KPQA: A Metric for Generative Question Answering Using Keyphrase Weights

Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt,
Doo Soon Kim, Trung Bui, Joongbo Shin and Kyomin Jung
Seoul National University MILAB & Adobe Research

NAACL 2021

Introduction

Problem

Context : ... , this process, called hypothesis testing, consists of **four steps**. , ...

Question : **How many steps** are involved in a hypothesis test?
Reference Answer : **Four steps** are involved in a hypothesis test.
Generated Answer : There are **seven steps** involved in a hypothesis test .

Human Judgment : 0.063

BLEU-1 : 0.778 **BLEU-1-KPQA** : 0.057
ROUGE-L : 0.713 **ROUGE-L-KPQA** : 0.127

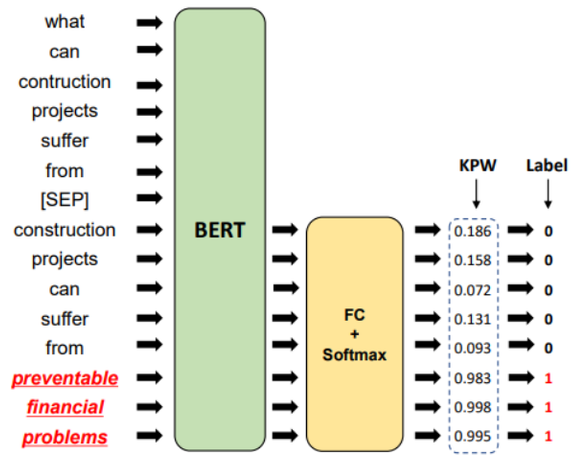
- Widely used **n-gram similarity metrics does not align with human judgments of correctness** in evaluating generative question answering(GenQA) systems.

Contributions

- We propose KPQA metric, an **importance weighting based evaluation metric for GenQA**.
- We collect high-quality human judgments of correctness for the model generated answers on MS-MARCO and AVSD, where those two GenQA datasets aim to generate sentence-level answers.
- We show that **our proposed metric has a dramatically higher correlation with human judgments** than the previous metrics for these datasets.
- We verify the robustness of our metric in various aspects such as question type and domain effect

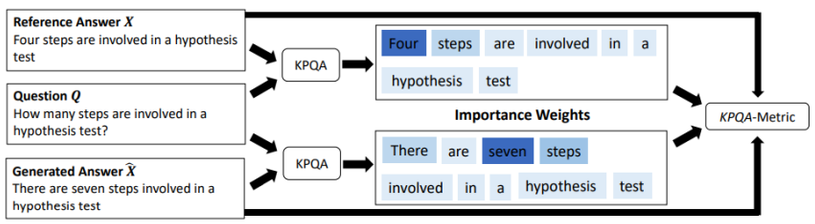
Methodology

KPQA



- KPQA(Keyphrase Predictor for Question Answering) classifies whether each word in the answer sentences is in the answer span for a given question.
- We use the **output probability of KPQA**, KPW(KeyPhrase Weights) as an importance weight.

KPQA-metric



- Importance weights are computed by pre-trained KPQA for each question-answer pair.
- Then, these **weights are integrated into existing metrics** such as BLEU, ROUGE and BERTScore to compute weighted similarity.

Experiments

Comparison with Existing Metrics

Dataset	MS-MARCO		AVSD		NarrativeQA		SemEval	
Metric	r	ρ	r	ρ	r	ρ	r	ρ
BLEU-1	0.349	0.329	0.580	0.562	0.634	0.643	0.359	0.452
BLEU-4	0.193	0.244	0.499	0.532	0.258	0.570	-0.035	0.439
ROUGE-L	0.309	0.301	0.585	0.566	0.707	0.708	0.566	0.580
METEOR	0.423	0.413	0.578	0.617	0.735	0.755	0.543	0.645
CIDEr	0.275	0.278	0.567	0.600	0.648	0.710	0.429	0.595
BERTScore	0.463	0.456	0.658	0.650	**0.785**	0.767	0.630	0.602
BLEU-1-KPQA	0.675	0.634	0.719	0.695	0.716	0.699	0.362	0.462
ROUGE-L-KPQA	**0.698**	0.642	0.712	0.702	0.774	0.750	**0.742**	**0.687**
BERTScore-KPQA	0.673	**0.655**	**0.729**	**0.712**	0.782	**0.770**	0.741	0.676

- Observe a significantly higher correlation score for our proposed KPQA-metric compared to existing metrics on all benchmark datasets.

Weight Visualization

Context	... , it can take 5-20 hours of walking to lose 1 pound ... , ...
Question	How long do i need to walk in order to loose a pound ?
Reference	Walk for 5 to 20 hours to lose 1 pound .
IDF	Walk for 5 to 20 hours to lose 1 pound .
KPW	Walk for 5 to 20 hours to lose 1 pound .

Human Judgment: 0.94, **BERTScore**: 0.72, **BERTScore-KPQA**: 0.93

UniLM	You need to walk for 5 to 20 hours in order to loose a pound .
IDF	You need to walk for 5 to 20 hours in order to loose a pound .
KPW	You need to walk for 5 to 20 hours in order to loose a pound .

- Compared to other imporatance weighting method IDF, our KPQA integrated metric assigns dynamic weights to words in the answer sentence using the context of the question.

Reference

[1] A. Chen et al., Evaluating Question Answering Evaluation, MRQA 2019
[2] T.Zhang et al., BERTScore: Evaluating Text Generation with BERT, ICLR 2020