KPQA: A Metric for Generative Question Answering Using Keyphrase Weights

Hwanhee Lee[1], Seunghyun Yoon[2], Franck Dernoncourt[2],

Doo Soon Kim[3], Trung Bui[2] , Joongbo Shin[1] , Kyomin Jung[1]

[1]Seoul National University, Seoul, Korea
[2]Adobe Research, San Jose, CA, USA
[3]Roku Inc., San Jose, CA, USA

Generative Question Answering Evaluation

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

- Different from extractive question answering, generation question answering(GenQA) is a task of **generating free-form answers** for the given question

- For this task, n-gram similarity metrics such as BLEU are widely used to evaluate the generated answers

Motivation

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 **fail to capture the correctness** of the generated answer because they equally consider each word in the sentence
- Although the answer is incorrect, it can receive high BLEU, ROUGE scores

Research Goal

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

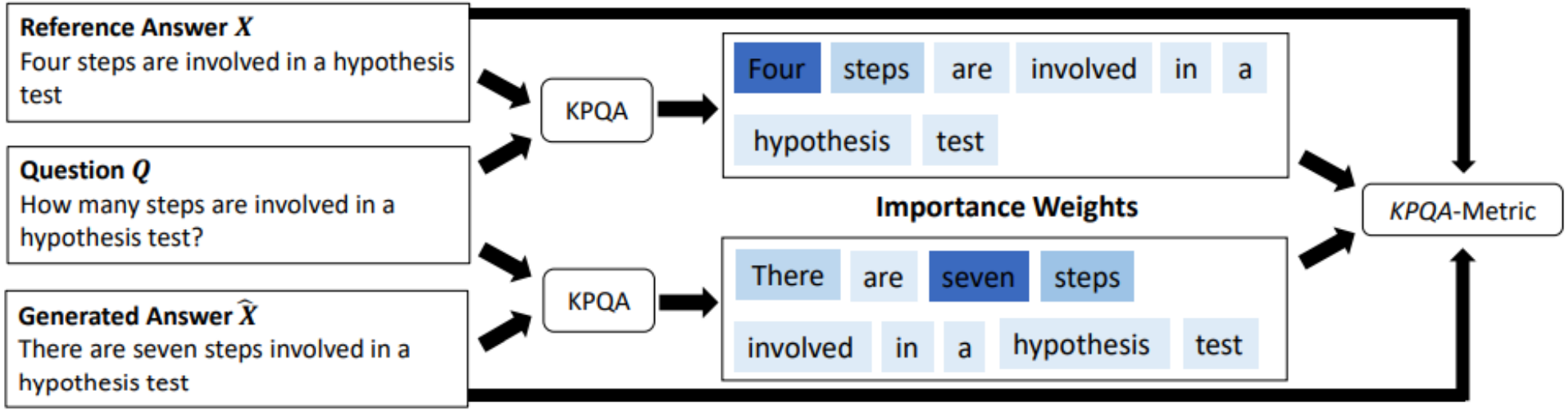
Question : **How many steps** are involved in a hypothesis test?
Reference Answer : **Four steps** are involved in a hypothesis test.
Generated Answer : There are **seven steps** involved in a hypothesis test .

Human Judgment : 0.063

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

- Develop an evaluation metric for GenQA that can consider the **importance of each word** in the generated answer
- By focusing on the keyphrases for the question, the metric can judge the **correctness** of the answer

