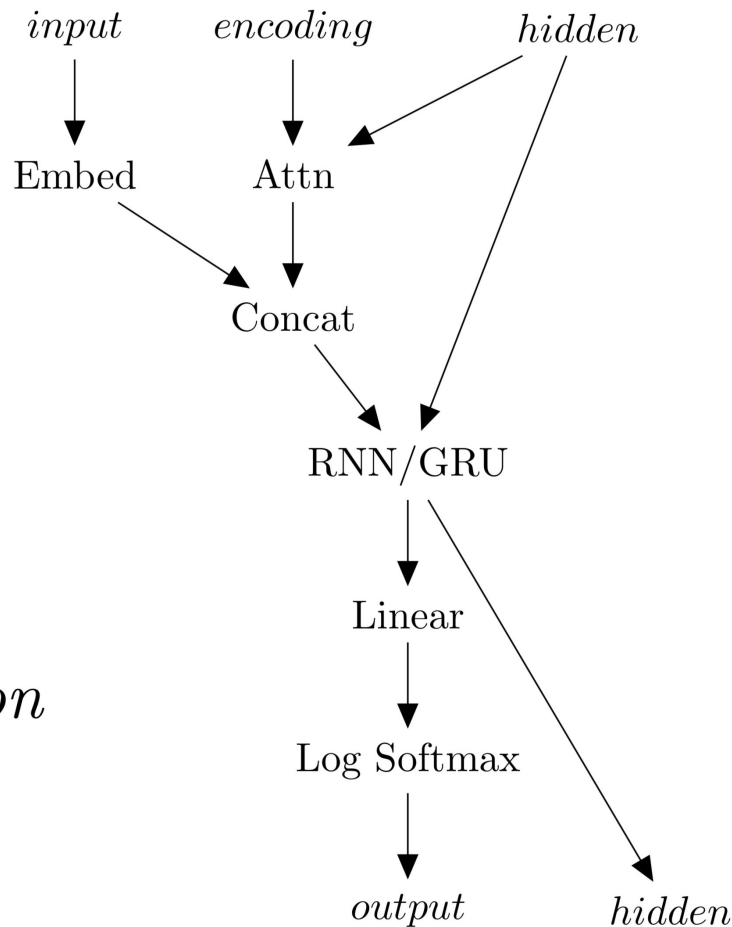
$$c_i = \sum_{j'} a(s_{i-1}, h_{j'}) \cdot h_{j'}$$



