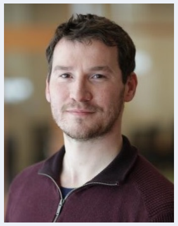
Classical Structured Prediction Losses for Sequence to Sequence Learning

Sergey Edunov*, Myle Ott*

Michael Auli, David Grangier, Marc'Aurelio Ranzato

Myle Ott



Michael Auli

David Grangier

Marc'Aurelio Ranzato

facebook
Artificial Intelligence

Training Seq2Seq models

Source:
Wir müssen unsere Einwanderungspolitik in Ordnung bringen.

Target:
We have to fix our immigration policy.

$$log\, p(\mathbf{t}|\mathbf{x}) = -\sum_{i=1}^{n} \log p(t_i|t_1, ..., t_{i-1}, \mathbf{x})$$

Training Seq2Seq models

Source:
Wir müssen unsere Einwanderungspolitik in Ordnung bringen.

Target:
We have to fix our immigration policy.

$$log\, p(\mathbf{t}|\mathbf{x}) = -\sum_{i=1}^{n} \log p(t_i|t_1, ..., t_{i-1}, \mathbf{x})$$

Decoding

Source:
Wir müssen unsere Einwanderungspolitik in Ordnung bringen.

Model output:
We need to fix our ....

$$u_i = arg\, \underset{u}{max}\, p(u|u_1, ..., u_{i-1}, \mathbf{x})$$

Decoding is autoregressive.

Exposure bias: training and testing are inconsistent

Evaluation

Training criterion (NLL) != Evaluation criterion (BLEU)

Evaluation criterion requires decoding

Evaluation criterion is not differentiable

Sequence level training with Neural Nets

Reinforcement Learning-inspired methods

- MIXER (Ranzato et al., ICLR 2016)
- Actor-Critic (Bahdanau et al., ICLR 2017)

Using beam search at training time:

- Beam search optimization (Wiseman et al. ACL 2016)
- Distillation based (Kim et al., EMNLP 2016)

Sequence level training before Neural Nets

How classical structure prediction compare to recent methods?

Classical losses for log-linear models, do they work for neural nets?

Bottou et al. “Global training of document processing systems with graph transformer networks” CVPR 1997

Collins “Discriminative training methods for HMMs” EMNLP 2002

Taskar et al. “Max-margin Markov networks” NIPS 2003

Tsochantaridis et al. “Large margin methods for structured and interdependent output variables” JMLR 2005

Och “Minimum error rate training in statistical machine translation” ACL 2003

Smith and Eisner “Minimum risk annealing for training log-linear models” ACL 2006

Gimpel and Smith “Softmax-margin CRFs: training log-linear models with cost functions” ACL 2010

Baseline: Token Level NLL

Source:
Wir müssen unsere Einwanderungspolitik in Ordnung bringen.

Target:
We have to fix our immigration policy.

$$\mathcal{L}_{\text{TokNLL}} = -\sum_{i=1}^{n} \log p(t_i | t_1, \ldots, t_{i-1}, \mathbf{x})$$

