chatgpt相关汇报

王永胜

Human-like Summarization Evaluation with ChatGPT

- 任务：在5个数据集上使用4种评价方法探索ChatGPT做类人工摘要的性能评价（人工评价耗时大、费用高）
- 现状：自动评价标准ROUGE\BERTScor\BARTScore\ FactCC\FEQA等性能表现、可用性等还远不及预期
- 设计4种prompt

Human-like Summarization Evaluation with ChatGPT

Evaluate the quality of summaries written for a news article. Rate each summary on four dimensions: {Dimension_1}, {Dimension_2}, {Dimension_3}, and {Dimension_4}. You should rate on a scale from 1 (worst) to 5 (best).

Article: {Article}
Summary: {Summary}

Figure 1: The template for Likert scale scoring.

Given a new article, which summary is better? Answer "Summary 0" or "Summary 1". You do not need to explain the reason.

Article: {Article}
Summary 0: {Summary_0}
Summary 1: {Summary_1}

Figure 2: The template for pairwise comparison.

You are given a summary and some semantic content units. For each semantic unit, mark "Yes" if it can be inferred from the summary, otherwise mark "No".

Summary: {Summary}
Semantic content units:
1. {SCU_1}
2. {SCU_2}
.....
n. {SCU_n}

Figure 3: The template for Pyramid.

Is the sentence supported by the article? Answer "Yes" or "No".

Article: {Article}
Sentence: {Sentence}

Figure 4: The template for binary factuality evaluation.

Human-like Summarization Evaluation with ChatGPT

Metric Name	consistency			relevance			fluency			coherence		
	sample	system	dataset	sample	system	dataset	sample	system	dataset	sample	system	dataset
ROUGE-1	0.153	0.744	0.137	0.326	0.744	0.302	0.113	0.730	0.080	0.167	0.506	0.184
ROUGE-2	0.179	0.779	0.129	0.290	0.621	0.245	0.156	0.690	0.062	0.184	0.335	0.145
ROUGE-L	0.111	0.112	0.109	0.311	0.362	0.284	0.103	0.306	0.079	0.128	0.138	0.141
BERTScore	0.105	-0.077	0.118	0.312	0.324	0.362	0.189	0.246	0.150	0.284	0.477	0.317
MoverScore	0.151	0.679	0.150	0.318	0.724	0.294	0.126	0.687	0.119	0.159	0.474	0.178
BARTScore_s_h	0.299	0.800	0.269	0.264	0.524	0.363	0.243	0.614	0.187	0.322	0.477	0.335
BARTScore_h_r	0.097	0.606	0.101	0.178	0.147	0.246	0.002	0.261	0.000	0.017	-0.115	0.064
BARTScore_r_h	-0.075	-0.556	-0.090	-0.081	-0.112	-0.136	0.013	-0.212	0.019	0.044	0.165	-0.010
BARTScore_cnn_s_h	0.367	0.435	0.334	0.356	0.765	0.394	0.349	0.746	0.285	0.448	0.700	0.408
BARTScore_cnn_h_r	0.171	0.771	0.106	0.320	0.456	0.244	0.111	0.561	0.066	0.153	0.174	0.130
BARTScore_cnn_r_h	0.001	-0.079	-0.004	0.146	0.312	0.221	0.107	0.297	0.145	0.228	0.506	0.236
ChatGPT	**0.435**	**0.833**	**0.425**	**0.433**	**0.901**	**0.445**	**0.419**	**0.889**	**0.410**	**0.561**	**0.832**	**0.557**

Table 1: Spearman's ρ of sample level, system level, and dataset level on SummEval.

Metric Name	coherence			fluency			informativeness			relevance		
	sample	system	dataset	sample	system	dataset	sample	system	dataset	sample	system	dataset
ROUGE-1	0.095	0.429	0.100	0.104	0.429	0.064	0.130	0.286	0.149	0.147	0.357	0.122
ROUGE-2	0.025	0.321	0.080	0.047	0.321	0.045	0.078	0.250	0.158	0.090	0.357	0.124
ROUGE-L	0.064	0.357	0.079	0.072	0.357	0.045	0.089	0.214	0.137	0.106	0.321	0.101
BERTScore	0.148	0.429	0.169	0.170	0.429	0.154	0.131	0.286	0.196	0.163	0.357	0.176
MoverScore	0.162	0.429	0.173	0.120	0.429	0.112	0.188	0.286	0.232	0.195	0.357	0.192
BARTScore_s_h	**0.679**	**0.964**	**0.656**	**0.670**	**0.964**	**0.615**	**0.646**	**0.821**	**0.645**	**0.604**	**0.893**	**0.588**
BARTScore_h_r	0.329	0.286	0.302	0.292	0.286	0.261	0.419	0.429	0.386	0.363	0.357	0.386
BARTScore_r_h	-0.311	-0.571	-0.249	-0.215	-0.571	-0.232	-0.423	-0.750	-0.346	-0.334	-0.607	-0.305
BARTScore_cnn_s_h	0.653	0.893	0.623	0.640	0.893	0.596	0.616	0.750	0.592	0.567	0.786	0.557
BARTScore_cnn_h_r	0.239	0.429	0.215	0.235	0.429	0.165	0.284	0.429	0.239	0.267	0.464	0.221
BARTScore_cnn_r_h	0.316	0.429	0.333	0.353	0.429	0.330	0.242	0.286	0.289	0.245	0.357	0.292
ChatGPT	0.484	0.821	0.476	0.480	0.607	0.471	0.521	0.607	0.508	0.524	0.714	0.521

Table 2: Spearman's ρ of sample level, system level, and dataset level on Newsroom.

