Improving Unsupervised Question Answering via Summarization-Informed Question Generation

Chenyang Lyu[1], Lifeng Shang[2], Yvette Graham[3], Jennifer Foster[1], Xin Jiang[2], Qun Liu[2]

[1]School of Computing, Dublin City University
[2]Huawei Noah's Ark Lab
[3]School of Computer Science and Statistics, Trinity College Dublin

EMNLP 2021

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

ML Labs

DCU Ollscoil Chathair Bhaile Átha Cliath Dublin City University

Science Foundation Ireland sfi For what's next

Outline

1. Background
2. Methodology
3. Experiments and Analysis
4. Conclusion

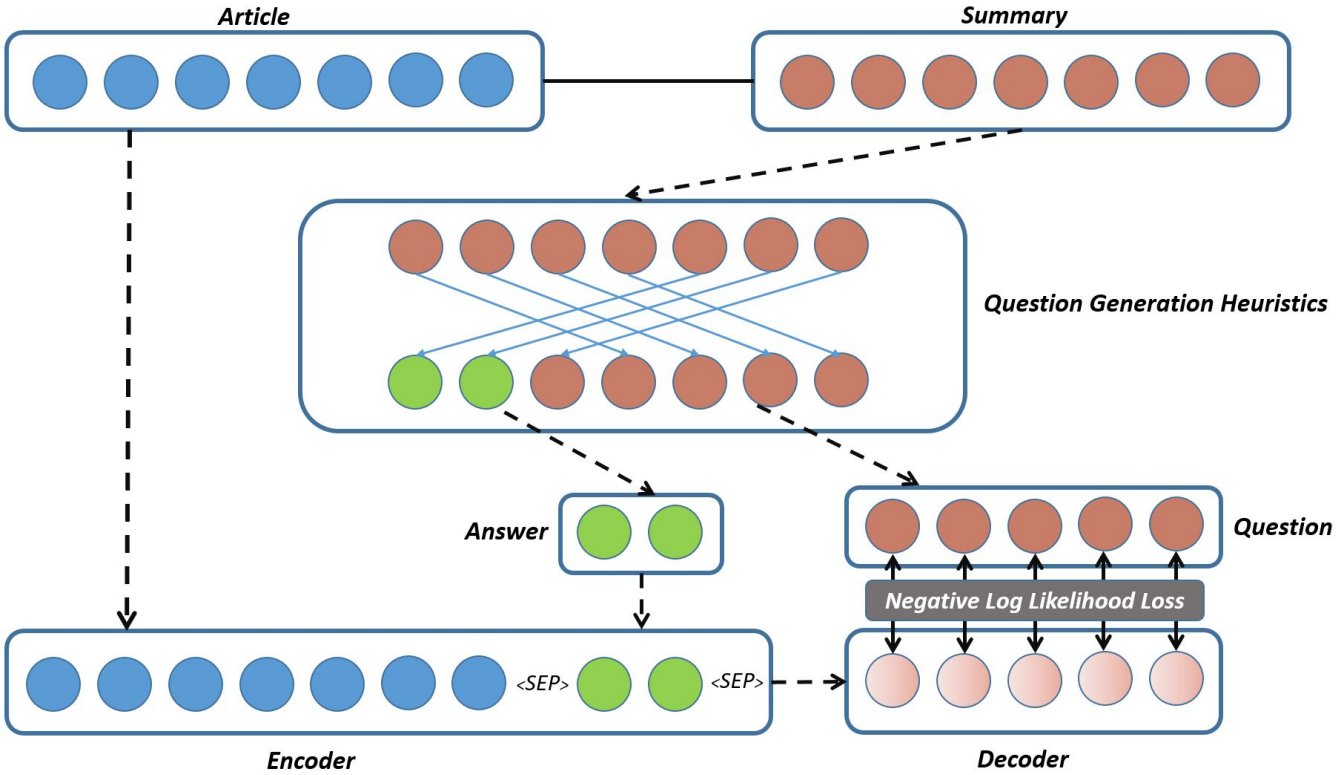
1. Background - Question Generation

- Template-based QG
  - Use hand-crafted rules induced from linguistic knowledge
  - *Shortcoming:*
    - *Generated questions have high lexical overlap with source text*
- Supervised QG
  - Use existing QA datasets to train a QG system (typically a neural model).
  - *Shortcoming:*
    - *Rely on the availability of QA dataset which is expensive to obtain and heavily tied to a certain domain and language.*

