Noisy Self-Knowledge Distillation for Text Summarization

Yang Liu, Sheng Shen and Mirella Lapata

NAACL 2021

Challenges in Automatic Summarization

- Maximum-likelihood training on single reference datasets
- Why this is not optimal?

Challenges in Automatic Summarization

- Maximum-likelihood training on single reference datasets
- Why this is not optimal?
  1. Human variation in summarization tasks

Challenges in Automatic Summarization

- Maximum-likelihood training on single reference datasets
- Why this is not optimal?
  1. Human variation in summarization tasks
     - Different people disagree on writing style and content selection
     - Summarization is naturally a multi-reference task

Challenges in Automatic Summarization

- Maximum-likelihood training on single reference datasets
- Why this is not optimal?
    1. Human variation in summarization tasks
    2. Most popular benchmarks are collated opportunistically

Challenges in Automatic Summarization

- Maximum-likelihood training on single reference datasets
- Why this is not optimal?
  1. Human variation in summarization tasks
  2. Most popular benchmarks are collated opportunistically
     - Their summaries only loosely correspond to the source input [1]
     - The inherent noise in the data collection hampers training, and models may be prone to hallucination

[1] Wojciech, K el al. Neural text summarization: A critical evaluation. EMNLP 2019

Challenges in Automatic Summarization

- Maximum-likelihood training on single reference datasets
- Why this is not optimal?
  1. Human variation in summarization tasks
  2. Most popular benchmarks are collated opportunistically

Challenges in Automatic Summarization

- Maximum-likelihood training on single reference datasets
- Why this is not optimal?
  1. Human variation in summarization tasks
  2. Most popular benchmarks are collated opportunistically
- Self-Knowledge Distillation can alleviate these challenges

Knowledge Distillation

- Teacher neural network $\mathcal{T}$, Student neural network $\mathcal{S}$
- To train a new student, Knowledge Distillation usually penalizes the difference between the trained teacher and the student

$$\mathrm{L}_{KD} = \sum_{x_i \in \mathrm{X}} l(f_T(x_i), f_S(x_i))$$

- Self-knowledge distillation refers to the special case: teacher and student have identical neural network architectures

Self-Knowledge Distillation for Text Summarization

- Teacher outputs provide softened distributions of the reference summaries
  - An enrichment of the single reference setting
  - A reweighting of gold summaries
  - Prevent the student from becoming over-confident in its predictions.

Self-Knowledge Distillation for Text Summarization

- NLL loss for abstractive summarization

$$\mathrm{L}_{\mathrm{NLL}} = -\sum_{t=1}^{T} log(p(y_t \mid y_1^{t-1}, x))$$

- KD loss for abstractive summarization

$$\mathrm{L}_{\mathrm{KD}} = \sum_{t=1}^{T} \mathrm{KL}(p_T(y_t \mid y_1^{t-1}, x), p_S(y_t \mid y_1^{t-1}, x))$$

- Final loss for abstractive summarization

$$\mathrm{L}_{\mathrm{FINAL}} = (1-\lambda)\mathrm{L}_{\mathrm{NLL}} + \lambda \mathrm{L}_{\mathrm{KD}}$$

Noisy Self-Knowledge Distillation

- To make summarization systems robust to noise in existing datasets
- Introduce noise to both distillation signals and training data
    1. Noisy Teacher

Noisy Self-Knowledge Distillation

- To make summarization systems robust to noise in existing datasets
- Introduce noise to both distillation signals and training data

1. Noisy Teacher
   - Dropout is kept active while generating teacher predictions
   - The teacher generates variable supervision labels
   - The teacher can also be considered as approximating an average ensemble from many neural networks

Noisy Self-Knowledge Distillation

- To make summarization systems robust to noise in existing datasets
- Introduce noise to both distillation signals and training data

  1. Noisy Teacher
     - Dropout is kept active while generating teacher predictions
  2. Noisy Student

Noisy Self-Knowledge Distillation

- To make summarization systems robust to noise in existing datasets
- Introduce noise to both distillation signals and training data

1. Noisy Teacher
   - Dropout is kept active while generating teacher predictions
2. Noisy Student
   - Inject noise into the training data
   - Word Drop, Word Replacement, Sentence Drop

Noisy Self-Knowledge Distillation

- To make summarization systems robust to noise in existing datasets
- Introduce noise to both distillation signals and training data
   1. Noisy Teacher
      - Dropout is kept active while generating teacher predictions
   2. Noisy Student
      - Inject noise into the training data

$$\mathrm{L}_{\mathrm{KD}} = \sum_{t=1}^{T} \mathrm{KL}(\tilde{p}_T{}^{\alpha}(y_t \mid y_1^{t-1}, x), p_S(y_t \mid y_1^{t-1}, \tilde{x}))$$