Proposed KPQA-Metric

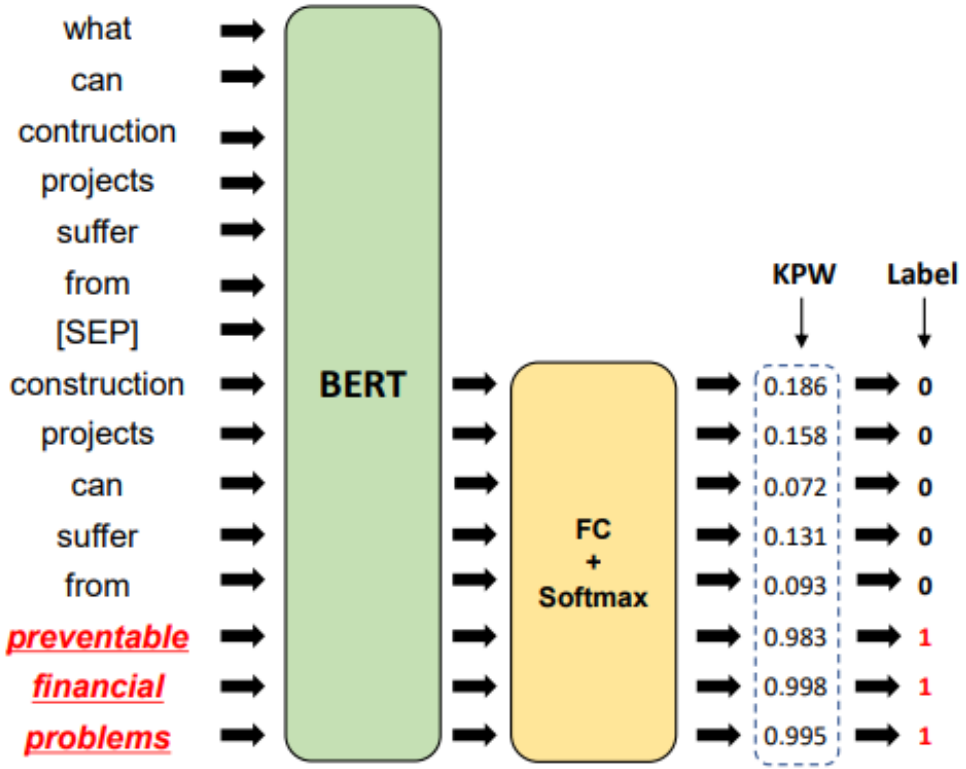


- Propose a new metric called **KPQA-metric** for evaluating GenQA systems.

- We first develop KPQA(Keyphrase Predictor for Question Answering), a **keyphrase prediction model to compute the importance weights**

- Using the weights from the KPQA, we **integrate them into existing metrics** such as BLEU, ROUGE and BERTScore to compute **weighted similarity.**

KPQA



- KPQA is a BERT based token classification model that classifies whether each token in the answer sentences is in the answer span for a given question.

- The answer sentences contain answer span for the given question like the answers in GenQA systems.

Constructing Training Set for KPQA

- Extracting the sentences that have answer span in “extractive QA” dataset such as SQuAD
- Construct *{question, answer span, answer sentence}* pair

Most early Greeks did not even consider 1 to be a number, so they could not consider it to be a prime. By the Middle Ages and Renaissance many mathematicians included 1 as the first prime number. In the mid-18th century Christian Goldbach listed 1 as the first prime in his famous correspondence with Leonhard Euler -- who did not agree. In the 19th century many mathematicians still considered the number 1 to be a prime. For example, Derrick Norman Lehmer's list of primes up to 10,006,721, reprinted as late as 1956, started with 1 as its first prime. Henri Lebesgue is said to be the last professional mathematician to call 1 prime. By the early 20th century, mathematicians began to accept that 1 is not a prime number, but rather forms its own special category as a "unit".

Question: Who included 1 as the first prime number in the mid 18th century?

Answer: Christian Goldbach

Answer Sentence: In the mid-18th century Christian Goldbach listed 1 as the first prime in his famous correspondence with Leonhard Euler -- who did not agree.

Constructing Training Set for KPQA

- We also use multi-hop QA dataset Hotpot QA to consider multiple sentence answers
- We extract supporting sentences after coreference resolution

