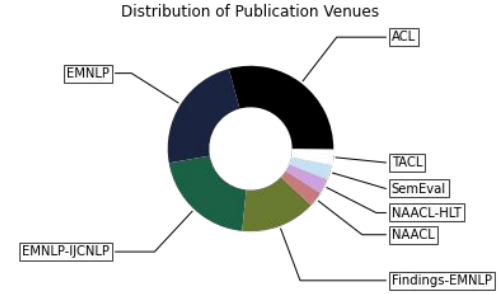
It's common sense, isn't it? Demystifying Human Evaluations in Commonsense-enhanced NLG systems

Miruna Clinciu[1*], Dimitra Gkatzia [2*] ✉, and Saad Mahamood [3*]
[1] Heriot-Watt University, Scotland, UK | [2] Edinburgh Napier University, Scotland, UK | [3] trivago N.V., Düsseldorf, Germany

Introduction

Enhancing **NLG systems with commonsense knowledge** has recently attracted a lot of attention. However, **evaluating such systems** remains challenging. In this work, we firstly look into how researchers have performed and reported human evaluations of such systems, and we then propose a **Commonsense Evaluation Card** (CEC) for reporting evaluations with commonsense-enhanced systems.



Commonsense Knowledge in NLG

Commonsense knowledge can be inserted in NLG systems through the following ways:

- External knowledge bases, such as knowledge graphs (e.g. ConceptNet), databases, documents.
- Pre-trained Language Models.
- Encoded in rules, e.g. in expert systems.

Data Collection & Annotations

We initially searched the ACL anthology for papers containing the terms "commonsense", "reasoning" etc. This left us with 129 papers, of which 55 were chosen for annotation. During annotation, we further discarded 21 papers that did not report an NLG system. For the paper annotation, we enhanced the annotation scheme provided by Howcroft et al. (2020) with the following categories:

- ❑ Definition of commonsense knowledge
- ❑ Type of commonsense knowledge
- ❑ Name of external knowledge used
- ❑ Was the knowledge evaluated in the generated text?
- ❑ Criterion name for evaluation of external knowledge

Analysis & Discussion

- 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 neither mentioned the type of common sense that their task was addressing.
- 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.
- To facilitate **reproducibility** as well as **understanding** of the reported work, we have devised the **Commonsense Evaluation Card (CEC) which offers authors a guide on how to report evaluations of commonsense-enhanced NLG systems**. It can also be used for NLP systems with minor modifications.

Commonsense Evaluation Card (CEC)

1. Commonsense Knowledge Definition: Basic definition of commonsense knowledge in the reported work.
– Definition
– Type of commonsense knowledge
– Example output of generated text that displays the intended commonsense capabilities

2. 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

3. 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

Scan QR Code for Resources