Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/coli_a_00561

LLM-based NLG Evaluation: Current Status and Challenges

Mingqi Gao*
Peking University
gaomingqi@pku.edu.cn

Xinyu Hu*
Peking University
huxinyu@pku.edu.cn

Xunjian Yin
Peking University
xjyin@pku.edu.cn

Jie Ruan
Peking University
ruanjie@stu.pku.edu.cn

Xiao Pu
Peking University
puxiao@stu.pku.edu.cn

Xiaojun Wan
Peking University
wanxiaojun@pku.edu.cn

Evaluating natural language generation (NLG) is a vital but challenging problem in natural language processing. Traditional evaluation metrics mainly capturing content (e.g. n-gram) overlap between system outputs and references are far from satisfactory, and large language models (LLMs) such as ChatGPT have demonstrated great potential in NLG evaluation in recent years. Various automatic evaluation methods based on LLMs have been proposed, including metrics derived from LLMs, prompting LLMs, fine-tuning LLMs, and human-LLM collaborative evaluation. In this survey, we first give a taxonomy of LLM-based NLG evaluation methods, and discuss their pros and cons, respectively. Lastly, we discuss several open problems in this area and point out future research directions.

Contents

1. Introduction

The evaluation of natural language generation (NLG) is an important but challenging issue. The lack of a single standard answer and the presence of multiple quality criteria make evaluating NLG more challenging than other NLP tasks. For example, in news summarization, a good summary should capture the key information from the source

[*] Equal contribution.