Generating Questions from Wikidata Triples

Kelvin Han[1], Thiago Castro Ferreira[2], Claire Gardent[1]

[1]*CNRS/LORIA, Université de Lorraine*

[2]*aiXplain, inc., Federal University of Minas Gerais*

1. Introduction

Teaching machines to generate a question from knowledge base (KB) triples has many potential uses:

- facilitating KB access by non-experts;
- improving dialog models;
- supporting intelligent tutoring;
- augmenting data for QA systems.

Our motivation is to generate varied questions from a single KB triple.

2. Task and data

Training/inference input format (Elsahar et al, 2018)

(M/03Y2SVR , ASTRONOMY/CELESTIAL_OBJECT/CATEGORY , M/0JVQ), celestial object, category, PLACEHOLDEROBJ designated as PLACEHOLDERSUB

What category of celestial object is 7624 Gluck?

Existing simple questions data: one KB triple-to-one question type, despite many ways of asking for the same answer.

WKDQG

- Questions for a data-to-text benchmark
- Unify three datasets from different KBs into one (Wikidata)
- Varied types and distribution of 346,439 (RDF, question) pairs

Dataset	Knowledge Base	Questions	Question focus	(RDF, Q) pairs	qtype/RDF
SimpleQ (SQ)	Freebase	Crowd-sourced	Object-only	53,624	1.0/1.0/2.0
WebNLG (WQ)	DBpedia		Object/Subject	10,272	2.26/1.0/4.0
ZeroshotRE (ZQ)	Wikidata	Crowd-sourced templates	Object-only	282,543	1.22/1.0/4.0
WKDQG	Wikidata	Varied	Varied	346,439	1.19/1.0/4.0

3. Our approach vs existing

- Encoder-decoder Transformer
  - Pre-trained weights (bart-base), subword tokenization
- Pre/Post-processing
  - ~~Delexicalization~~
  - ~~Property lexicalizations~~
- Input:
  - triple (Wikidata labels)
  - question type (qtype) control
  - question focus position + semantic type

Training/inference input format

(7624 GLUCK , INSTANCE OF , ASTEROID), WHAT, ANSOBJ, category

- Compare with Elsahar et al 2018

SQ: No zero shot constraints

Model	B-4	BSc	R-L	M
RDF-only				
Elsahar	34.01	64.85	61.51	32.67
BART$_{rdf}$	37.05	69.42	65.12	34.22
BART$_{rdf,mtl}$	$\boxed{37.91}$	69.68	65.20	34.55
BART$_{rdf,qt}$	**41.95**	**73.51**	**71.21**	**36.78**
RDF+NL				
Elsahar$_{nl}$	$\boxed{38.13}$	68.63	65.48	34.74
BART$_{rdf+nl}$	38.38	70.00	65.67	34.87
BART$_{rdf+nl,mtl}$	38.10	70.17	65.57	34.73
BART$_{rdf+nl,qt}$	**42.67**	**73.78**	**71.50**	**37.28**

SQ: Zero shot property and entity types

			Sub-type		Obj-type	
Model	B-4	R-L	B4	R-L	B4	R-L
RDF-only						
Elsahar	14.24 (±2.48)	44.30 (±2.66)	29.96 (±2.10)	58.46 (±2.29)	23.94 (±4.34)	53.54 (±3.23)
BART$_{rdf,qt}$	**28.35** ±3.33	**60.84** ±2.67	**37.30** (±1.68)	**67.48** (±1.38)	**35.05** (±3.03)	**66.11** (±1.97)

Overview

Pre-trained language model fine-tuned with varied data and question type control allows:

- generation of varied questions;
- comparable performance (auto metrics and human judgements), without (i) noisy distant supervision for properties; (ii) heavy pre- and post-processing (delexicalization);
- data augmentation for training more robust QA systems.

4. Generating varied questions with WKDQG

- Finetune bart-base with WKDQG
- Replace (qtype) control for test set samples with an alternative:
  - BERT-based question type predictor using WKDQG data (question focus's entity type + position in triple, S or O);
  - Pick most probable alternative qtype and generate a new question with this.

Model — Metric	B-4	BSc	R-L	M
Test$_O$				
BART$_{rdf,qt}$	41.95	73.51	71.21	36.78
BART$_{rdf,qt,wkdqg}$	41.31	72.63	70.28	36.62
Test$_A$				
BART$_{rdf,qt}$	26.60	60.15	49.53	29.27
BART$_{rdf,qt,wkdqg}$	26.37	59.48	49.29	29.3

Choice — Measure	D	A	N	E
BART$_{rdf,qt}$ Test$_O$	14%	12%	18%	2%
BART$_{rdf,qt,wkdqg}$ Test$_A$	76%	4%	10%	4%
Same	8%	80%	62%	92%
No Majority Vote	2%	4%	10%	2%

Results: As expected, alternative questions score poorly on the automatic metrics, but they have comparable quality based on human judgements (3 annotators, Fleiss' kappa: 0.521)

5. Downstream QA performance

- Alternative test questions also lead to degraded performance of QA systems — Huang et al 2019 (KEQA), Mohammed et al 2018 (BuboQA) — due to distribution shift.
- However, enriching training data with set of questions of plausible qtype for the given triple leads to:
  - a reversal of the degradation on alternative test set;
  - robust performance maintained on original test set.

Split/Model	SQ_o	SQ_o+e	SQ_w0	SQ_w1	SQ_w2	SQ_w3
Train	O, (75,722)	O+E, (173,063)	O_w, (37,521)	O_w, (37,521)	O_w+E, (149,710)	O_w+E, (149,710)
Dev	O, (10,815)	O+E, (24,664)	O_w, (5,360)	O_w, (5,360)	O_w+E, (21,380)	O_w+E (21,380)
Test	O, (21,687)	O, (21,687)	O_w, (10,726)	O_w-A, (10,726)	O_w-A, (10,726)	O_w (10,726)
BuboQA Acc@1	74.63	74.03	**85.12**	81.42	85.08	84.57
KEQA Acc@1	75.30	74.76	**86.85**	81.15	83.79	86.44

Our code for the experiments: https://gitlab.inria.fr/hankelvin/wikidataqg