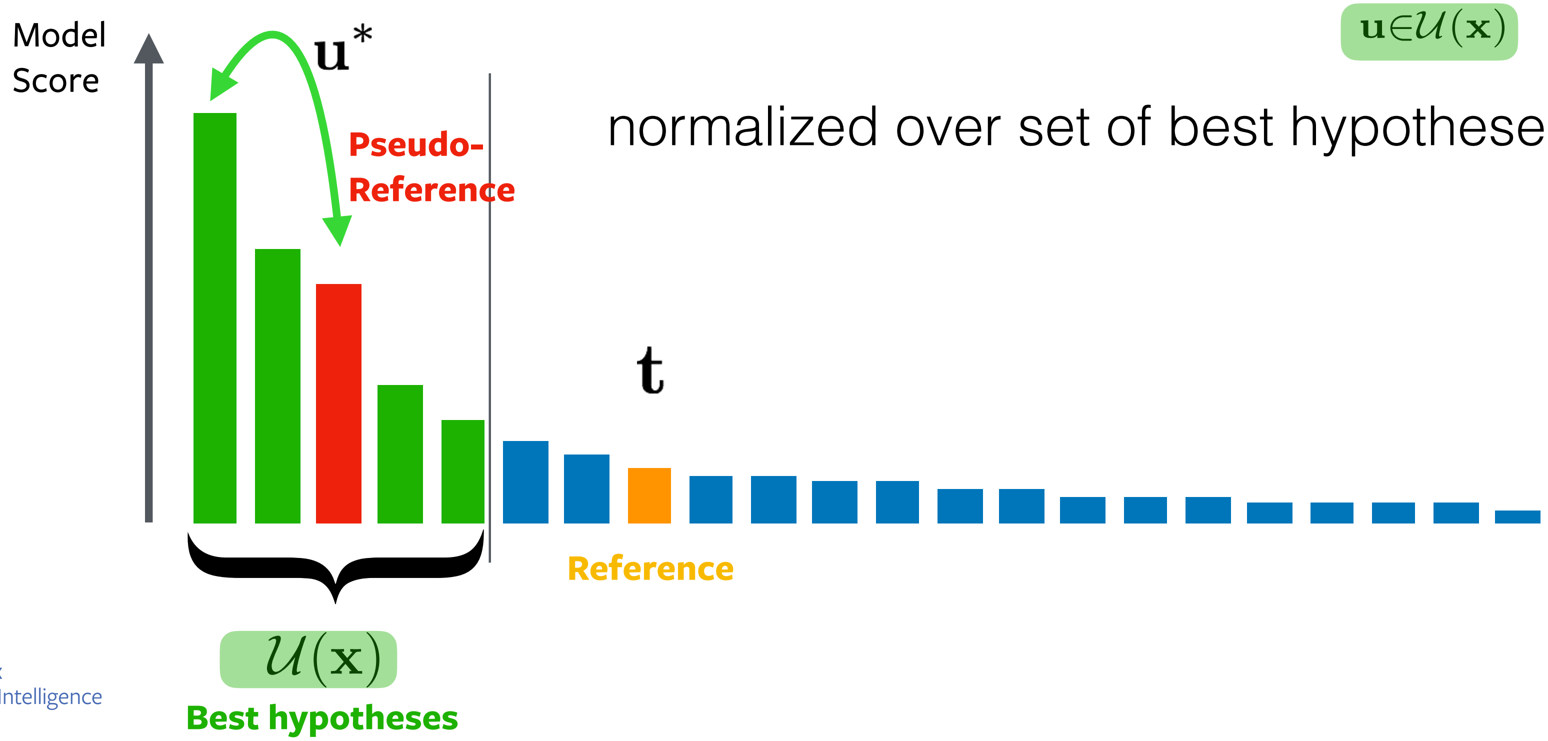
target
conditioning

'Locally' normalized over vocabulary.

Sequence Level NLL

$$\mathcal{L}_{\text{SeqNLL}} = -\log p(\mathbf{u}^*|\mathbf{x}) + \log \sum_{\mathbf{u} \in \mathcal{U}(\mathbf{x})} p(\mathbf{u}|\mathbf{x})$$

normalized over set of best hypotheses $\mathcal{U}(\mathbf{x})$



Sequence Level NLL

$$\mathcal{L}_{\text{SeqNLL}} = -\log p(\mathbf{u}^*|\mathbf{x}) + \log \sum_{\mathbf{u} \in \mathcal{U}(\mathbf{x})} p(\mathbf{u}|\mathbf{x})$$

Source:
Wir müssen unsere Einwanderungspolitik in Ordnung bringen.

Target:
We have to fix our immigration policy.

Beam:

$\mathcal{U}(\mathbf{x})$

BLEU	Model score	
45.5	-0.23	We should fix our immigration policy.
75.0	-0.30	We need to fix our immigration policy. → $\mathbf{u}^*$
36.9	-0.36	We need to fix our policy policy.
66.1	-0.42	We have to fix our policy policy.
66.1	-0.44	We've got to fix our immigration policy.

Sequence Level NLL

$$\mathcal{L}_{\text{SeqNLL}} = -\log p(\mathbf{u}^*|\mathbf{x}) + \log \sum_{\mathbf{u}\in\mathcal{U}(\mathbf{x})} p(\mathbf{u}|\mathbf{x})$$

Source:
Wir müssen unsere Einwanderungspolitik in Ordnung bringen.

