Keyphrase Generation with Fine-Grained Evaluation-Guided Reinforcement Learning

Yichao Luo*, Yige Xu*, Jiacheng Ye, Xipeng Qiu, Qi Zhang†

Fudan NLP Group, Fudan University, China

Findings of EMNLP 2021

Outline

- Background and Motivation
- Our Method and Framework
- Main Results
- Conclusion

Background and Motivation

- Task Description for Keyphrase Generation
  - The source document is a summary of an article (e.g., abstract for scientific papers)
  - The target is some keywords that represent the core information of the source document
  - The **red** words represent the present keyphrases that appear in the source document
  - The **blue** words represent the absent keyphrases that cannot be found in the source document

Document: Rental software valuation in it investment decisions. The growth of application service providers (asps) is very rapid, leading to a number of options to organizations interested in developing new information technology services. ... Likewise, newer risks are associated with asps, including pricing variability. Some of the more common capital budgeting models may not be appropriate in this volatile marketplace. However, option models allow for many of the quirks to be considered. ...

Keyphrase labels: (present keyphrases) application service providers; options; capital budgeting; (absent keyphrases) information technology investment; stochastic processes; risk analysis

Background and Motivation

➢ Motivation

- Traditional F1 score only considers the exact match predictions

Score(“natural language processing”, “language understanding”) = Score(“natural language processing”, “apple tree”) = 0

⇒ **Is this suitable?**

- Keyphrases are short, therefore it is not suitable for n-gram-based metrics

⇒ **Is there any fine-grained metric for a smooth evaluation?**

Our Method and Framework

- Fine-Grained Score (FG-Score)
  - Token-level F1 Score
  - Token-level Edit Distance
  - Repetition Rate Penalty
  - Generation Quantity Penalty

Our Method and Framework

- Fine-Grained Score (FG-Score)
  - Token-level F1 Score
    - For the predicted keyphrase and the ground truth, compute the F1 score in token level
  - Token-level Edit Distance
  - Repetition Rate Penalty
  - Generation Quantity Penalty

Our Method and Framework

- Fine-Grained Score (FG-Score)
  - Token-level F1 Score
  - Token-level Edit Distance
    - Use dynamic programming to compute the edit distance in token level and then re-normed by the target length
  - Repetition Rate Penalty
  - Generation Quantity Penalty

Our Method and Framework

- Fine-Grained Score (FG-Score)
  - Token-level F1 Score
  - Token-level Edit Distance
  - Repetition Rate Penalty
    - Prevent from generating similar keyphrases
    - Penalize when the predicted words appear more times than that in the ground truth
  - Generation Quantity Penalty

Our Method and Framework

- Fine-Grained Score (FG-Score)
  - Token-level F1 Score
  - Token-level Edit Distance
  - Repetition Rate Penalty
  - Generation Quantity Penalty
    - Prevent from generating keyphrases only with high confidence
    - Penalize when the number of the predicted keyphrases is not equal to the number of the ground truth

Our Method and Framework

- Fine-Grained Score (FG-Score)

Token-level:

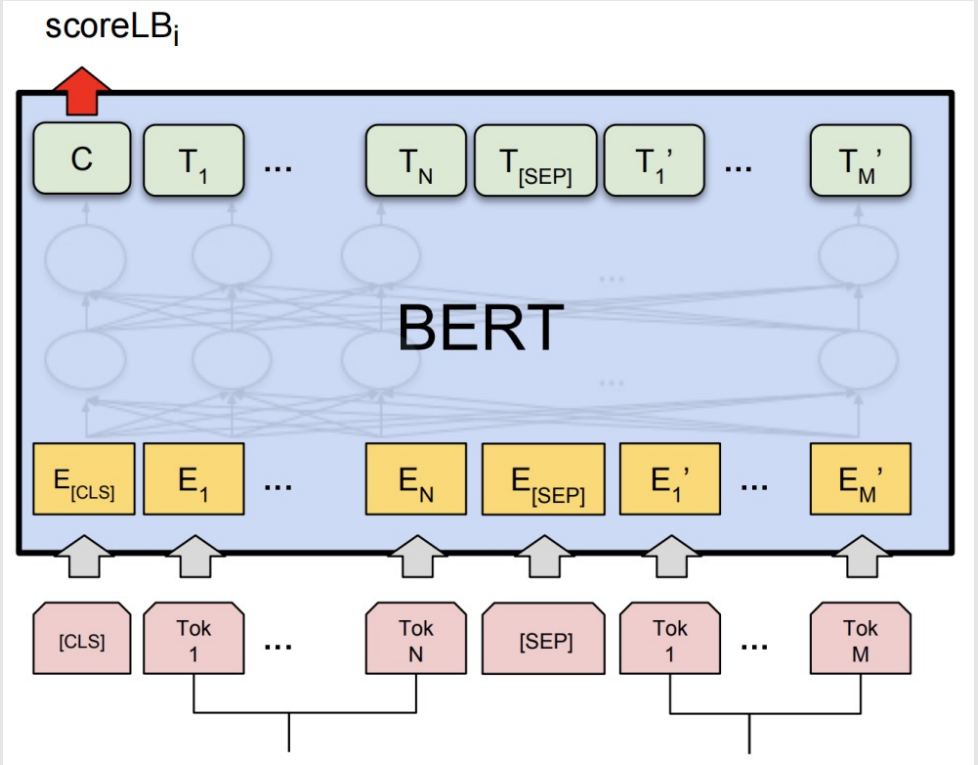
- Token-level F1 Score
- Token-level Edit Distance