Experiments

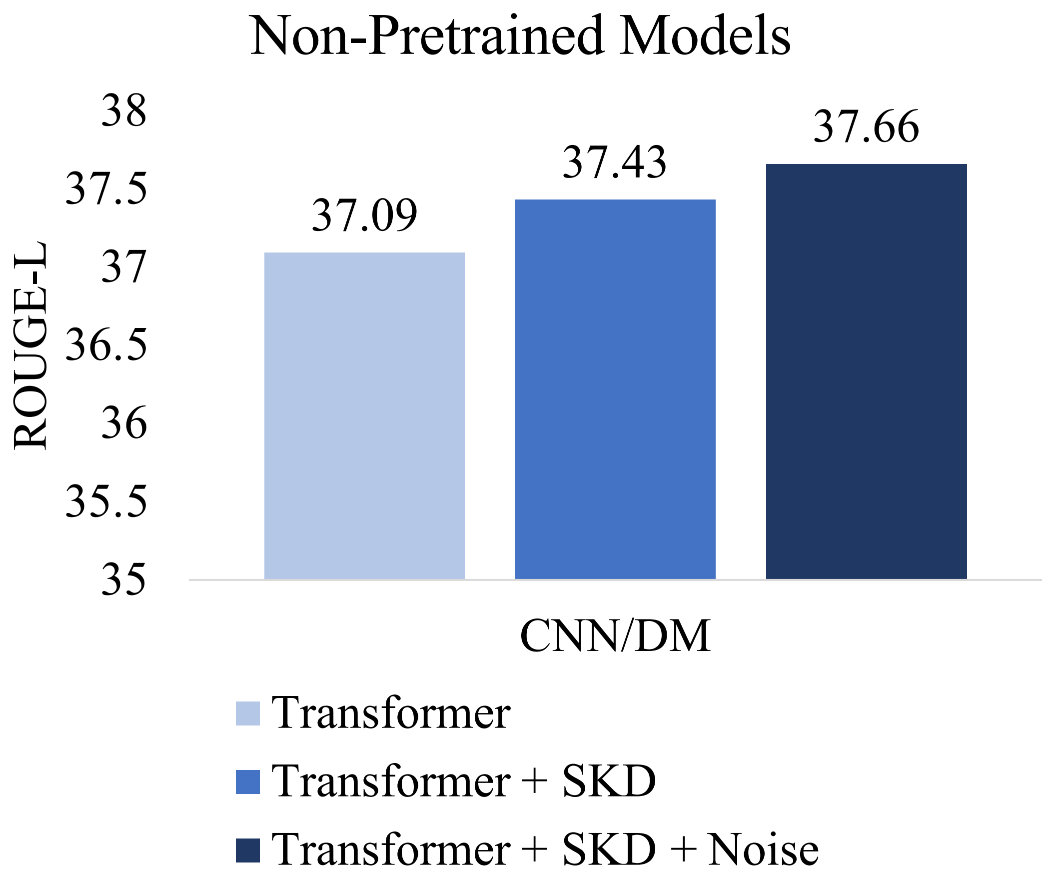
Single-document Summarization Datasets

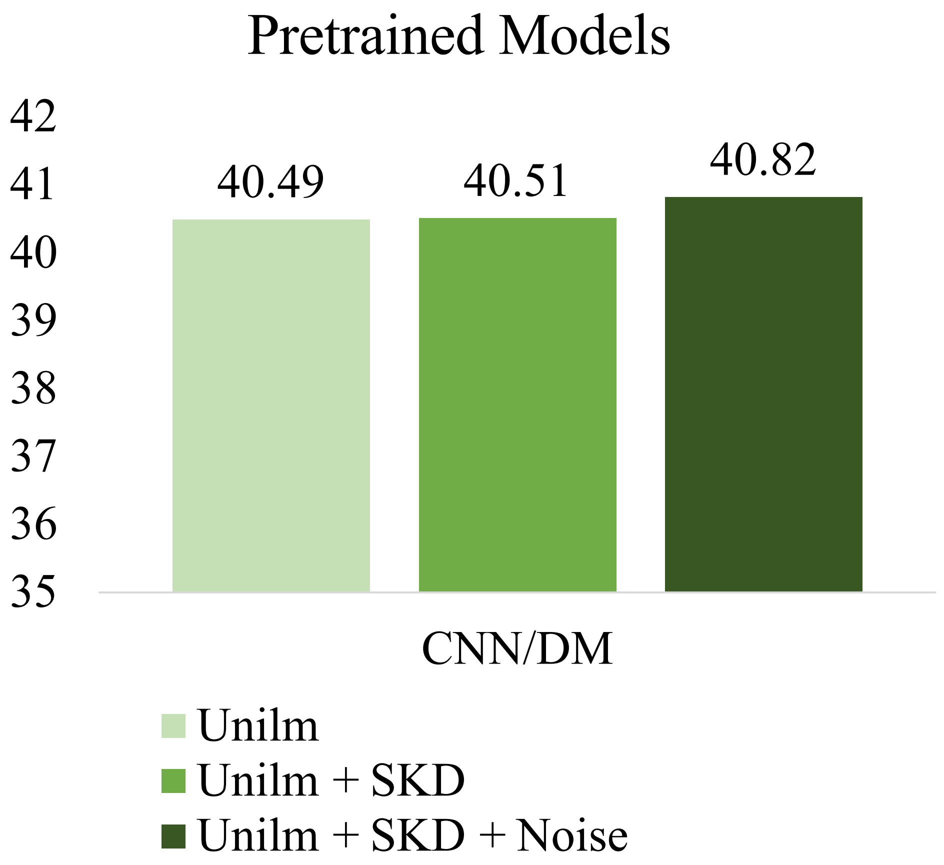
	# docs (train/val/test)	avg. doc length	avg. summary length
CNN	90,266/1,220/1,093	760.50	45.70
DailyMail	196,961/12,148/10,397	653.33	54.65