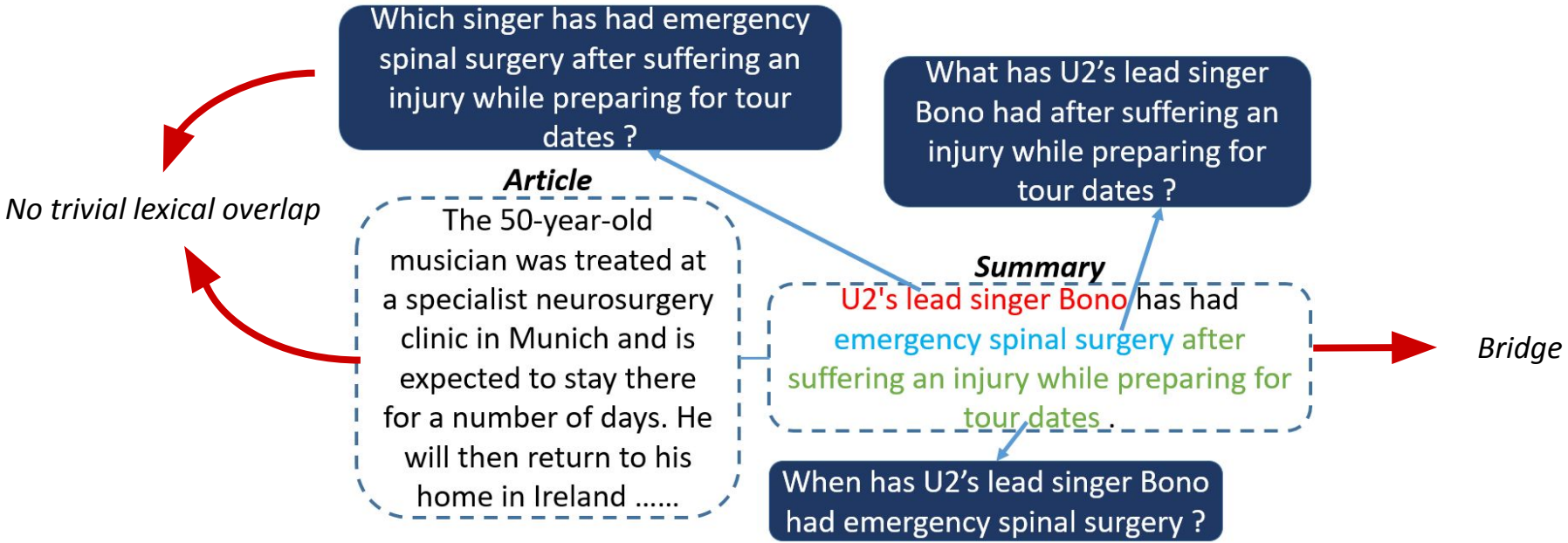
2. Methodology - Summarisation-informed QG

- We propose a summarisation-informed unsupervised QG approach:
  - Employ summary data as a bridge between passage and question
  - Generate questions based on summaries using heuristics
  - Train a QG system using data created above

2. Methodology - Summarisation-informed QG



2. Methodology - Summarisation-informed QG



2. Methodology - Summarisation-informed QG

- Parse the summaries using:
  - *dependency parsing*
  - *named-entity recognition*
  - *semantic role labeling*
- Semantic role labeling:
  - [U2's lead singer Bono ***ARG-0***] has [had ***VERB***] [emergency spinal surgery ***ARG-1***] [after suffering an injury while preparing for tour dates ***ARG-TMP***]

2. Methodology - Summarisation-informed QG

Which singer has had emergency spinal surgery after suffering an injury while preparing for tour dates?

