SEQ^3: Differentiable Sequence-to-Sequence-to-Sequence Autoencoder for Unsupervised Abstractive Sentence Compression

Christos Baziotis, Ion Androutsopoulos, Ioannis Konstas, Alexandros Potamianos

NAACL-HLT 2019, Minneapolis, USA

Machine Translation

the big black cat ...

η μεγάλη μαύρη γάτα...

Dialogue

A: What do you want to do tonight?

B: Let's go for a movie!

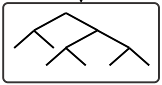
Text to Code

sort a list of numbers

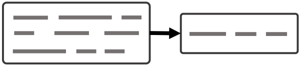
```
for i in range(len(A)):
    min_idx = i
    for j in range(i+1, len(A)):
        if A[min_idx] > A[j]:
            min_idx = j
    A[i], A[min_idx] = A[min_idx], A[i]
```

Text to Tree

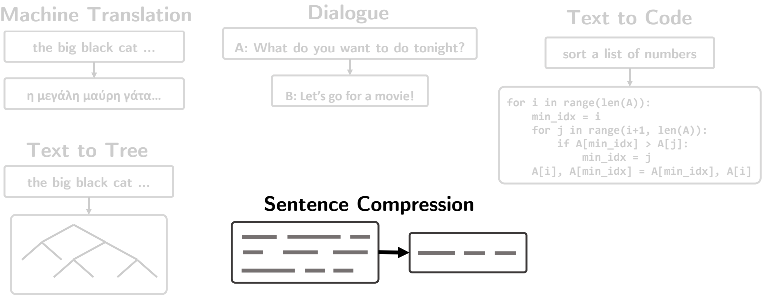
the big black cat ...



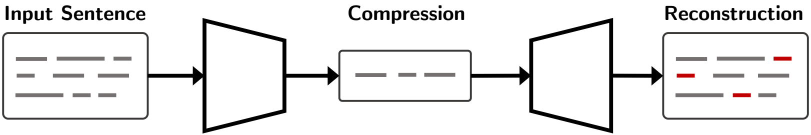
Sentence Compression



Introduction

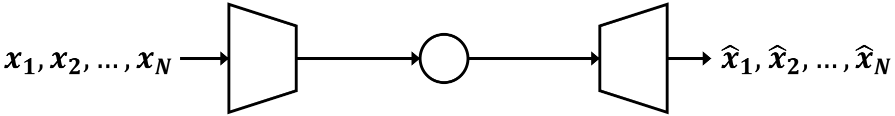


SEQ^3: Sequence-to-Sequence-to-Sequence Autoencoder



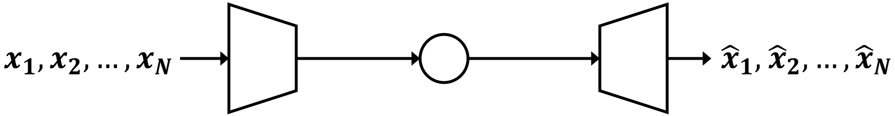
Unsupervised Models for Language

Vanilla Autoencoders

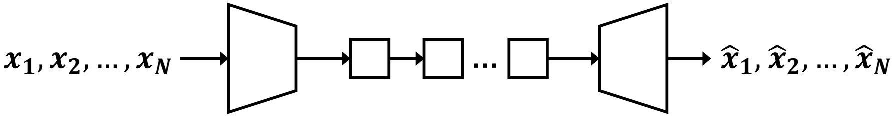


Unsupervised Models for Language

Vanilla Autoencoders



Discrete Latent Variable Autoencoders



- \+ Model the **discreteness** of language
- – Sampling is **not differentiable**
- – REINFORCE: sample **inefficient** and **unstable**

Contributions

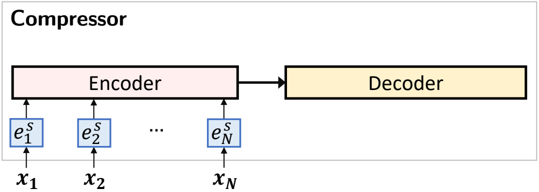
Model	Supervision	Abstractive	Differentiable	Latent
Miao & Blunsom (2016)	semi			✓
Wang & Lee (2018)	weak	✓		✓
Fevry & Phang (2018)	*none*		✓	
SEQ^3	none	✓	✓	✓

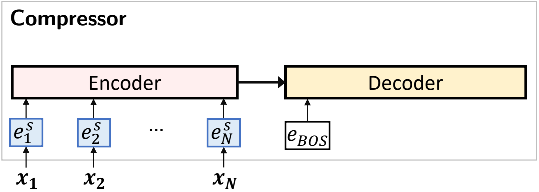
SEQ^3 Features (+ contributions)

+ Fully **unsupervised** and **abstractive**
+ Fully **differentiable** (continuous approximations)
+ **Topic**-grounded compressions
- **Human-readable** compressions via **LM prior**
- **User-defined** flexible compression ratio

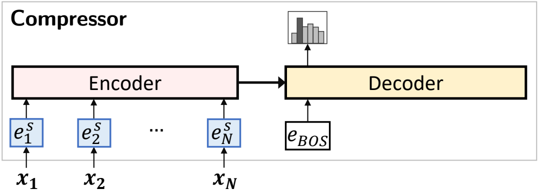
SOTA in unsupervised sentence compression

