Beam Search

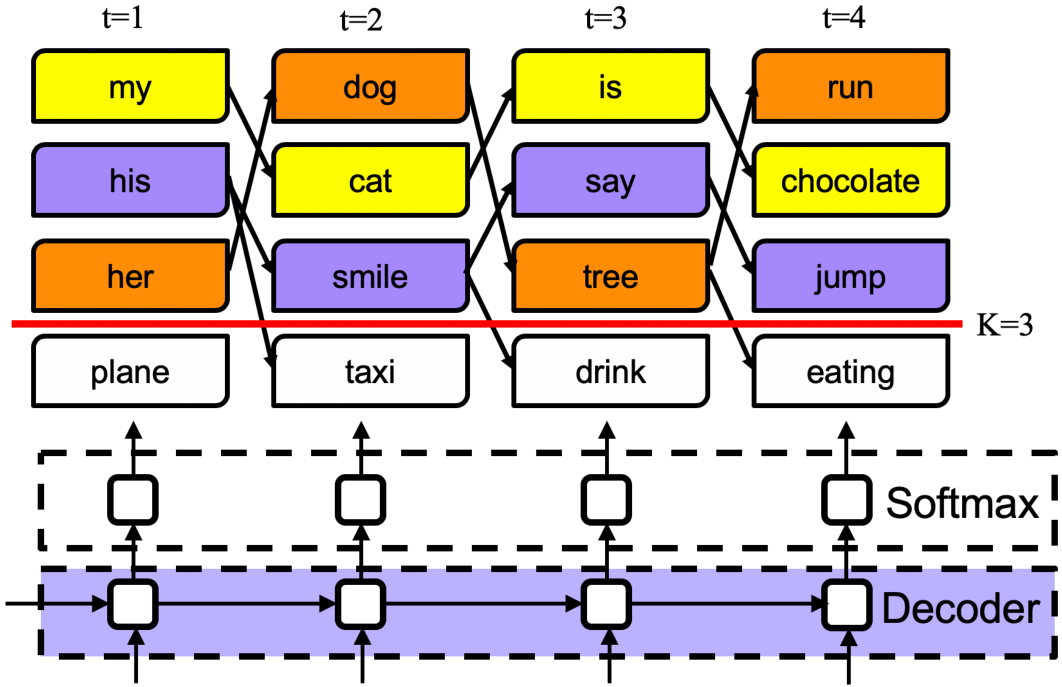
Shahrzad Kiani and Zihao Chen

CSC2547 Presentation

Beam Search

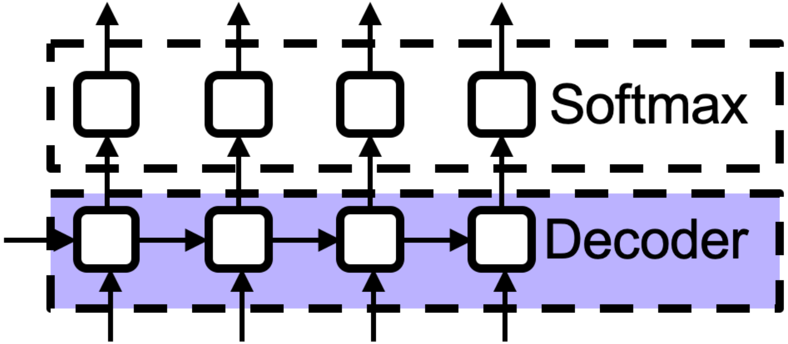
Greedy Search: Always go to **top 1** scored sequence (seq2seq)

Beam Search: Maintain the **top K** scored sequences (this paper)



Seq2Seq Train and Test Issues

gold sequence $y_{1:t} = [y_1, \ldots, y_t]$
predicted sequence $\hat{y}_{1:t} = [\hat{y}_1, \ldots, \hat{y}_t]$



Word level

- $p_{train}(\hat{y}_t | y_{1:t-1}) = Softmax(decoder(y_{1:t-1}))$
- $p_{test}(\hat{y}_t | \hat{y}_{1:t-1}) = Softmax(decoder(\hat{y}_{1:t-1}))$