What has U2's lead singer Bono had emergency spinal surgery after suffering an injury while preparing for tour dates?

[U2's lead singer Bono ***ARG-0***] has [had ***VERB***] [emergency spinal surgery ***ARG-1***] [after suffering an injury while preparing for tour dates ***ARG-TMP***]

When has U2's lead singer Bono had emergency spinal surgery ?

2. Methodology - Summarisation-informed QG

- QG heuristics:

$S = summary$
$srl\_frames = SRL(S)$
$ners = NER(S)$
$dps = DP(S)$
$examples = []$
for *frame in srl_frames* **do**
  $root\_verb = dps_{root}$
  $verb = frame_{verb}$
  if *root_verb* ***equal to*** *verb* **then**
    for *arg in frame* **do**
      $wh* =$
        $identify\_wh\_word(arg, ners)$
      $base\_verb, auxs =$
        $decomp\_verb(arg, dps, root\_verb)$
      $Q_{arg} =$
        $wh\_move(S, wh*, base\_verb, auxs)$
      $Q_{arg} = post\_edit(Q_{arg})$
      $examples.append(context, Q_{arg}, arg)$
    end
  end
end

2. Methodology - Summarisation-informed QG

- Use <***passage***, *question, extracted answer*> triples to train QG system
  - Input: *passage <SEP> extracted answer <SEP>*
  - Output target: *question*
  - Training objective:

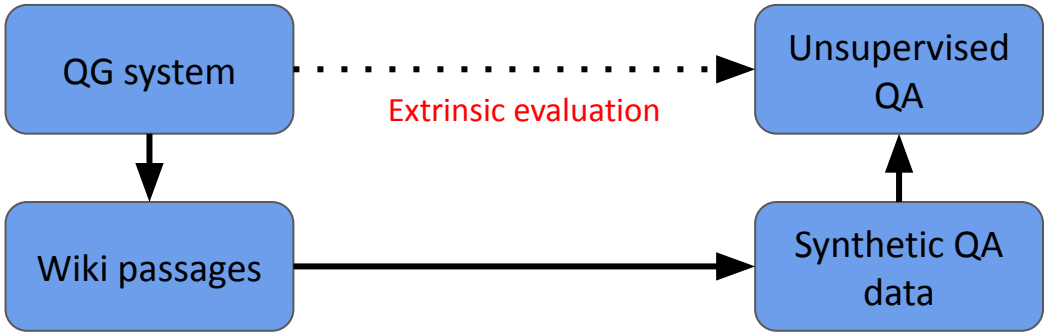
$$L = -\sum_{i=1}^{N} logP(q_i|C, A)$$

3. Experiments and Analysis

- Datasets
  - Summarisation data: XSUM (free available from BBC News)
  - QA data:
    - Wikipedia-based: SQuAD1.1, Natural Questions, TriviaQA
    - Other domain: NewsQA, DuoRC, BioASQ

3. Experiments and Analysis

- How to evaluate our QG system?
  - BLEU, ROUGE and Meteor are not suitable for evaluating QG since with a <passage, answer> pair there could be multiple plausible questions
- We **extrinsically** evaluate our QG system using unsupervised QA
  - We train a QG system using the data generated by QG heuristics
  - Then we use wikipedia passages to generate synthetic QA data using the QG system



3. Experimental Results

We employ our synthetic QA dataset (20k samples) to fine-tune a BERT-large model then evaluate it on SQuAD, Natural Questions and TriviaQA (in-domain)

Models	SQuAD1.1 EM	SQuAD1.1 F-1
SUPERVISED MODELS		
Match-LSTM	64.1	73.9
BiDAF	66.7	77.3
BERT-base	81.2	88.5
BERT-large	84.2	91.1
UNSUPERVISED MODELS		
Lewis et al. (2019)	44.2	54.7
Li et al. (2020)	62.5	72.6
Our Method	**65.6**	**74.5**