SEQ^3 Overview

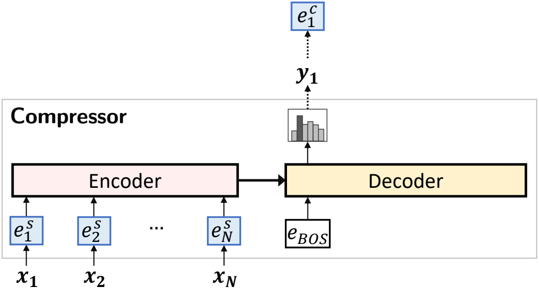



Compressor
Encoder
Decoder
e_1^s
e_2^s
...
e_N^s
e_{BOS}
x_1
x_2
x_N

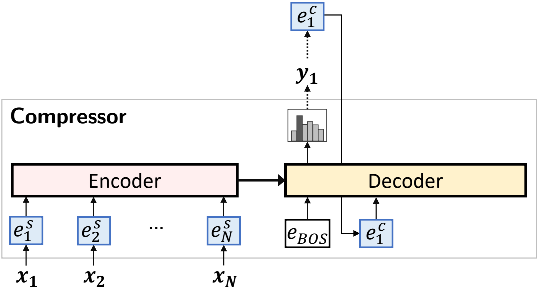
SEQ^3 Overview



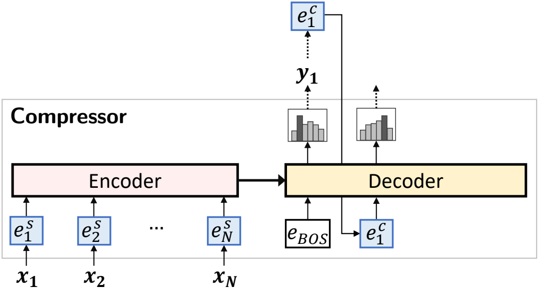
SEQ^3 Overview



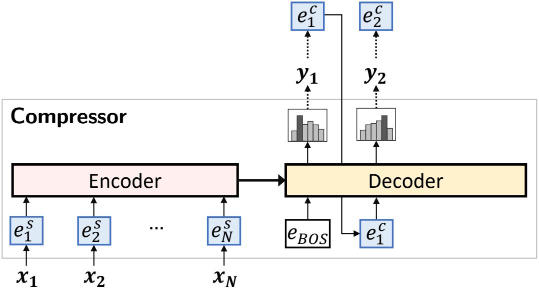
SEQ^3 Overview



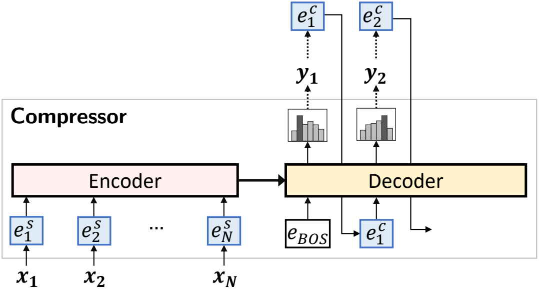
SEQ^3 Overview



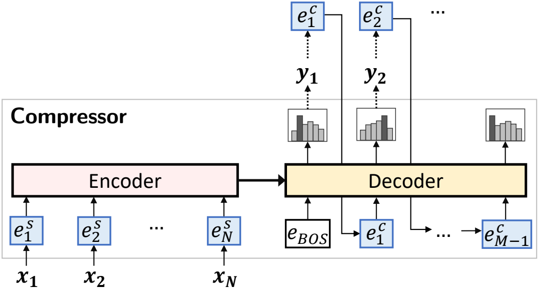
SEQ^3 Overview

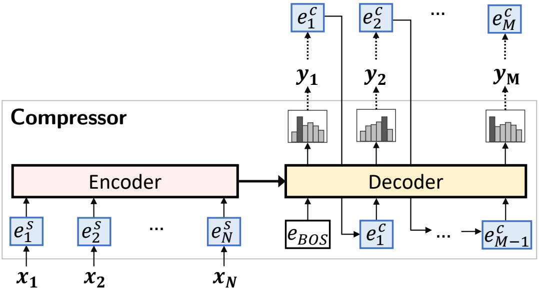


SEQ^3 Overview

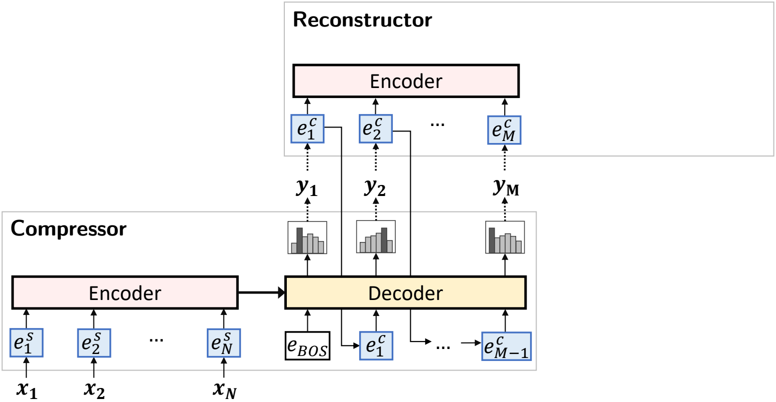


SEQ^3 Overview

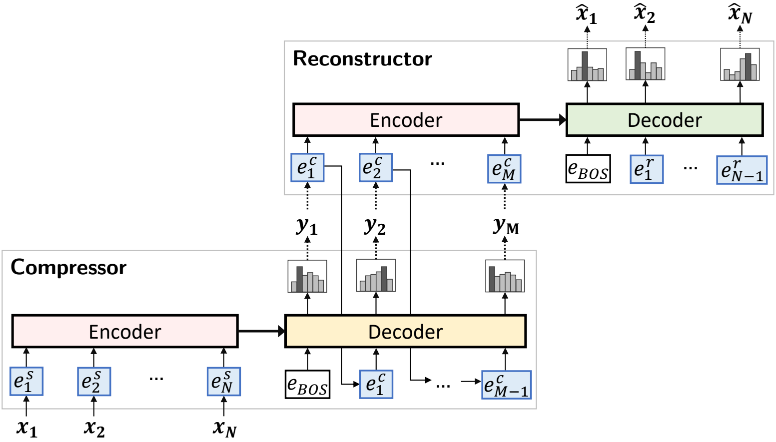



Compressor
Encoder
Decoder
e_1^s
e_2^s
...
e_N^s
x_1
x_2
x_N
e_{BOS}
e_1^c
e_{M-1}^c
y_1
y_2