1.Exposure Bias

Sentence level

- $p_{train}(\hat{y}_{1:t} = y_{1:t}) = \prod_{t=1}^{T} p(\hat{y}_t = y_t \, | y_{1:t-1})$

Seq2Seq Train and Test Issues (continued)

Training Loss

- Maximize $p_{train}(\hat{y}_{1:t} = y_{1:t}) = \prod_{t=1}^{T} p(\hat{y}_t = y_t \mid y_{1:t-1})$
- Minimize Negative Log Likelihood (NLL)

$$NLL = -ln \prod_{t=1}^{T} p(\hat{y}_t = y_t \mid y_{1:t-1}) = -\sum_{t} \ln(p(\hat{y}_t = y_t \mid y_{1:t-1}))$$

Testing Evaluation

- **Sequence level** metrics like BLEU

Seq2Seq Train and Test Issues (continued)

Training Loss

- Maximize $p_{train}(\hat{y}_{1:t} = y_{1:t}) = \prod_{t=1}^{T} p(\hat{y}_t = y_t \mid y_{1:t-1})$
- Minimize Negative Log Likelihood (NLL)

$$NLL = -ln \prod_{t=1}^{T} p(\hat{y}_t = y_t \mid y_{1:t-1}) = -\sum_{t} \ln(p(\hat{y}_t = y_t \mid y_{1:t-1}))$$

Testing Evaluation

- **Sequence level** metrics like BLEU

word level loss

2.Loss-Evaluation Mismatch

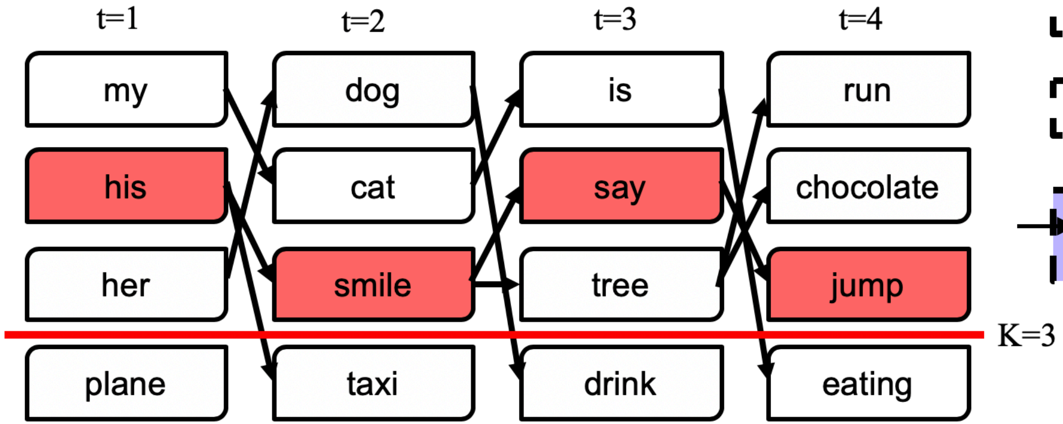
Optimization Approach

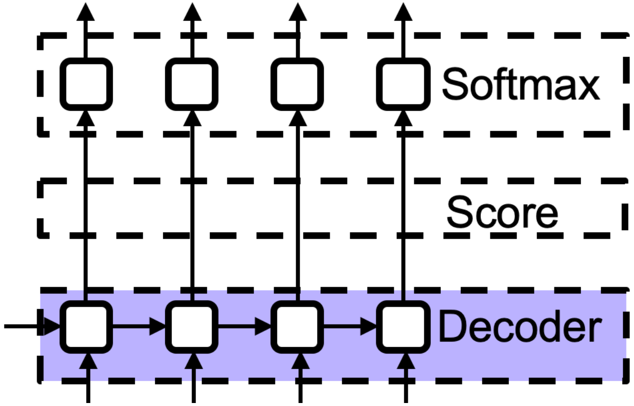
1. **Exposure Bias:** model is not exposed at its errors at training
    - Train with beam search
2. **Loss-Evaluation Mismatch**: loss on word level, evaluation on sequence
    - Define score for sequence
    - Define search-based sequence loss