Target:
We have to fix our immigration policy.

Beam:

$\mathcal{U}(\mathbf{x})$

BLEU	Model score	
45.5	-0.23	We should fix our immigration policy.
75.0	-0.30	We need to fix our immigration policy.
36.9	-0.36	We need to fix our policy policy.
66.1	-0.42	We have to fix our policy policy.
66.1	-0.44	We've got to fix our immigration policy.

$\mathbf{u}^*$

Expected Risk

$$\mathcal{L}_{\text{Risk}} = \sum_{\mathbf{u} \in \mathcal{U}(\mathbf{x})} \text{cost}(\mathbf{t}, \mathbf{u}) \frac{p(\mathbf{u}|\mathbf{x})}{\sum_{\mathbf{u}' \in \mathcal{U}(\mathbf{x})} p(\mathbf{u}'|\mathbf{x})}$$

Source:
Wir müssen unsere Einwanderungspolitik in Ordnung bringen.

Target:
We have to fix our immigration policy.

$$cost(\mathbf{t}, \mathbf{u}) = 1 - BLEU(\mathbf{t}, \mathbf{u})$$

Beam:

$\mathcal{U}(\mathbf{x})$

BLEU	Model score	
45.5	-0.23	We should fix our immigration policy.
75.0	-0.30	We need to fix our immigration policy.
36.9	-0.36	We need to fix our policy policy.
66.1	-0.42	We have to fix our policy policy.
66.1	-0.44	We've got to fix our immigration policy.

Ayana et al. (2016)
Shen et al. (2016)

Expected Risk

$$\mathcal{L}_{\text{Risk}} = \sum_{\mathbf{u} \in \mathcal{U}(\mathbf{x})} \text{cost}(\mathbf{t}, \mathbf{u}) \frac{p(\mathbf{u}|\mathbf{x})}{\sum_{\mathbf{u}' \in \mathcal{U}(\mathbf{x})} p(\mathbf{u}'|\mathbf{x})}$$

Source:
Wir müssen unsere Einwanderungspolitik in Ordnung bringen.

Target:
We have to fix our immigration policy.

$$cost(\mathbf{t}, \mathbf{u}) = 1 - BLEU(\mathbf{t}, \mathbf{u})$$

Beam:

$\mathcal{U}(\mathbf{x})$

BLEU	Model score	
45.5	-0.23	We should fix our immigration policy.
75.0	-0.30	We need to fix our immigration policy.
36.9	-0.36	We need to fix our policy policy.
66.1	-0.42	We have to fix our policy policy.
66.1	-0.44	We've got to fix our immigration policy.

(expected BLEU=58)

Ayana et al. (2016)
Shen et al. (2016)

Other sequence level training losses

Check our paper!

- Max-Margin
- Multi-Margin
- Softmax-Margin

Results on IWSLT'14 De-En

	TEST
TokNLL (Wiseman et al. 2016)	24.0
BSO (Wiseman et al. 2016)	26.4
Actor-Critic (Bahdanau et al. 2016)	28.5
Phrase-based NMT (Huang et al. 2017)	29.2

Results on IWSLT'14 De-En

	TEST
TokNLL (Wiseman et al. 2016)	24.0
BSO (Wiseman et al. 2016)	26.4
Actor-Critic (Bahdanau et al. 2016)	28.5
Phrase-based NMT (Huang et al. 2017)	29.2
our TokNLL	31.8