y_M
e_1^c
e_2^c
e_M^c

SEQ^3 Overview



SEQ^3 Overview

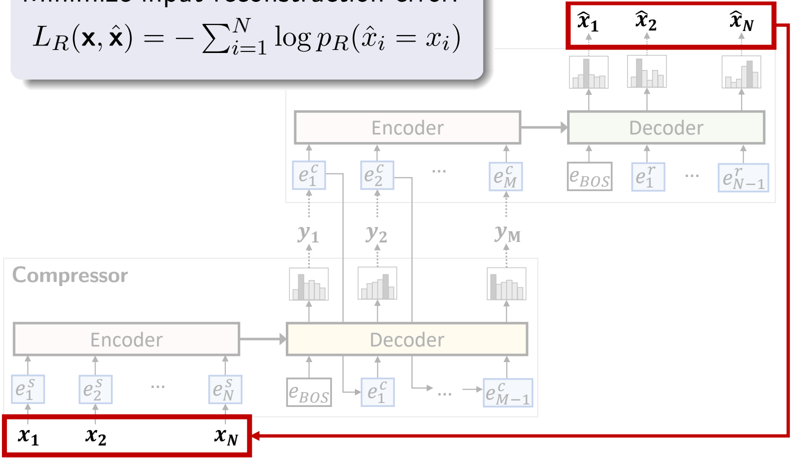


- **Reconstruction** loss: **distill** input into the latent sequence

Reconstruction Loss

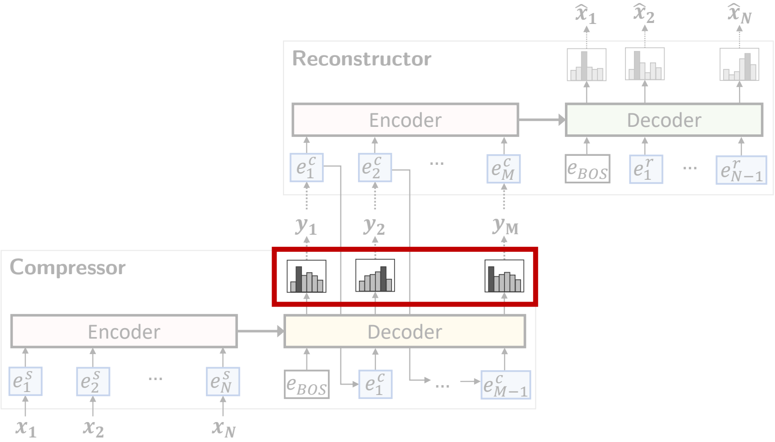
Minimize input reconstruction error:

$$L_R(\mathbf{x}, \hat{\mathbf{x}}) = -\sum_{i=1}^{N} \log p_R(\hat{x}_i = x_i)$$



SEQ^3 Overview

- **Reconstruction** loss: **distill** input into the latent sequence
- **LM Prior** loss: **human-readable** compressions



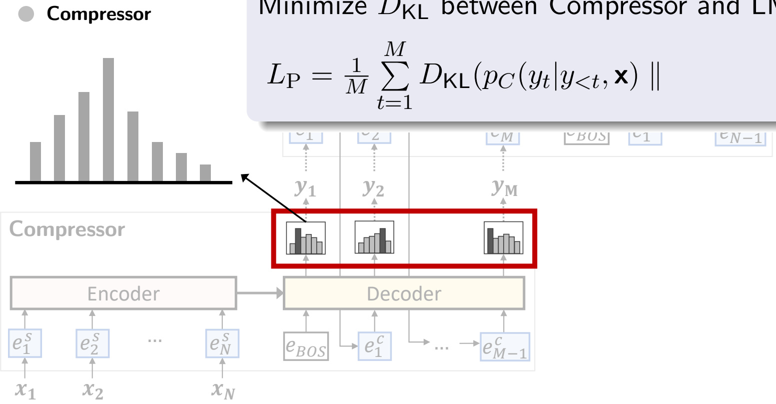
SEQ^3 Overview

- **Reconstruction** loss: **distill** input into the latent sequence
- **LM Prior** loss: **human-readable** compressions

