It’s common sense, isn’t it? Demystifying Human Evaluations in Commonsense-enhanced NLG systems

Miruna Clinciu*, Dimitra Gkatzia*, and Saad Mahamood*

Overview

- What is common sense?
- How do we instill commonsense knowledge to NLG systems?
- How is commonsense knowledge evaluated in current NLG systems?
- Introducing the *Commonsense Evaluation Card*

What is common sense?

“simple wisdom”
(Oxford English Dictionary)

“the ability to use good judgment in making decisions and to live in a reasonable and safe way”
(Cambridge dictionary)

“sound and prudent judgment based on a simple perception of the situation or facts”
(Mirriam Webster)

Commonsense knowledge forms in NLG

Rules

- Expert domain NLG systems such as BabyTalk
  (Portet et al., 2008)

Commonsense knowledge forms in NLG

Rules

- Expert domain NLG systems such as BabyTalk
  (Portet et al., 2008)

External knowledge

- ConceptNet
  (e.g. Speer et al., 2016)
- ATOMIC
  (Sap et al., 2019)
- COMET
  (Bosselut et al., 2019)

Commonsense knowledge forms in NLG

Rules

- Expert domain NLG systems such as BabyTalk
  (Portet et al., 2008)

External knowledge

- ConceptNet
  (e.g. Speer et al., 2016)
- ATOMIC
  (Sap et al., 2019)
- COMET
  (Bosselut et al., 2019)

PTLMs

- These approaches assume that common sense is present in pre-trained language models
  (e.g. Zhou et al., 2018)

Commonsense-enhanced NLG systems

- “a program has common sense if it automatically deduces for itself a sufficiently wide class of immediate consequences of anything it is told and what it already knows” *(McCarthy, 1959)*

- Commonsense can further refer to domain knowledge, stylistic attributes (sarcasm, humour, etc.) or reasoning among others

- This makes the evaluations of commonsense-enhanced NLG systems difficult, as the definition of commonsense is **context-dependent**.

Why is evaluation of commonsense knowledge difficult?

- Different definitions
- Commonsense in generated text
- Commonsense in external knowledge

How is commonsense knowledge evaluated in current NLG systems?

Methodology

Methodology

- We considered all papers published in ACL venues in the past three years (2018–2020)
- We screened the papers using the following search terms in their title: commonsense, generation, reasoning, domain knowledge, expert, expertise, sensible, ontology, knowledge
- We randomly chose 55 to annotate, and ended up with 34 papers

Methodology

Papers

- We considered all papers published in ACL venues in the past three years (2018–2020)
- We screened the papers using the following search terms in their title: commonsense, generation, reasoning, domain knowledge, expert, expertise, sensible, ontology, knowledge
- We randomly chose 55 to annotate, and ended up with 34 papers

Annotations

Annotation scheme from Howcroft et al. (2020) plus:

- Definition of commonsense knowledge
- Type of commonsense knowledge
- External knowledge: free text field.
- Was the knowledge evaluated in the generated text? (Yes/No)
- Criterion name for evaluation of external knowledge

Analysis

Methodology

Papers

- We considered all papers published in ACL venues in the past three years (2018–2020)
- We screened the papers using the following search terms in their title: commonsense, generation, reasoning, domain knowledge, expert, expertise, sensible, ontology, knowledge
- We randomly chose 55 to annotate, and ended up with 34 papers

Annotations

Annotation scheme from Howcroft et al. (2020) plus:

- Definition of commonsense knowledge
- Type of commonsense knowledge
- External knowledge: free text field.
- Was the knowledge evaluated in the generated text? (Yes/No)
- Criterion name for evaluation of external knowledge

Analysis

- At least 37 different evaluation criteria names were given in the 34 papers.
- Almost half of the papers did not contain a definition of common sense.
- External knowledge was evaluated less than half of the time.
- Most papers do not report evaluation details, which makes it harder to reproduce and evaluate the reported work.

Results

- At least 37 different evaluation criteria names were given in the 34 papers.
- Almost half of the papers did not contain a definition of common sense.
- External knowledge was evaluated less than half of the time.
- Most papers do not report evaluation details, which makes it harder to reproduce and evaluate the reported work.

Commonsense Evaluation Card

Commonsense Evaluation Card (CEC)

Commonsense Knowledge Definition: Basic definition of commonsense knowledge in the reported work.

- Definition
- Type of commonsense
- Example output of generated text that displays the intended commonsense capabilities.

External Knowledge: Basic information regarding the use of external knowledge and its evaluation

- Structured Knowledge
- Pre-trained Language Models
- Other
- Metrics for Evaluation of External Knowledge

Commonsense Knowledge in Generated Text: Evaluation Settings

- Automatic Metrics for Evaluation of commonsense knowledge in generated text
- Human Evaluation of commonsense knowledge in generated text

Conclusions

- This paper presented a human evaluation analysis on works describing systems that incorporate commonsense knowledge or other external knowledge

- As a solution for the large variability on how systems are evaluated we encourage the following:

    - Evaluate the reasoning ability of NLG systems (in addition to standard NLG metrics)

    - Provide definition(s) of commonsense knowledge to evaluators

    - Validate external knowledge bases to ensure that any errors present in generated output are not derived from the underlying knowledge.

    - Present commonsense knowledge errors in a more structured way

Дякую тобі

Thank you

EPSRC

Engineering and Physical Sciences Research Council

Scan QR Code for Resources