Instance-level:

- Repetition Rate Penalty
- Generation Quantity Penalty

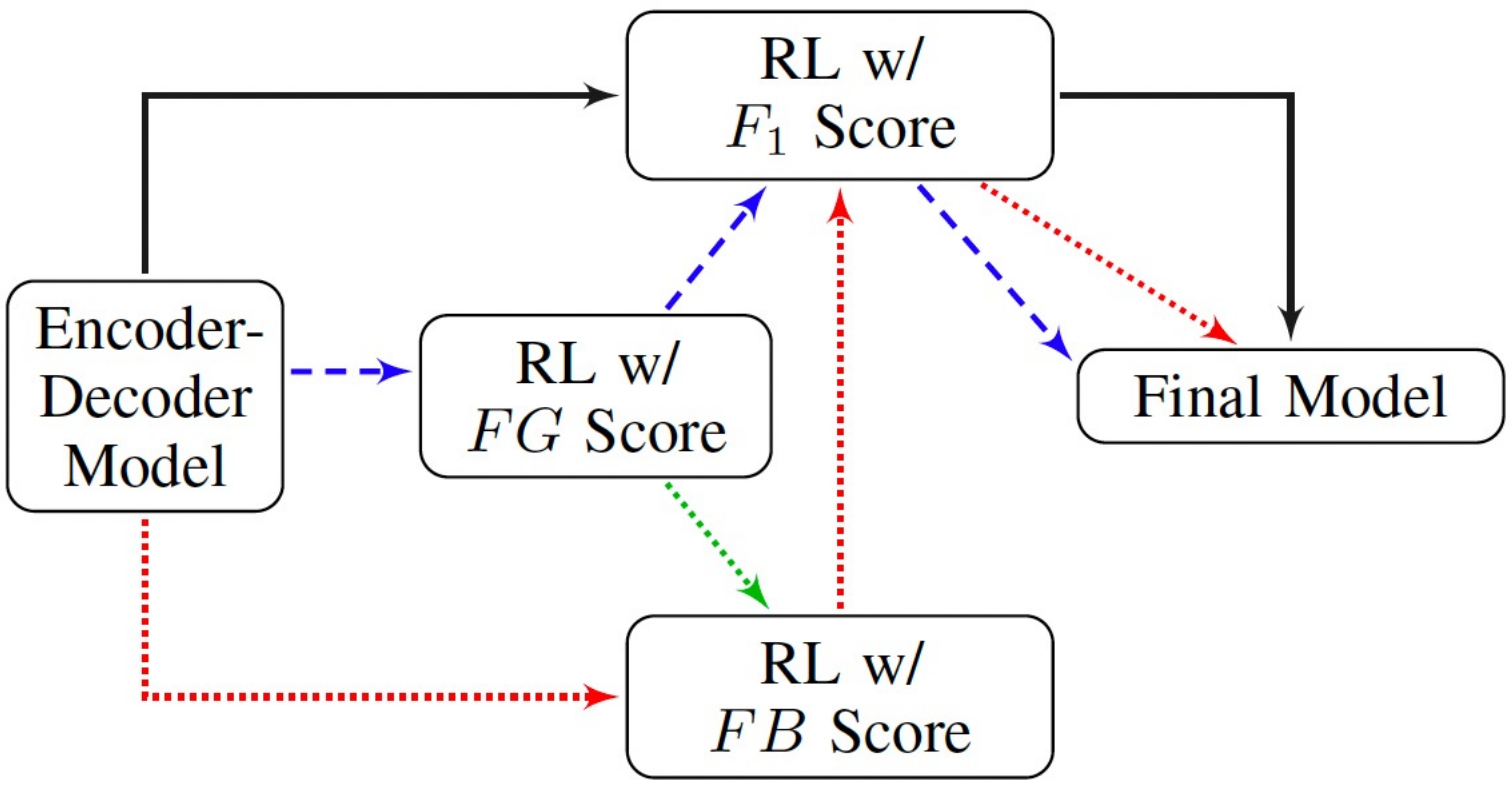
Our Method and Framework

- Continuous Scorer (FB-Score)
  - Smoother
  - Higher score in synonyms



Our Method and Framework

- Reinforcement Learning Framework



Main Results

Model	Inspec			Krapivin			KP20k		
	$F_1@M$	$F_1@5$	FG	$F_1@M$	$F_1@5$	FG	$F_1@M$	$F_1@5$	FG
catSeq(Yuan et al., 2020)	0.262	0.225	0.381	0.354	0.269	0.352	0.367	0.291	0.371
catSeqD(Yuan et al., 2020)	0.263	0.219	0.385	0.349	0.264	0.350	0.363	0.285	0.369
catSeqCorr(Chen et al., 2018)	0.269	0.227	0.391	0.349	0.265	0.360	0.365	0.289	0.374
catSeqTG(Chen et al., 2019)	0.270	0.229	0.391	0.366	0.282	0.344	0.366	0.292	0.369
SenSeNet(Luo et al., 2020)	0.284	0.242	0.393	0.354	0.279	0.355	0.370	0.296	0.373
ExHiRD-h(Chen et al., 2020)	0.291	0.253	0.395	0.347	0.286	0.354	0.374	0.311	0.375
Utilizing RL (Chan et al., 2019)									
catSeq+RL(F_1)	0.300	0.250	0.382	0.362	0.287	0.360	0.383	0.310	0.369
catSeqD+RL(F_1)	0.292	0.242	0.380	0.360	0.282	0.357	0.379	0.305	0.377
catSeqCorr+RL(F_1)	0.291	0.240	0.392	0.369	0.286	0.376	0.382	0.308	0.377
catSeqTG+RL(F_1)	0.301	0.253	0.389	0.369	0.300	0.344	0.386	0.321	0.370
Ours									
catSeq*+RL(FG)	0.252	0.201	0.460	0.359	0.228	0.413	0.365	0.290	0.440
catSeq*+RL(FB)	0.254	0.200	**0.463**	0.354	0.230	**0.416**	0.366	0.291	**0.444**
catSeq*+2RL(FG)	0.308	0.266	0.425	**0.375**	0.304	0.389	0.391	0.327	0.381
catSeq*+2RL(FB)	**0.310**	**0.267**	0.430	0.374	**0.305**	0.390	**0.392**	**0.330**	0.383

Table 1: Result of present keyphrase prediction on three datasets. "RL" denotes that a model is trained by one-stage reinforcement training. "2RL" denotes that a model is trained by two-stage RL training. The notation in parentheses denotes the reward function in first RL training stage. All second reward function in two-stage RL training is F_1 score. "catSeq*" represents that we select the best model of four different catSeq-based baseline models. FB indicates that the reward is computed by the continuous BERT scorer. The underline numbers represent the best result in previous work. FG is the metric we propose.

- RL with the F1 score has similar FG score compared to the baseline model.
- First train with FG or FB score and then train with F1 score can achieve the best result.
- Using continuous FB score usually performs better.

Conclusion

- We formulate a **Fine-Grained Evaluation Score (FG-Score)** for Keyphrase Generation.
- We propose a **two-stage reinforcement learning** training framework with our fine-grained evaluation metric for Keyphrase Generation task.
- **SOTA results** on several datasets. Moreover, it proves that it is necessary to deal with the semantic similarities between predictions and targets.

Thank You for Listening

Code:

https://github.com/xuyige/FGRL4KG