LM Prior Loss

Minimize D_{KL} between Compressor and LM:

$$L_{\mathrm{P}} = \frac{1}{M} \sum_{t=1}^{M} D_{\mathsf{KL}}(p_C(y_t|y_{<t}, \mathbf{x}) \parallel \qquad)$$



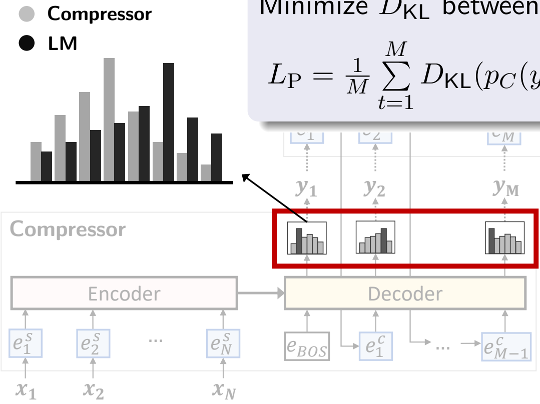
SEQ^3 Overview

- **Reconstruction** loss: **distill** input into the latent sequence
- **LM Prior** loss: **human-readable** compressions

LM Prior Loss

Minimize D_{KL} between Compressor and LM:

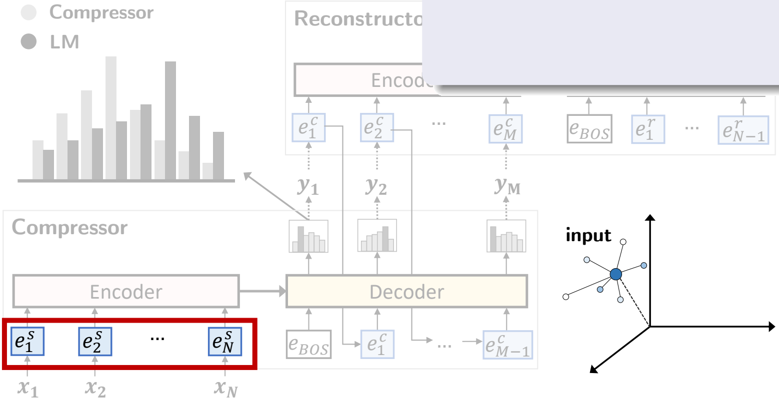
$$L_{\mathrm{P}} = \frac{1}{M}\sum_{t=1}^{M} D_{\mathrm{KL}}(p_C(y_t|y_{<t},\mathbf{x}) \parallel p_{LM}(y_t|y_{<t}))$$



- **Reconstruction** loss: **distill** input into the latent sequence
- **LM Prior** loss: **human-readable** compressions
- **Topic** loss: similar **topic** as input

Topic Loss

$\mathbf{v}^{\mathrm{x}}$: IDF-weighted average of $e_i^{\mathbf{s}}$



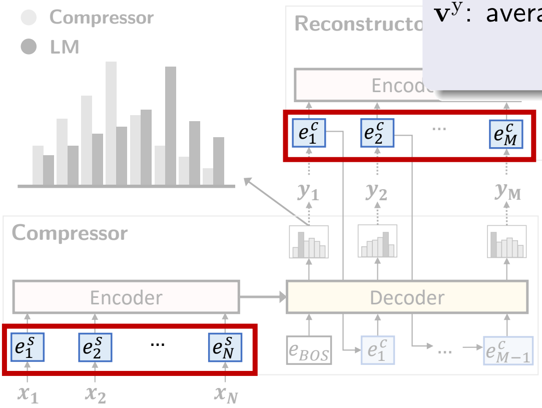
SEQ^3 Overview

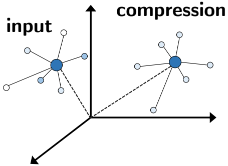
- **Reconstruction** loss: **distill** input into the latent sequence
- **LM Prior** loss: **human-readable** compressions
- **Topic** loss: similar **topic** as input

Topic Loss

$\mathbf{v}^{\mathrm{x}}$: IDF-weighted average of $e_i^{\mathbf{s}}$

$\mathbf{v}^{\mathrm{y}}$: average of $e_i^{\mathbf{c}}$





SEQ^3 Overview

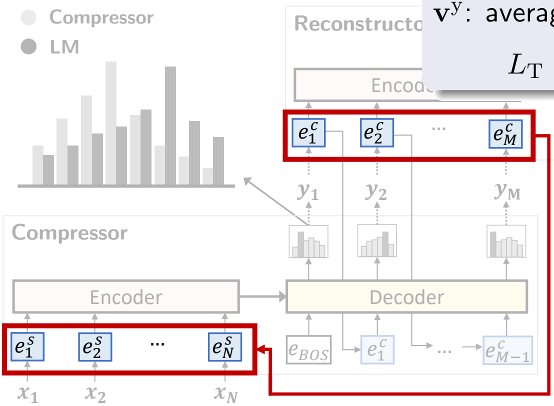
- **Reconstruction** loss: **distill** input into the latent sequence
- **LM Prior** loss: **human-readable** compressions
- **Topic** loss: similar **topic** as input