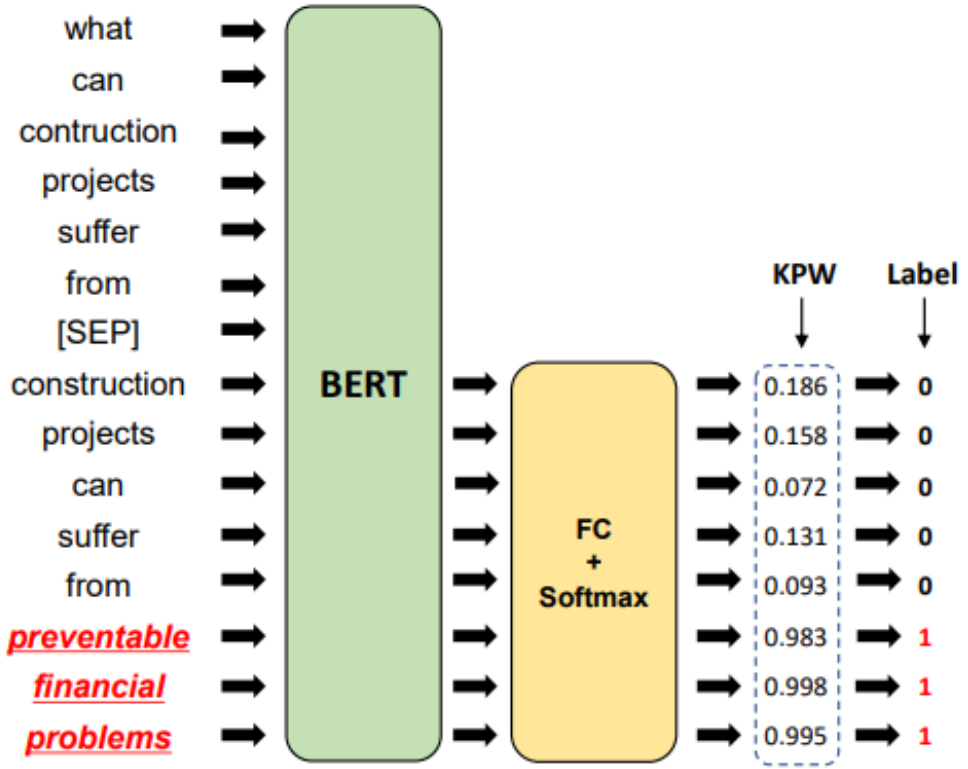
The Glory of Tang Dynasty is a 2017 Chinese television series starring Jing Tian and Allen Ren. It is based on the novel "The Concubine of Tang: Legend of Pearl by Cang Mingshui; and tells the fictional love story of Emperor Daizong and Consort Shen, aided by the grandiose historical background of the Shi Rebellion (755-763).
Jing Tian is a Chinese actress. **She** graduated from the Beijing Dance Academy and Beijing Film Academy. She is known for her roles in war epic "The Warring States" (2011) and the action films "Special ID" and "Police Story 2013" She is part of the cast for three Legendary Pictures films, including a prominent role in "The Great Wall" (2016) as well as "" (2017) and the upcoming.

Question: What dance academy did the starring actress from The Glory of Tang Dynasty graduate from?

Answer: Beijing Dance Academy

Answer Sentences: The Glory of Tang Dynasty is a 2017 Chinese television series starring Jing Tian and Allen Ren. **Jing Tian** graduated from the Beijing Dance Academy and Beijing Film Academy.