Decoder - With Attention

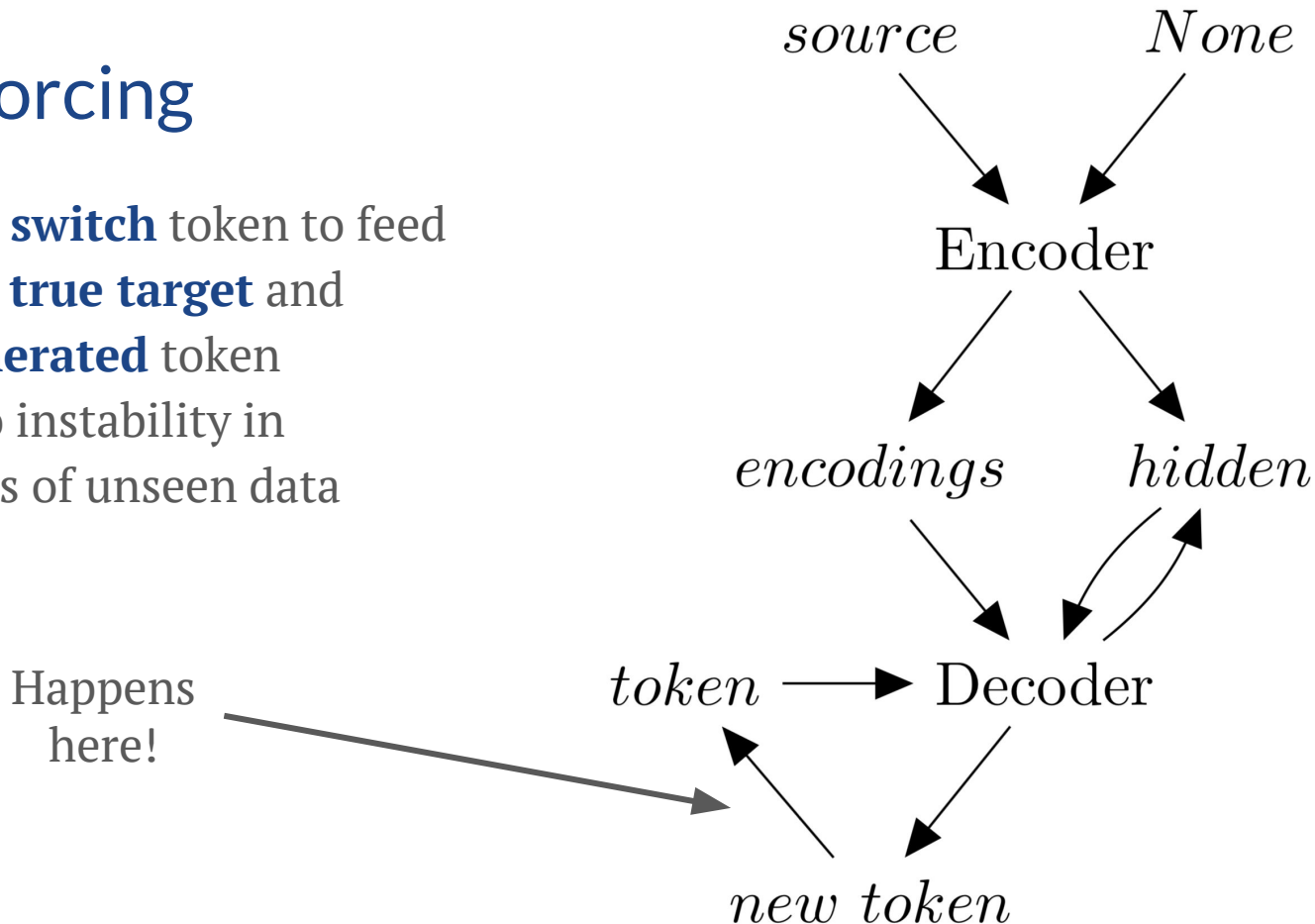
- **Concatenate** the input token embedding with the context vector before the recurrent layer

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



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	Tonight, my lord?
RNN	What you you,
GRU	How, my lord?
BiGRU	My, my lord?
BiGRU + General + 100% TF	Brother, my lord?
BiGRU + Concat + 100% TF	Tonight, my lord?
Best Model	Tonight, my lord?
Target	Tonight, my lord?

Sample Translation 3

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 4

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?

Sample Translation 5

Model	Sentence
Source	Some wine, ho?
RNN	What, ho!
BiGRU	And, ho!
BiGRU + General + 100% TF	Some three, a blessing in like a!
BiGRU + Concat + 50% TF	Let me fight!
BiGRU + Concat + 100% TF	Some whiskey, hello!
Best Model	Some wine, ho!
Target	Some wine, ho!

Sample Translation 6

Model	Sentence
Source	Down, strumpet!
RNN	oh,
GRU	Hey, hey prostitute
BiGRU + General + 100% TF	prostitute!
BiGRU + Concat + 50% TF	Peace, prostitute!
BiGRU + Concat + 100% TF	Down, prostitute!
Best Model	Down, prostitute!
Target	Down, prostitute!

Sample Translation 7

Model	Sentence
Source	Let each man render me his bloody hand.
BiGRU + General	Let me, where where to his..
BiGRU + General + 50% TF	Let me leave his hand away.
BiGRU + Concat	Let give they give me me his hand hand.
BiGRU + Concat + 50% TF	Let the man from give me his honorable.
BiGRU + Concat + 100% TF	Let go get saying any man.
Best Model	Let me shake his beard hand.
Target	Let each man give me his bloody 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

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.

Questions?