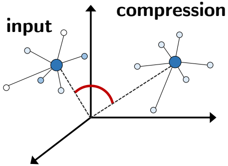
Topic Loss

$\mathbf{v}^{\mathrm{x}}$: IDF-weighted average of $e_i^{\mathbf{s}}$

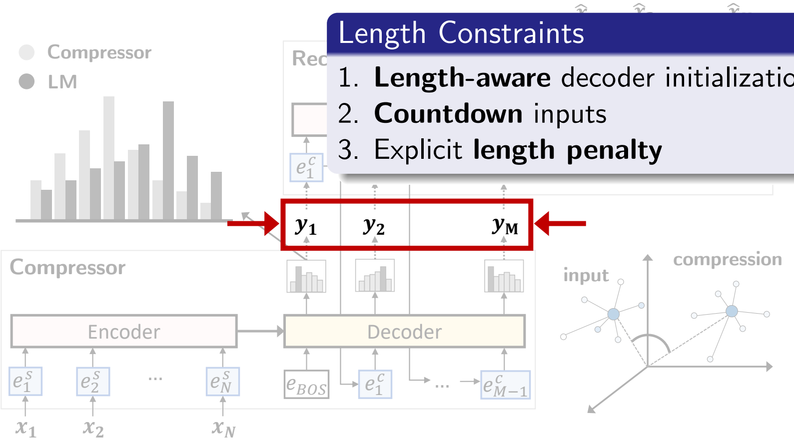
$\mathbf{v}^{\mathrm{y}}$: average of $e_i^{\mathbf{c}}$

$$L_{\mathrm{T}} = 1 - \cos(\mathbf{v}^{\mathrm{x}}, \mathbf{v}^{\mathrm{y}})$$





- **Reconstruction** loss: **distill** input into the latent sequence
- **LM Prior** loss: **human-readable** compressions
- **Topic** loss: similar **topic** as input
- **Length** constraints: user-defined **shorter** length



Length Constraints

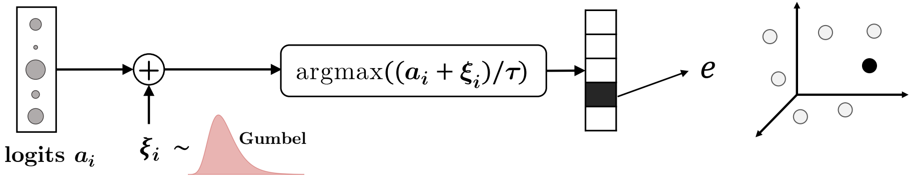
1. **Length-aware** decoder initialization
2. **Countdown** inputs
3. Explicit **length penalty**

Differentiable Sampling

Straight-Through + Gumbel-softmax

(Bengio et al.,2013, Maddison et al.,2017; Jang et al.,2017)

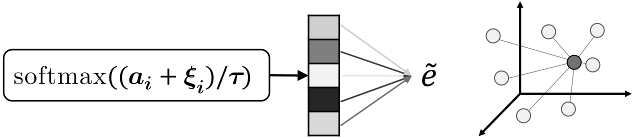
Forward-pass: **Discrete** embedding (Gumbel-max trick)



Backward-pass: **Mixture** of embeddings (Gumbel-softmax *approx.*)

Gradient

$$\nabla_\theta e \approx \nabla_\theta \tilde{e}$$

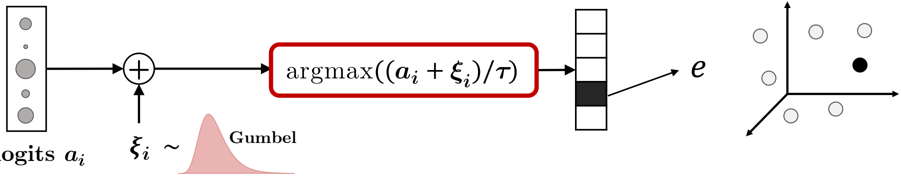


Differentiable Sampling

Straight-Through + Gumbel-softmax

(Bengio et al.,2013, Maddison et al.,2017; Jang et al.,2017)

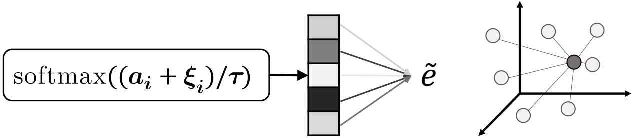
Forward-pass: **Discrete** embedding (Gumbel-max trick)



Backward-pass: **Mixture** of embeddings (Gumbel-softmax *approx.*)

Gradient

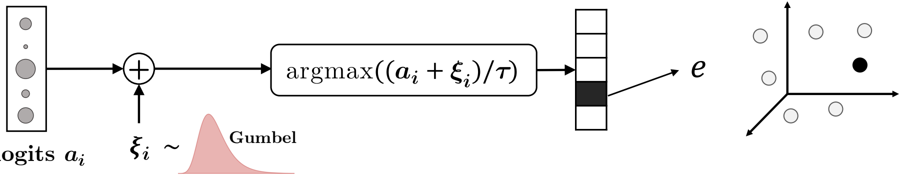
$$\nabla_\theta e \approx \nabla_\theta \tilde{e}$$