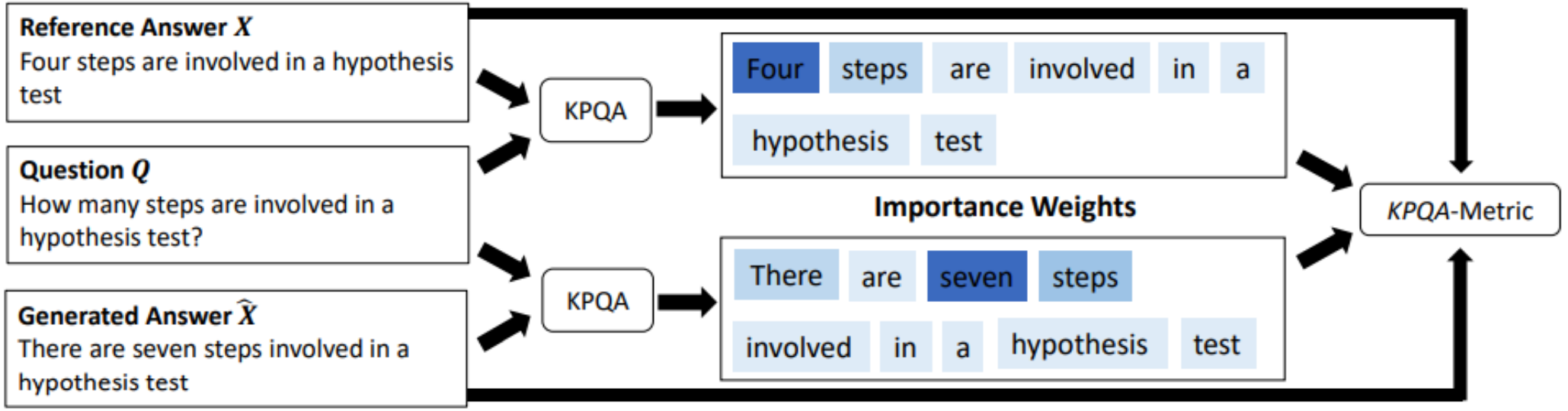
KPQA



- Using the *question, answer span, and answer sentences,* KPQA is trained to mark each token of the answer span in the *answer sentences*

- When we comupute KPQA-metric, we use the output probability of KPQA, KPW(KeyPhrase Weights) as an importance weight.

Computing KPQA-Metric



- Then, we **integrate importance weights into existing metrics** such as BLEU, ROUGE and BERTScore to compute **weighted similarity.**

KPQA-Metric Variants

BLEU-1 KPQA

$$B1 = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} \cdot I(i,j)}{m}$$

1-gram precision

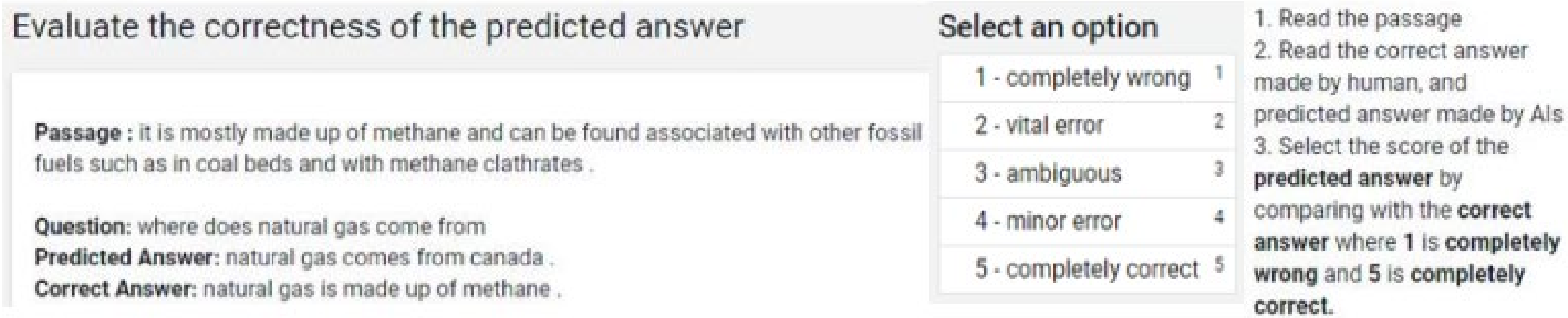
$$B1^{KPQA} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} \mathrm{KPW}_i^{(Q,X)} \cdot I(i,j)}{\sum_{i=1}^{m} \mathrm{KPW}_i^{(Q,X)}}$$

weighted 1-gram precision

I(i,j) : check whether each token is same

- We modify previous metrics to weighted similarity metrics by integrating the importance weights from KPQA
- We introduce BLEU-1-KPQA, ROUGE-L-KPQA and BERTScore-KPQA

Dataset Collection



- To evaluate the proposed metric, we create new human judgments of correctness for model generated answers on two GenQA datasets, MS-MARCO and AVSD.

- We generate answers using state-of-the-art GenQA models on MS-MARCO and AVSD where the target answers are natural sentences rather than short phrases.

- We then collect human judgements of correctness over the 1k generated answers for each dataset.