XSum	204,045/11,332/11,334	431.07	23.26

Experiments
Multi-document Summarization Dataset

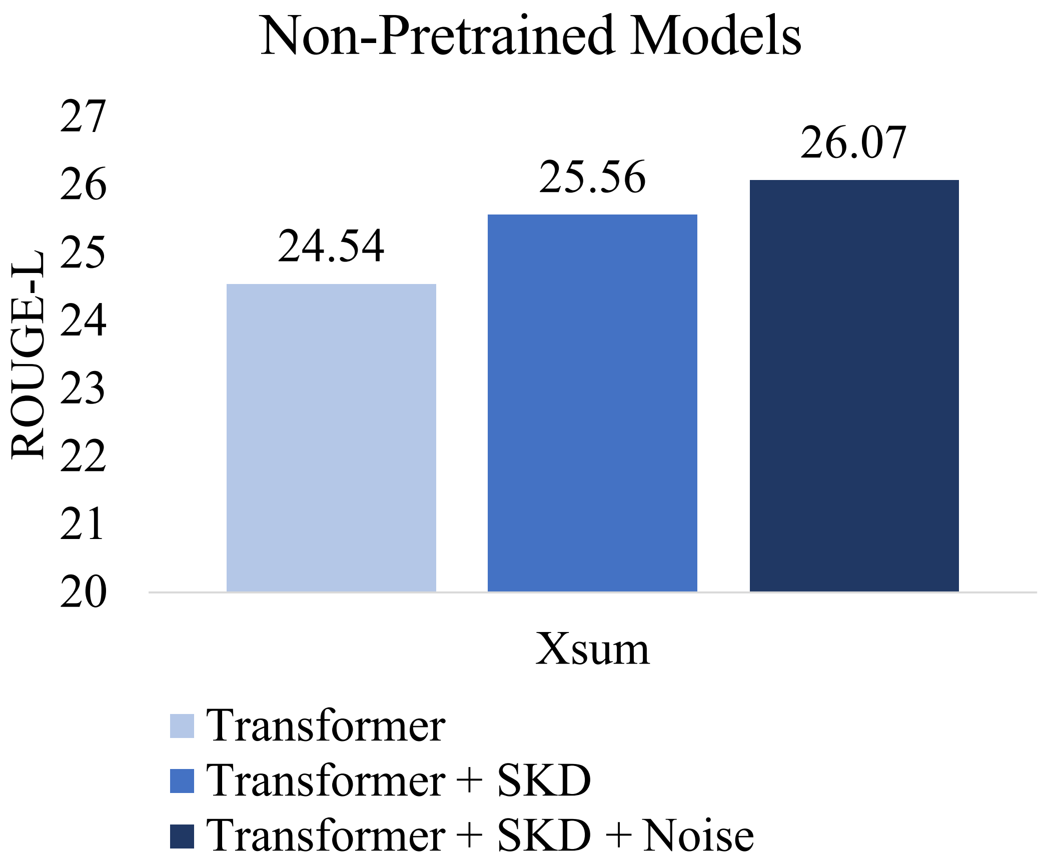
	Category	# instances	avg. summary length	
			sents	words
WikiCatSum	Company	62,545	5.09	124.20
	Film	59,973	4.17	98.16
	Animal	60,816	4.71	92.69