Straight-Through + Gumbel-softmax

(Bengio et al.,2013, Maddison et al.,2017; Jang et al.,2017)

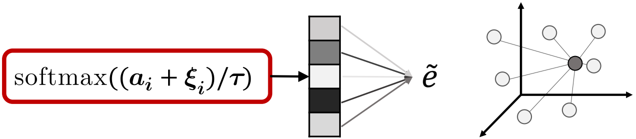
Forward-pass: **Discrete** embedding (Gumbel-max trick)



Backward-pass: **Mixture** of embeddings (Gumbel-softmax *approx.*)

Gradient

$$\nabla_\theta e \approx \nabla_\theta \tilde{e}$$



Experimental Setup

Dataset	Training	Evaluation
Gigaword (English)	✓ (source sentences)	✓
DUC-2003		✓
DUC-2004		✓

Training

- Train LM (LM prior) → Train SEQ^3
- **Never** exposed to target sentences (compressions)
- Vocabulary: 15K most frequent words in source sentences

Metrics

- Average F1 of ROUGE-1, ROUGE-2, ROUGE-L

Results on Gigaword

Supervision	Model	R-1	R-2	R-L
Unsupervised	Lead-8 (Rush et al., 2015)	21.86	7.66	20.45
	Pretrained Generator (Wang & Lee,2018)	21.26	5.60	18.89
	SEQ^3	**25.39**	**8.21**	**22.68**

Table: Results on (English) Gigaword for sentence compression.

Results on Gigaword

Supervision	Model	R-1	R-2	R-L
Unsupervised	LEAD-8 (Rush et al., 2015)	21.86	7.66	20.45
	Pretrained Generator (Wang & Lee,2018)	21.26	5.60	18.89
	SEQ^3	**25.39**	**8.21**	**22.68**
Weak	Adv. REINFORCE (Wang & Lee,2018)	28.11	9.97	25.41
Supervised	ABS (Rush et al.,2015)	29.55	11.32	26.42
	SEASS (Zhou et al., 2017)	36.15	17.54	33.63
	words-lvt5k-1sent (Nallapati et al.,2016)	<u>36.40</u>	<u>17.70</u>	<u>33.71</u>

Table: Results on (English) Gigaword for sentence compression.

Ablation

Model	R-1	R-2	R-L
SEQ^3 (Full)	25.39	8.21	22.68
SEQ^3 w/o LM	24.48 (-0.91)	6.68 (-1.53)	21.79 (-0.89)
SEQ^3 w/o TOPIC	3.89	0.10	3.75

Table: Ablation results on Gigaword.

Both topic and LM losses work in **synergy**

- **LM** prior loss: **how** words should be included
- **Topic** loss: **what** words to include

Model Outputs

INPUT the central election commission (cec) on monday decided that taiwan will hold another election of national assembly members on may # .

GOLD national <unk> election scheduled for may

SEQ^3 the central election commission (cec) announced elections .

INPUT dave bassett resigned as manager of struggling english premier league side nottingham forest on saturday after they were knocked out of the f.a. cup in the third round , according to local reports on saturday .

GOLD forest manager bassett quits

SEQ^3 dave bassett resigned as manager of struggling english premier league side UNK forest on knocked round press

Conclusions and Future Work

Conclusions

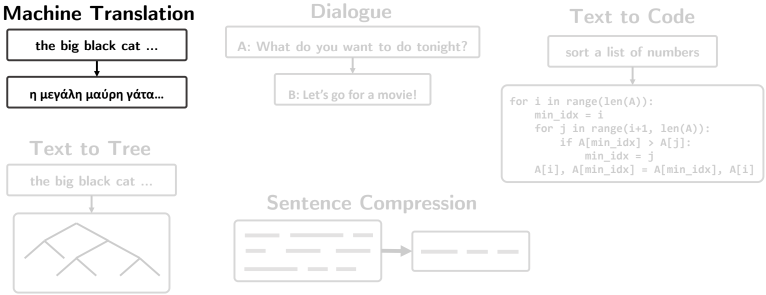
- Fully **differentiable** seq2seq2seq (SEQ^3) autoencoder
- SOTA in unsupervised abstractive sentence compression
- **Topic** loss is essential for convergence
- **LM prior** improves **readability**

Conclusions and Future Work

Conclusions

- Fully **differentiable** seq2seq2seq (SEQ^3) autoencoder
- SOTA in unsupervised abstractive sentence compression
- **Topic** loss is essential for convergence
- **LM prior** improves **readability**

Next Step: unsupervised machine translation



Questions?