Results on IWSLT'14 De-En

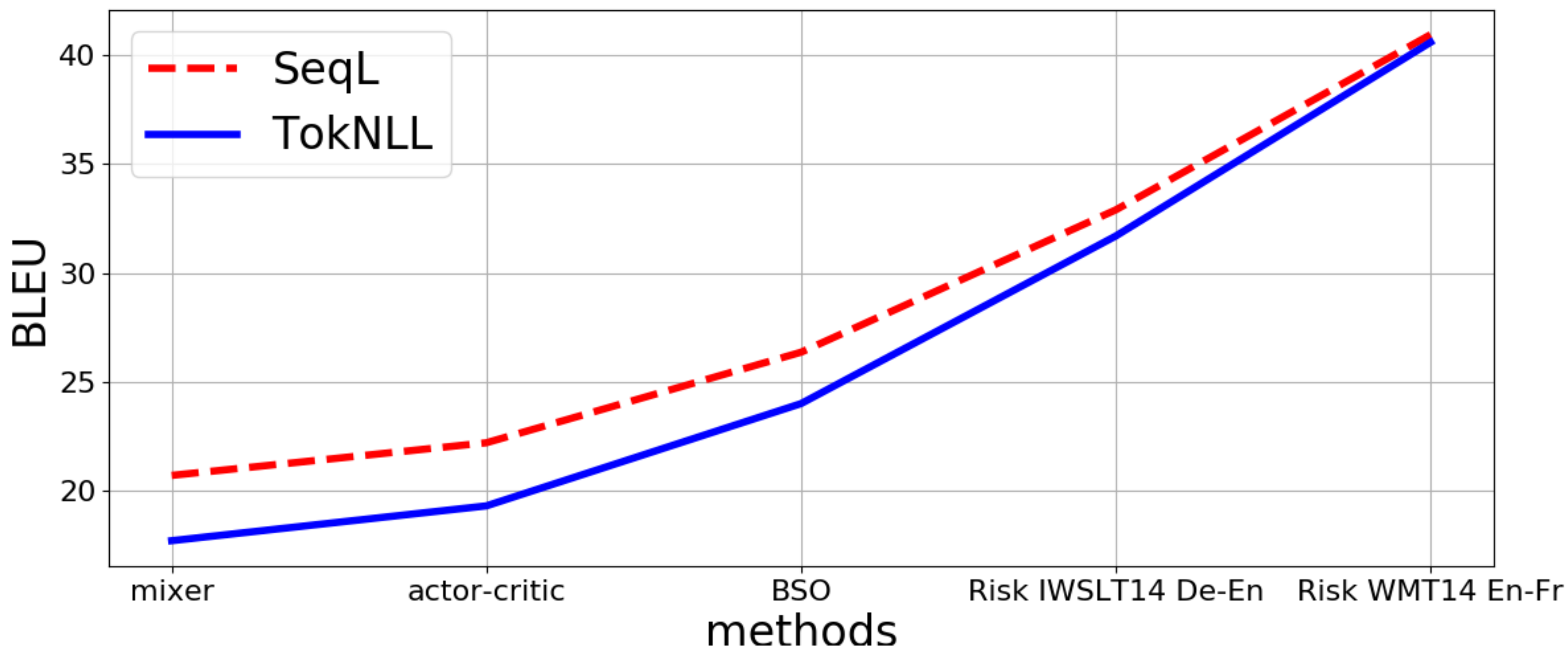
	TEST
TokNLL (Wiseman et al. 2016)	24.0
BSO (Wiseman et al. 2016)	26.4
Actor-Critic (Bahdanau et al. 2016)	28.5
Phrase-based NMT (Huang et al. 2017)	29.2
our TokNLL	31.8
SeqNLL	32.7
Risk	**32.8**
Max-Margin	32.6

Fair Comparison to BSO

	TEST
TokNLL (Wiseman et al. 2016)	24.0
BSO (Wiseman et al. 2016)	26.4
Our re-implementation of their TokNLL	23.9
Risk on top of the above TokNLL	26.7

Methods are comparable once the baseline is the same…

Diminishing Returns



On WMT'14 En-Fr, TokNLL gets 40.6 while Risk gets 41.0
The stronger the baseline, the less to be gained.

Practical Tip #1

Better if pre-trained model had **label smoothing**.

		valid	test
base	TokNLL	32.96	31.74
	Risk init with TokNLL	33.27	32.07
	Δ	0.31	0.33
label smoothing	TokLS	33.11	32.21
	Risk init with TokLS	33.91	32.85
	Δ	0.8	0.64

Practical Tip #2

Accuracy vs speed trade-off: **offline/online** generation of hypotheses.