Models	NQ EM	NQ F-1	TriviaQA EM	TriviaQA F-1
SUPERVISED MODELS				
BERT-base	66.1	78.5	65.1	71.2
BERT-large	69.7	81.3	67.9	74.8
UNSUPERVISED MODELS				
Lewis et al. (2019)	27.5	35.1	19.1	23.8
Li et al. (2020)	31.3	48.8	27.4	38.4
Our Method	**46.0**	**53.5**	**36.7**	**43.0**

3. Experimental Results

To investigate the transferability of our synthetic QA data, we further apply it on three **out-of-domain** QA datasets, NewsQA, DuoRC and BioASQ

	NewsQA		BioASQ		DuoRC	
	EM	F-1	EM	F-1	EM	F-1
Lewis et al. (2019)	19.6	28.5	18.9	27.0	26.0	32.6
Li et al. (2020)	33.6	46.3	30.3	38.7	32.7	41.1
Our Method	**37.5**	**50.1**	**32.0**	**43.2**	**38.8**	**46.5**

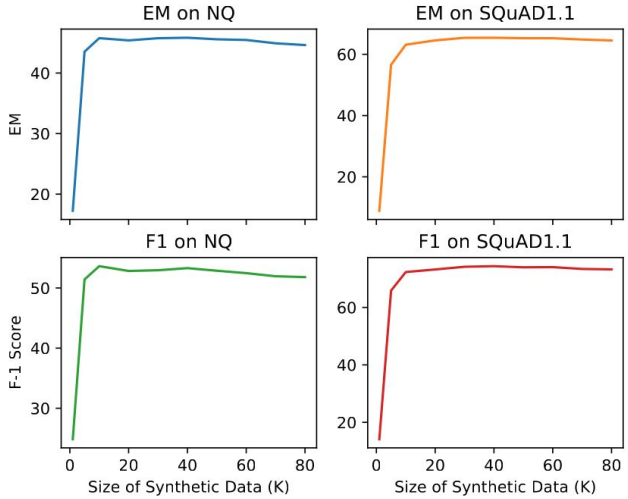
3. Experiments and Analysis

- Naive-QG: no summary, same as a template-QG system
- Summary-QG: employing summary data.
  - Main-verb: only generate questions based on the main verb
  - Wh-movement: move the wh-word to the beginning of sentence
  - Deomp-verb: decompose the verb to its base form and auxiliaries
  - NER-Wh: use NER tags to obtain appropriate question words

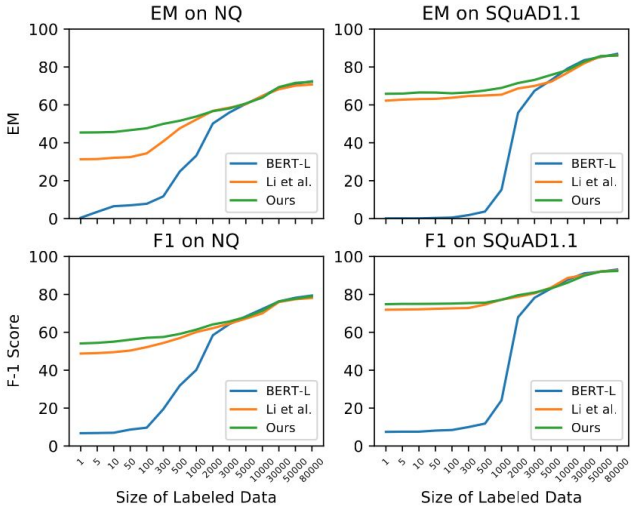
Heuristics	EM	F-1
Naive-QG	31.1	43.3
Summary-QG	50.9	59.4
+Main Verb	53.8	63.6
+Wh-Movement	59.5	67.7
+Decomp-Verb	64.1	73.9
+NER-Wh	**65.4**	**74.8**

3. Experiments and Analysis

Effect of data amount



Few-shot learning



4. Conclusion

- We propose an unsupervised question generation method which uses summarization data
  - 1) To minimize the lexical overlap between passage and question
  - 2) To provide a QA-dataset-independent way of generating questions
- Our unsupervised QA extrinsic evaluation shows that our method substantially outperforms previous methods
- Our synthetic QA data transfers well to out-of-domain datasets

Thanks for listening!