Source code

https://github.com/cbaziotis/seq3

Contact me

christos.baziotis@gmail.com

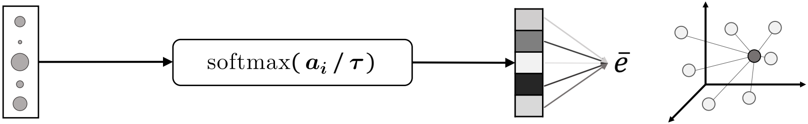
@cbaziotis

Appendix

Bonus Slides

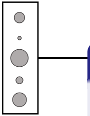
Soft-argmax: Weighted sum of embeddings from peaked softmax

(Goyal et al.,2017)

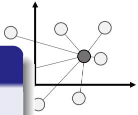


logits a_i

Soft-argmax: Weighted sum of embeddings from peaked softmax (Goyal et al.,2017)



logits a_i



Gumbel-Softmax

Gumbel-max trick:

$$
\begin{aligned}
\mathbf{y} &\sim \mathrm{softmax}(a_i) \\
&= \mathrm{argmax}(a_i + \xi_i), \quad \xi_i \sim \mathrm{Gumbel}
\end{aligned}
$$

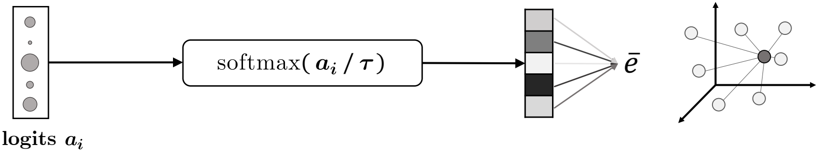
Gumbel-softmax relaxation:

$$\hat{\mathbf{y}} = \mathrm{softmax}(a_i + \xi_i), \quad \xi_i \sim \mathrm{Gumbel}$$

Differentiable Sampling (Extended)

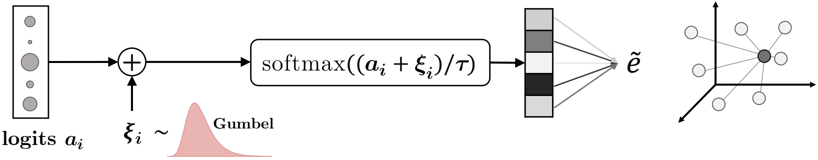
Soft-argmax: Weighted sum of embeddings from peaked softmax

(Goyal et al.,2017)



Gumbel-softmax: Differentiable approximation to sampling

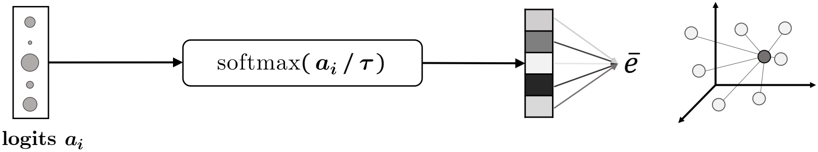
(Maddison et al.,2017; Jang et al.,2017)



Differentiable Sampling (Extended)

Soft-argmax: Weighted sum of embeddings from peaked softmax

(Goyal et al.,2017)

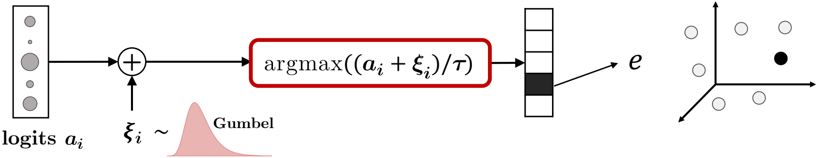


Gumbel-softmax: Differentiable approximation to sampling

(Maddison et al.,2017; Jang et al.,2017)

Straight-Through: forward-pass: one-hot, backward-pass: soft

(Bengio et al.,2013)



Out of Vocabulary (OOV) Words

We **copy OOV** words using the approach of Fevry and Phang (2018). Simpler alternative to pointer networks (See et al., 2017).

1. We use a set of **special OOV tokens**: OOV_1, OOV_2, ..., OOV_N.
2. We **replace** the $i\mathrm{th}$ unknown word in the input with the OOV_i token.
3. If all the OOV tokens are used, we use the generic UNK token.
4. In inference, we replace the special tokens with the original words.

OOV Handling Example

RAW	"John arrived in Rome yesterday. While in Rome, John had fun."
INPUT	"OOV_1 arrived in OOV_2 yesterday. While in OOV_2, OOV_1 had fun."
OOVs	John, Rome

Temperature for Gumbel-Softmax

Temperature τ does not affect the forward pass, but it **affects gradients**.

1. Jang et al. (2017) anneal $\tau \to 0$.
2. Gulcehre et al. (2017) **learn** τ:

$$\tau(h_t^c) = \frac{1}{\log(1 + \exp(w_\tau^\intercal\, h_t^c)) + 1}$$

3. Havrylov & Titov (2017) tune bound τ_0:

$$\tau(h_t^c) = \frac{1}{\log(1 + \exp(w_\tau^\intercal\, h_t^c)) + \boldsymbol{\tau_0}}$$

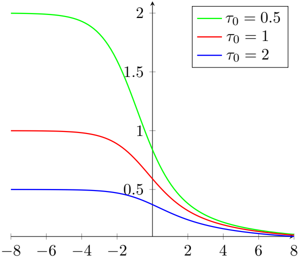


Figure: Values of τ_0 bound.

In our experiments the learned temperature lead to **instability**.
We **fix** $\tau = 0.5$ following (Gu et al., 2018).