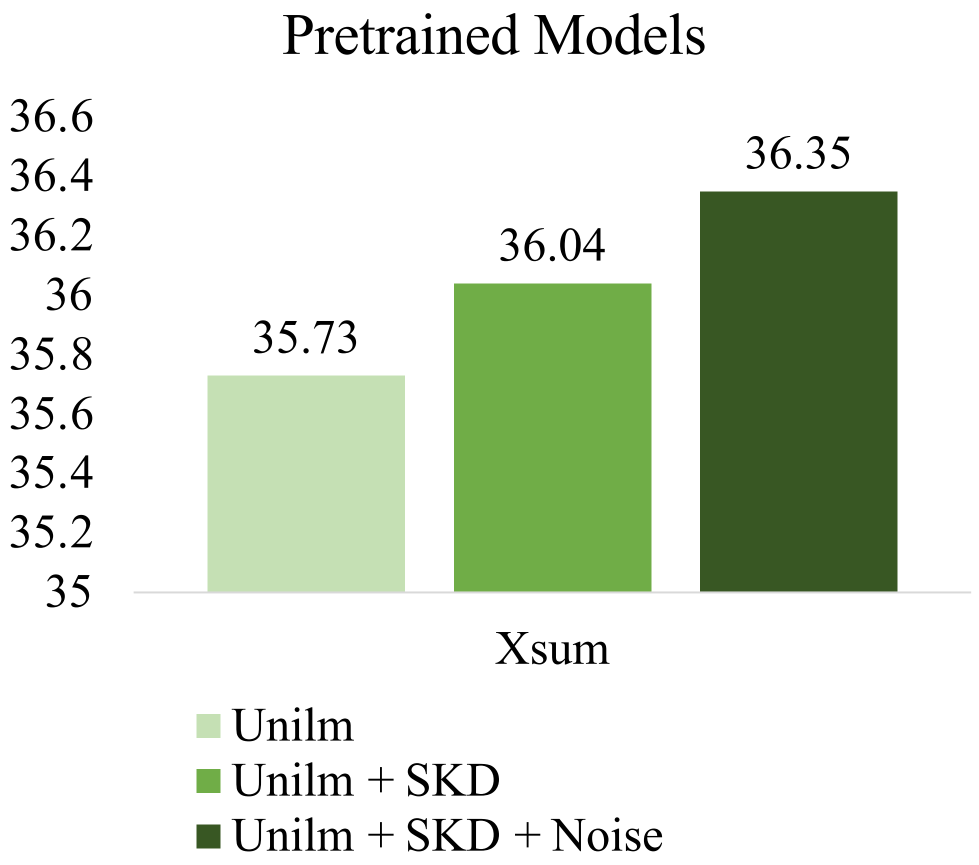
Experiments
CNN/DM Results



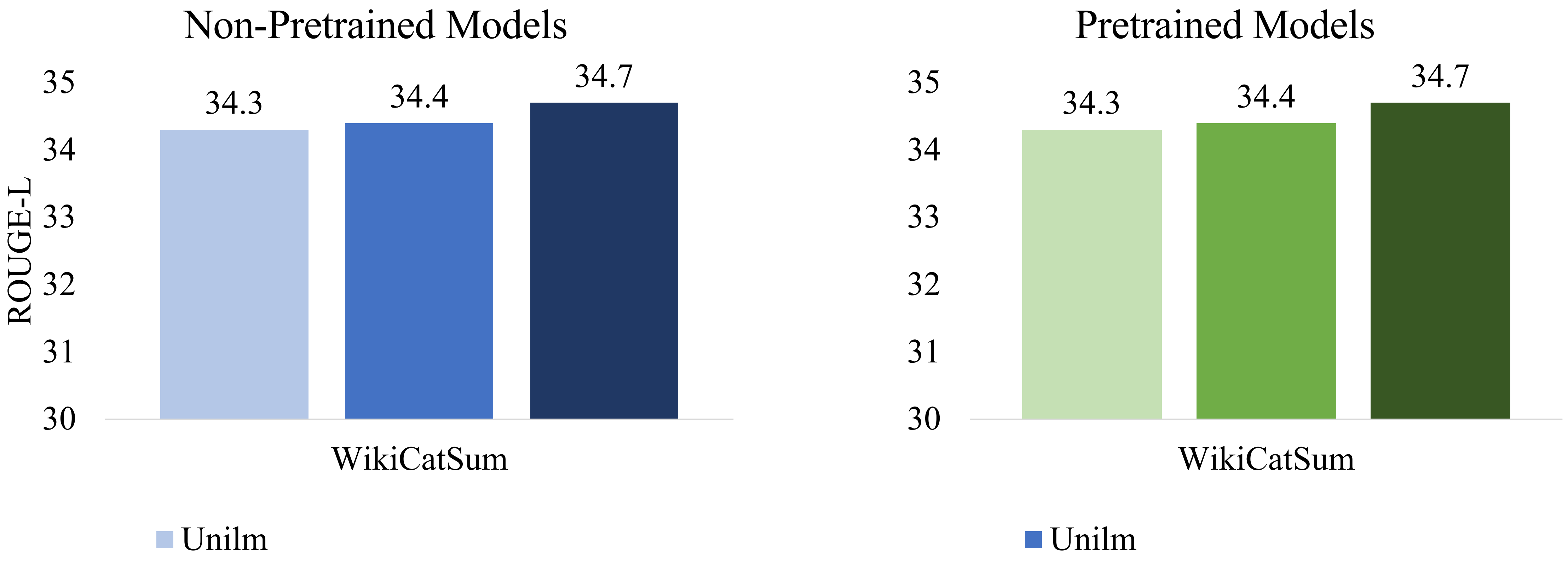


Experiments
XSum Results



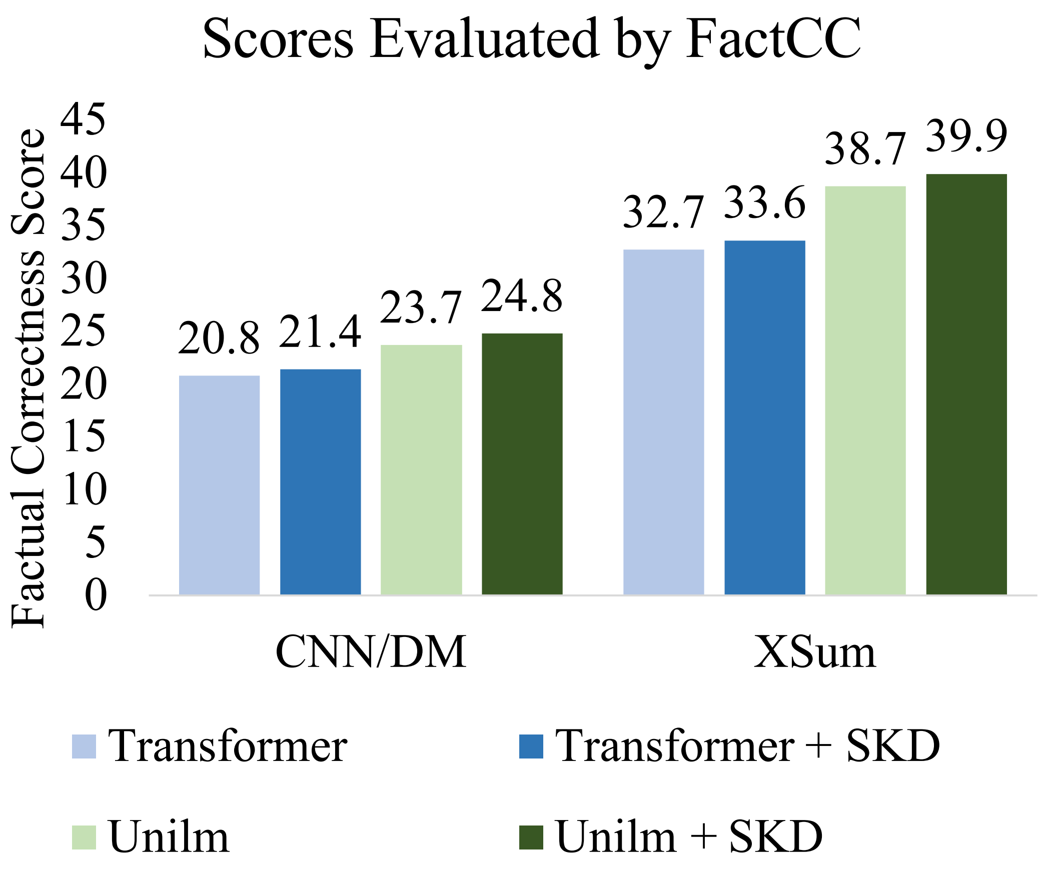


Experiments
WikiCatSum Results



Experiments

Factual Correctness Evaluation



Conclusions

- Self-Knowledge Distillation can alleviate problems associated with maximum-likelihood training in summarization tasks.
- Noise Injection (in the training signal and training data) can help regularize training and further boost performance.