Experiments: Benchmark Datasets

Dataset	Answer Length (avg.)	# Samples
MS MARCO	16.6	183k
AVSD	9.4	118k
Narrative QA	4.7	47k
SemEval	2.5	14k

- We evaluate our proposed metrics using four datasets on human judgments of correctness including the two datasets from previous works

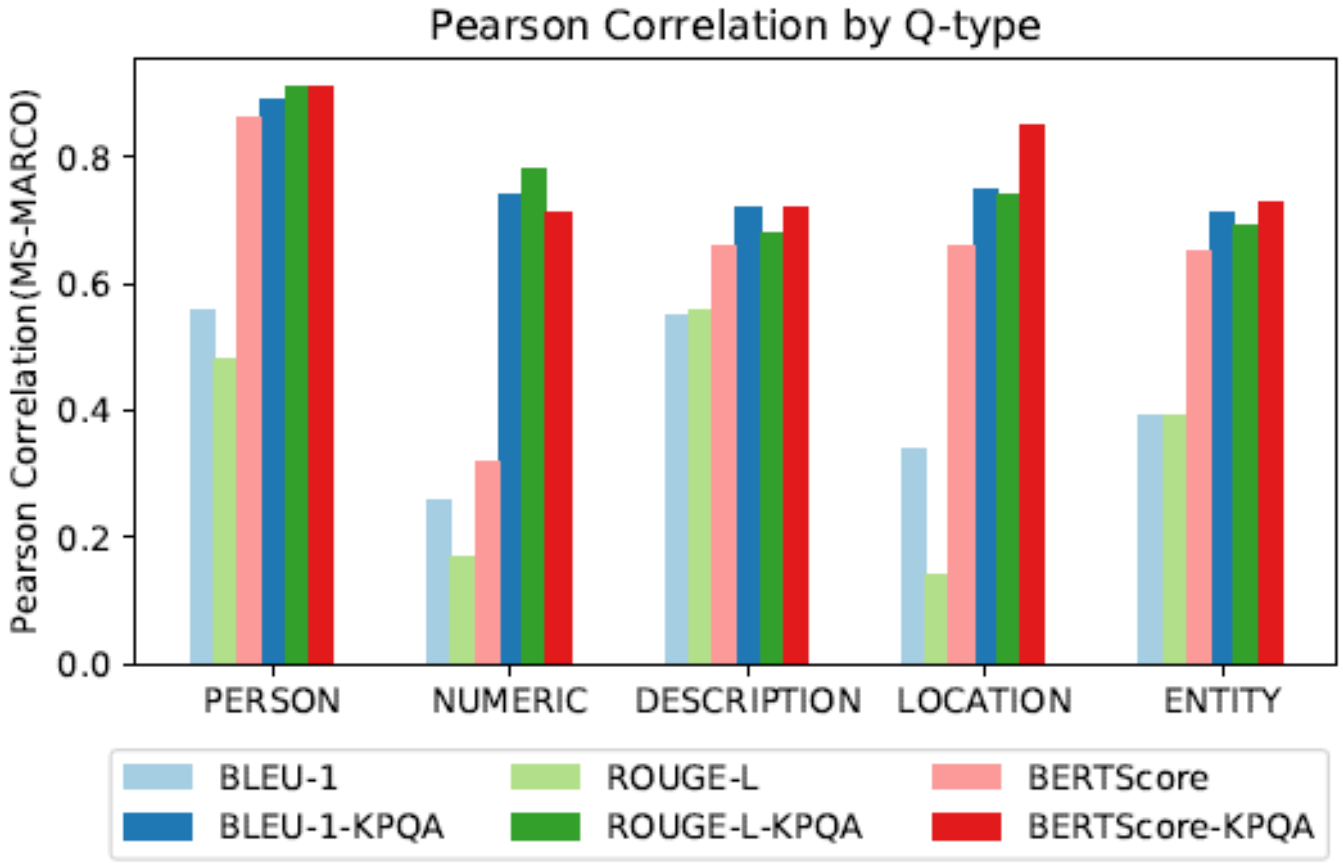
- Our proposed datasets MS-MARCO, AVSD have more longer and abstractive answers than NarrativeQA and SemEval

Experiments: Correlation with Human Judgments

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.

Experiments: Correlation by Question Type



- We split the dataset into five question types and measure the performance of various metrics with Pearson correlation coefficients.

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.

Closing Remarks

- We propose KPQA-metric, an **importance weighting based evaluation metric for GenQA**.

- We collect 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.

Contact: wanted1007@snu.ac.kr

Code: https://github.com/hwanheelee1993/KPQA