	valid	test
Online generation	33.91	32.85
Offline generation*	33.52	32.44

***Offline is 26x faster than online**

Practical Tip #3

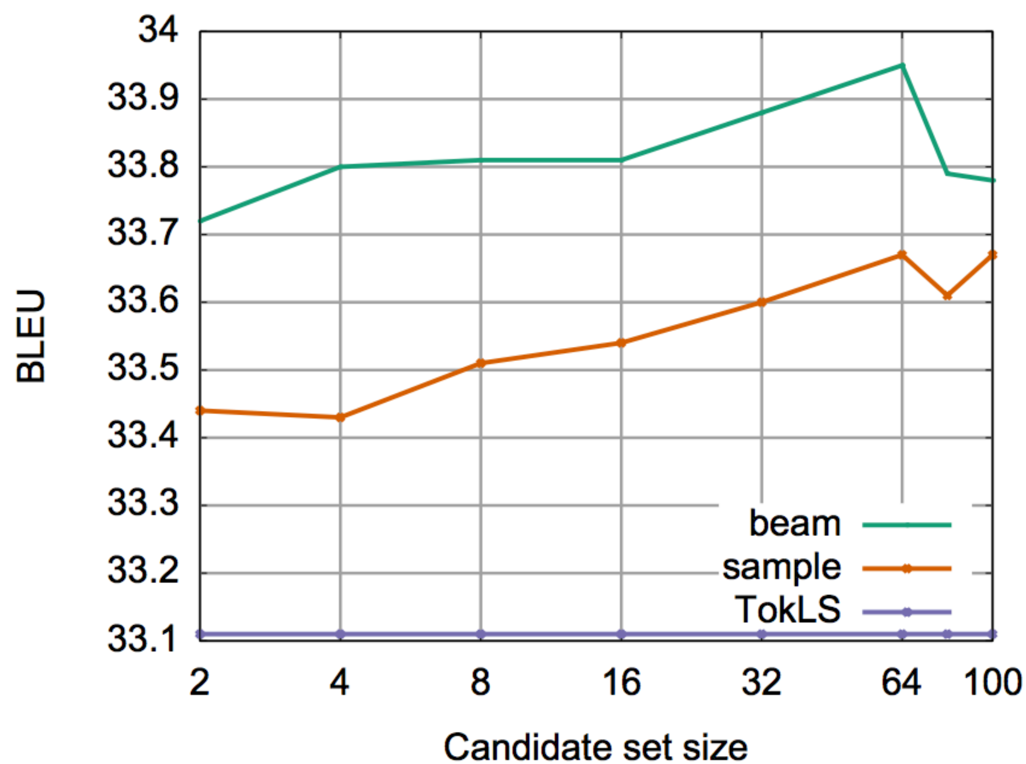
better result when **combining** token-level + sequence-level loss

		valid	test
Single Task	TokLS	33.11	32.21
	Risk only	33.55	32.45
	Δ	0.44	0.24
Combined	Weighted Risk + TokLS	33.91	32.85
	Δ	0.8	0.64

Practical Tip #4

Bigger search space size = better performance

It is also more computationally expensive



Practical Tip #5

All structural losses are **comparable**

	test
TokNLL	31.78
TokNLL+Smoothing	32.23
Sequence NLL	32.68
Risk	32.84
Max Margin	32.55
Multi Margin	32.59
Softmax Margin	32.71

Summary

Initialize from a model pre-trained at the token level.
Training with search is excruciatingly **slow**…

Sequence level training does **improve**, but with diminishing returns.

Specific loss to train at the sequence level does not matter.

Important to use **pseudo-reference** as opposed to real reference.

Code at:
https://github.com/pytorch/fairseq/tree/classic_seqlevel

Questions?


Diolch
Kiitos
umesc
Sheun
Kasih
Todah
Mamnoon
Shnorhakalutiun
Shokriya
Dziękuję
Shokrun
Spaas
Mul
Ači
Gamsahapnida
Dekuju/Dekujeme
Hvala
Ngiyabonga
Xie
Dank
Takk
Cam
or
Gra
Kop
Salamat
Merci
al
Dhanyavaad
Dakujem
Waad
Dhanyavaadaalu
Dankie
Go
Grazie
Faleminderit
Daw
Dhanyavaadalu
Dhanyavad
Khopjai
Kruthagnathalu
krap
ThankYou
kun
Arigatou
Tack
Shukriya
Or
Dhonnobaad
ederim
Asante
Grazzi
raibh
Gracias
Nandree
Blagodariya
Gomapsupnida
Fyrir
Terima
Enkosi
Danke
dank
Euxaristo
Kun
Hain
Dhan
daa