Sequence Score

- $\text{score}(\hat{y}_{1:T}) = decoder(t)$
- Hard constraint $score(\hat{y}_{1:t}) = -\infty$

  Constrained Beam Search Optimization(ConBSO)
- Sequence with K-th ranked score $\hat{y}_{1:t}^{(K)}$



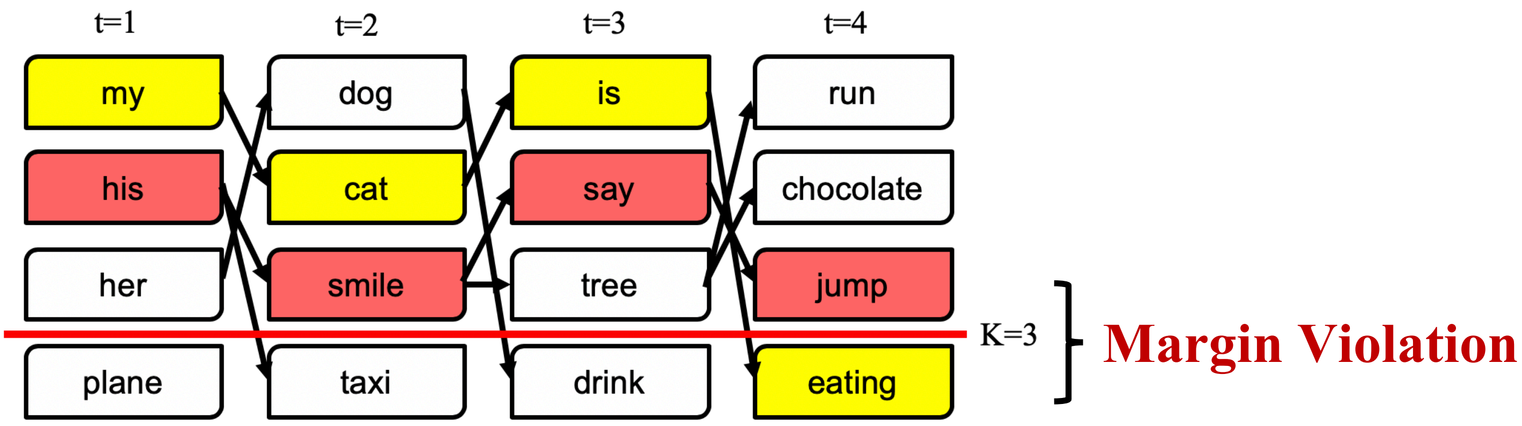


Search-Based Sequence Loss

$$\mathcal{L}(\theta) = \sum_t \Delta\left(\hat{y}_{1:t}^{(K)}\right)\left[1 + score\left(\hat{y}_{1:t}^{(K)}\right) - score(y_t)\right]$$

When $1 + score\left(\hat{y}_{1:t}^{(K)}\right) - score(y_t) > 0$:

- The gold sequence $y_{1:t}$ doesn't have a K highest score
- **Margin Violation**



Search-Based Sequence Loss (continued)

$$\mathcal{L}(\theta) = \sum_t \Delta\left(\hat{y}_{1:t}^{(K)}\right)\left[1 + score(\hat{y}_{1:t}^{(K)}) - score(y_t)\right]$$

$\Delta\left(\hat{y}_{1:t}^{(K)}\right)$

- scaling factor of penalizing the prediction
- = 1 when margin violation; = 0 when no margin violation

Goals:

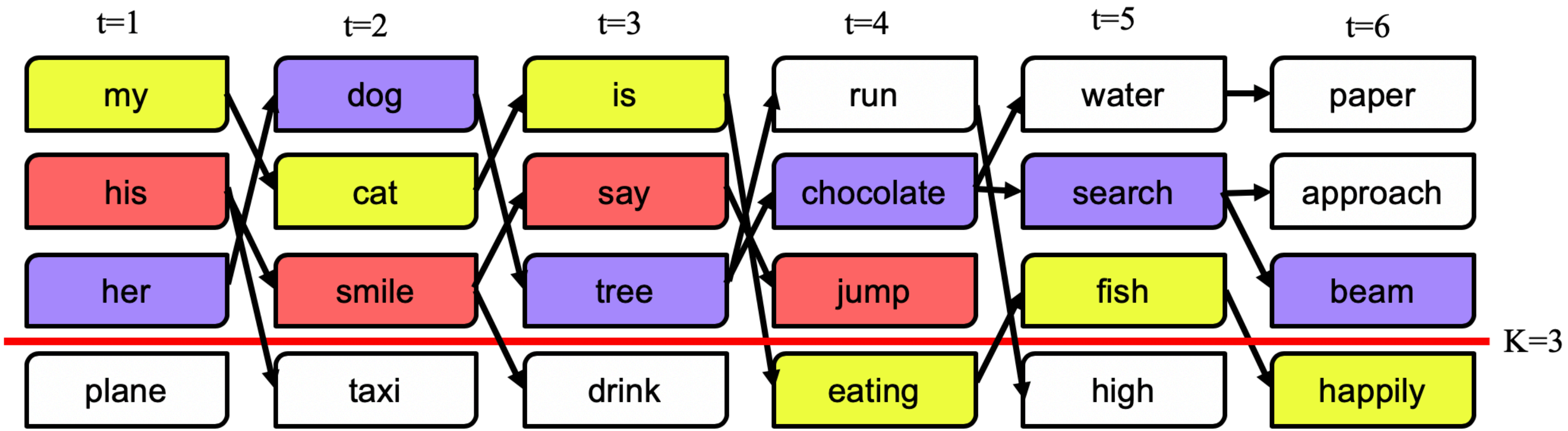
- When t<T, avoid margin violation, force the gold sequence to be **top K**
- When t=T, force the gold sequence to be **top 1**, so set K = 1

Backpropagation Through Time (BPTT)

- Recall loss function:

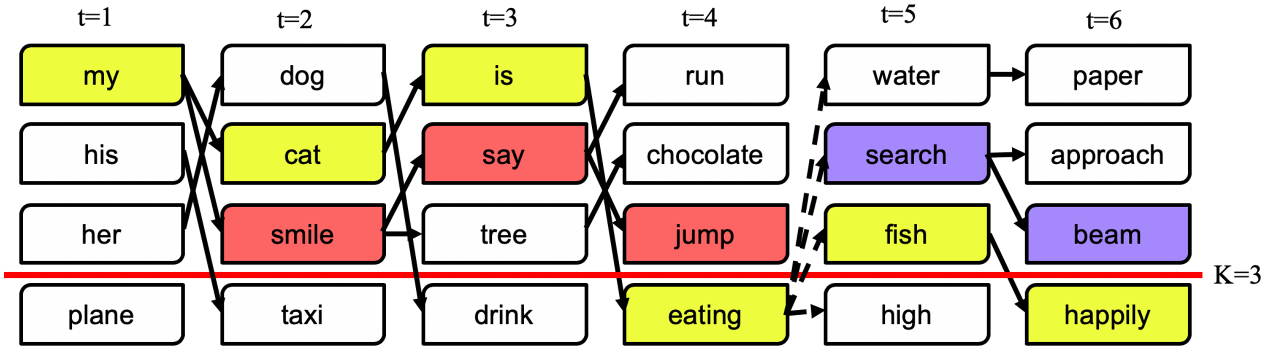
$$\mathcal{L}(\theta) = \sum_t \Delta\left(\hat{y}_{1:t}^{(K)}\right)\left[1 + score(\hat{y}_{1:t}^{(K)}) - score(y_t)\right]$$