Human-like Summarization Evaluation with ChatGPT

Metric Name	Accuracy
ROUGE-1	0.5869
ROUGE-2_f	0.4997
ROUGE-L_f	0.5647
BARTScore	0.5674
MoverScore	0.5864
BARTScore_s_h	0.5858
BARTScore_h_r	0.6151
BARTScore_r_h	0.5317
BARTScore_cnn_s_h	0.5880
BARTScore_cnn_h_r	0.5934
BARTScore_cnn_r_h	0.5089
ChatGPT	**0.6178**

Table 3: Accuracy of pairwise comparison on TLDR.

Metric Name	Accuracy
DAE	0.6304
FactCC	0.5362
ChatGPT	**0.6436**

Table 4: Accuracy of the binary determination of SCUs on REALSumm.

	QAGS_CNN	QAGS_XSUM
DAE	0.8459	0.6360
FactCC	0.7731	0.4937
ChatGPT	**0.8488**	**0.7573**

Table 5: Accuracy of binary factuality evaluation on QAGS.

Human-like Summarization Evaluation with ChatGPT

- 不同prompt差异较大

	consistency			relevance			fluency			coherence		
	sample	system	dataset	sample	system	dataset	sample	system	dataset	sample	system	dataset
ChatGPT	0.435	0.833	0.425	0.433	0.901	0.445	0.419	0.889	0.410	0.561	0.832	0.557
ChatGPT+def	0.471	0.786	0.479	0.453	0.877	0.479	0.347	0.606	0.341	0.568	0.802	0.570
ChatGPT+def+ins	0.338	-0.149	0.302	0.396	-0.079	0.433	0.349	0.016	0.325	0.501	0.338	0.494
ChatGPT+sys_prompt	0.414	0.007	0.376	0.334	0.268	0.365	0.390	0.149	0.362	0.473	0.552	0.470
Annotator_0	**0.843**	**0.990**	**0.902**	0.748	**0.968**	0.816	0.740	**0.960**	0.775	0.845	0.929	0.884
Annotator_1	0.813	0.965	0.881	**0.767**	0.953	**0.823**	**0.847**	0.843	**0.876**	**0.889**	**0.982**	**0.913**
Annotator_2	0.712	0.973	0.797	0.743	0.944	0.747	0.613	0.923	0.700	0.790	0.932	0.820

Table 6: Spearman's ρ of sample level, system level, and dataset level on SummEval. Annotator_0, Annotator_1, Annotator_2 refer to the three expert annotators. We compute the correlation coefficient between the score given by a particular annotator and the average score of the three. "+def" means adding dimension definitions in the prompt. "+ins" means adding step instructions in the prompt. Please see the example in Figure 5 for dimension definitions and step instructions. "+sys_prompt" denotes setting system prompt.

Human-like Summarization Evaluation with ChatGPT

- 评价chatgpt自动生成评价解释的质量
- --chatgpt可以自圆其说，但不一定准确
- chatgpt产生无效回应（1%），包括：拒绝评价、无关评价、生成新的摘要、继续生成摘要，生成无效的原因需进一步探索

Linguistic ambiguity analysis in ChatGPT

- NLP里常见三种歧义：缺少上下文，词汇（一词多义）、句法（分词等）、语义（指代）
- 1、关于谐音词（相同写法和读法但意义不同）结论： prompt很重要，过度检测歧义，并指出上下文是消歧的关键因素，当没有足够信息来源生成可靠响应时，会分配一个默认输出。
- 2、关于一词多义结论：ChatGPT 非常擅长对无歧义句子中的多义词的含义进行分类，即使定义不在上下文中也是如此。 在歧义句中ChatGPT 的表现不如同谐音词中的表现，但随着更好的提示，结果会有所改善。
- 3、关于句法歧义结论：尽管ChatGPT很难检测到句法歧义，但是可以通过多轮对话指出模型学习的关键因素，以便更好的prompt
- 4、指代歧义结论：有明显性别提示时，ChatGPT可成功识别，但是当句子无歧义时（从语法角度），ChatGPT也会将性别偏见置于语法之上

Linguistic ambiguity analysis in ChatGPT

- 性能： accuracy of 0.6061 and an F1 of 0.48

Ambiguity type	True Positive	True Negative	False Positive	False Negative
Homonymy	3	2	7	1
Polysemy	0	10	0	2
Syntactic	2	0	0	3
Semantic	1	2	0	0

Table 1: Ambiguity detection results