Implementation Details

Hyper-Parameters

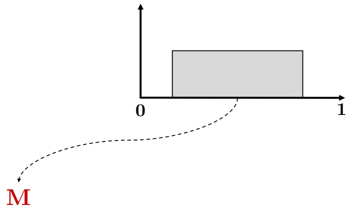
- Encoders: 2-layer bidirectional LSTM with size 300
- Decoders: 2-layer unidirectional LSTM with size 300
- Embedding: initialize with $100\mathrm{d}$ GloVe (Pennington et al., 2014)

Parameter Sharing

- **Tied encoders** of the compressor and reconstructor.
- **Shared embedding** layer for all encoders and decoders.
- **Tied embedding-output** layers of both decoders.

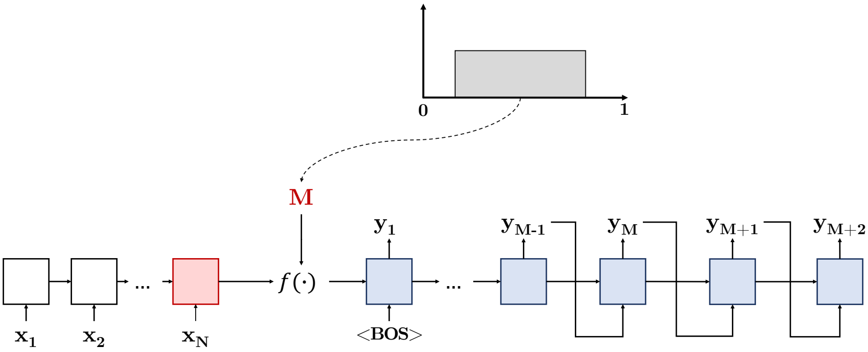
Length Control

1. **Sample** target length M.



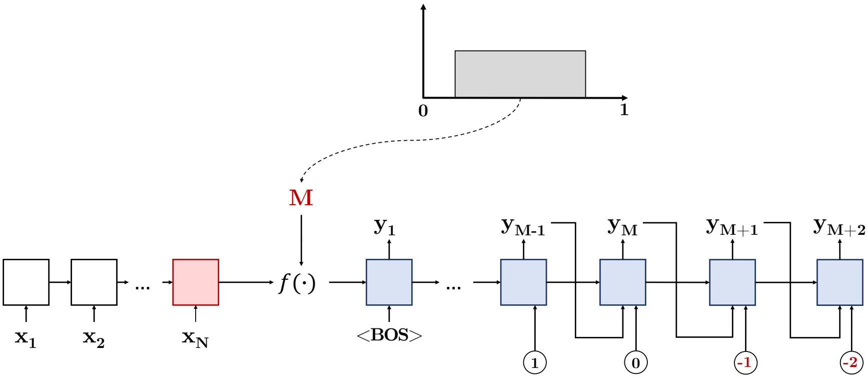
Length Control

1 **Sample** target length M.
2 Decoder's state **length-aware initialization**.



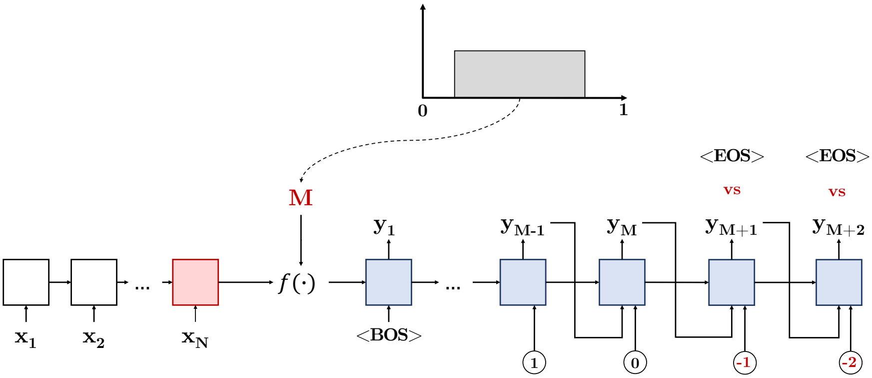
Length Control

1 **Sample** target length M.
2 Decoder's state **length-aware initialization**.
3 **Countdown** input.



Length Control

1 **Sample** target length M.
2 Decoder's state **length-aware initialization**.
3 **Countdown** input.
4 Explicit length **penalty**.



Results on DUC Shared Tasks

Model	R-1	R-2	R-L
TOPIARY (Zajic et al., 2007)	25.12	6.46	20.12
(Woodsend et al., 2010)	22.00	6.00	17.00
ABS (Rush et al., 2015)	28.18	8.49	23.81
PREFIX	20.91	5.52	18.20
SEQ^3 (Full)	**22.13**	**6.18**	**19.3**

Table: Results on the DUC-2004

Model	R-1	R-2	R-L
ABS (Rush et al., 2015)	28.48	8.91	23.97
PREFIX	**21.3**	**6.38**	**18.82**
SEQ^3 (Full)	20.90	6.08	18.55

Table: Results on the DUC-2003

Model Output (Extra)

INPUT	the american sailors who thwarted somali pirates flew home to the u.s. on wednesday but without their captain , who was still aboard a navy destroyer after being rescued from the hijackers .
GOLD	us sailors who thwarted pirate hijackers fly home
SEQ^3	the american sailors who foiled somali pirates flew home after crew hijacked .