- When margin violation, backpropagate for $score(\hat{y}_{1:t}^{(K)})$ and $score(y_t)$:$\boldsymbol{O(T)}$
- A margin violation at each time step: worst case $\boldsymbol{O(T^2)}$



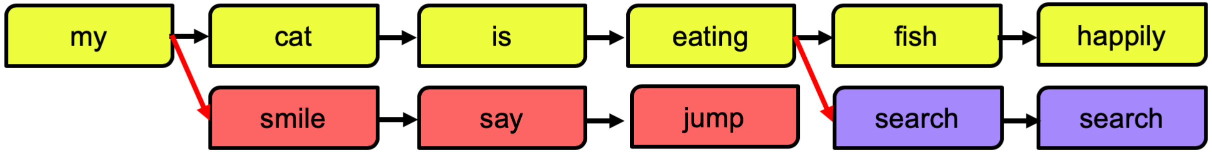
Learning as Search Optimization (LaSO)

- Normal case: update beam with $\hat{y}_{1:t}^{(K)}$
- Margin violation case: update beam with $y_{1:t}$ instead



Each incorrect sequence is an extension of the partial gold sequence

Only maintain two sequences, $O(2T) = \boldsymbol{O(T)}$



Experiment on Word Ordering

fish cat eat -> cat eat fish

Features

- Non-exhaustive search
- Hard constraint

Settings

- Dataset: PTB dataset
- Metrics: BLEU
- $\Delta(\hat{y}_{1:t}^{K})$ scaler: 0/1

	Word Ordering (BLEU)		
	K_{te} = 1	K_{te} = 5	K_{te} = 10
seq2seq	25.2	29.8	31.0
BSO	28.0	33.2	34.3
ConBSO	**28.6**	**34.3**	**34.5**
LSTM-LM	15.4	-	26.8

Table 1: Word ordering. BLEU Scores of seq2seq, BSO, constrained BSO, and a vanilla LSTM language model (from Schmaltz et al, 2016). All experiments above have $K_{tr} = 6$.

[Image credit: Sequence-to-Sequence Learning as Beam Search Optimization, Wiseman et al., EMNLP' 16]

Conclusion

Alleviate the issues of seq2seq

- Exposure Bias: Beam Search
- Loss-Evaluation Mismatch: sequence level cost function with O(T) BPTT with hard constraint

A variant of seq2seq with beam search training scheme

Related Works and References

- Wiseman, Sam, and Alexander M. Rush. "Sequence-to-Sequence Learning as Beam-Search Optimization." Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016.Sbs
- Kool, Wouter, Herke Van Hoof, and Max Welling. "Stochastic Beams and Where To Find Them: The Gumbel-Top-k Trick for Sampling Sequences Without Replacement." *International Conference on Machine Learning*. 2019.
- https://guillaumegenthial.github.io/sequence-to-sequence.html
- https://medium.com/@sharaf/a-paper-a-day-2-sequence-to-sequence-learning-as-beam-search-optimization-92424b490350
- https://www.facebook.com/icml.imls/videos/welcome-back-to-icml-2019-presentations-this-session-on-deep-sequence-models-inc/895968107420746/
- https://icml.cc/media/Slides/icml/2019/hallb(13-11-00)-13-11-00-4927-stochastic_beam.pdf
- https://vimeo.com/239248437
- Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." *Advances in neural information processing systems*. 2014.
  - **Propose Sequence-to Sequence learning with deep neural networks**
- Daumé III, Hal, and Daniel Marcu. "Learning as search optimization: Approximate large margin methods for structured prediction." *Proceedings of the 22nd international conference on Machine learning*. ACM, 2005.
  - **Propose a framework for learning as search optimization, and two parameter updates with convergence theorems and bounds**
- Gu, Jiatao, Daniel Jiwoong Im, and Victor OK Li. "Neural machine translation with gumbel-greedy decoding." *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.
  - **Propose the Gumbel-Greedy Decoding, which trains a generative network to predict translation under a trained model**