POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Evaluation of Metrics for Neural-network-based Summarization

Laurea Magistrale in Computer Science Engineering - Ingegneria Informatica

Author: Yasmin Awad

Advisor: Prof. Maristella Matera

Co-advisor: Emanuele Pucci, Marco Fisichella

Academic year: 2022-2023

1. Introduction

The increasing use of visual design in web interfaces presents challenges for individuals with visual impairments. With over 2.2 billion people globally affected by vision impairment, ensuring web accessibility is crucial. Screen readers are commonly used by visually impaired individuals, but they face difficulties when web developers don't adhere to accessibility standards. To address this, Politecnico di Milano developed ConWeb, a conversational web browser, to facilitate browsing without visual exploration. During testing, users emphasized the importance of summaries for smoother browsing. However, concerns about summary quality led to research on selecting an appropriate summarization model for ConWeb. Taking into account Italian Web articles, this research aims to choose a Large Language Model capable of generating high-quality summaries for ConWeb. Additionally, we assessed a range of automatic metrics commonly used for evaluating summaries, highlighting their limitations when applied to advanced language models like GPT [4]. The latter has been achieved through a comparison with human evaluations, considered a golden standard.

1.1. Contributions

The contributions of this thesis include:

- Utilizing automatic metrics to evaluate summaries of Italian web articles.
- Developing a human evaluation form for assessing summaries.
- Comparing the performance of various Large Language Models against human-generated reference summaries, with both an extractive and abstractive style.
- Providing evidence of the inadequacy of automatic metrics for the evaluation of Large Language Models like GPT [4].
- Validating the quality and performance of four Large Language Models for summarizing Italian Web articles, a novel contribution to the existing literature.

2. Literature Review

Text summarization is a critical task in natural language processing, aiming to condense large amounts of text into concise summaries. Two main approaches, extractive and abstractive summarization, are used, with the latter being more challenging as it involves understanding and rephrasing the source text. Transformer models have shown exceptional performance in abstractive summarization tasks, leveraging

self-attention mechanisms to capture global dependencies in the source text. Over recent years, Large Language Models (LLMs) like GPT [4] have achieved state-of-the-art results in various natural language processing tasks.

However, evaluating the quality of summaries poses a challenge due to the complexity of language. Traditional automatic evaluation metrics like ROUGE [3] focus on lexical matching and may favor extractive approaches, failing to capture the semantic meaning and coherence of abstractive summaries. Recent studies [2, 5] have shown that these metrics are inadequate for evaluating summaries generated by LLMs, leading to a reliance on human evaluation, which is slow and costly. Human evaluation remains the golden standard for assessing the quality of summaries despite its limitations.

3. Set Up

To proceed with our experiments, we have taken into account 4 **models**: mbart-summarization-mlsum, mbart-summarization-ilpost, BART, GPT-3.5. From now on called: *mlsum*, *ilpost*, *bart*, *gpt*. While the first three models were used in previous research at Politecnico di Milano, gpt's inclusion stems from its advanced NLP features and integration within ConWeb.

The **dataset** is made of 76 articles taken from Italian websites, divided into 4 topics: Culture, Politics, Science, and Sport. For each article, one reference summary is present, and 4 summaries have been generated utilizing the before-said models.

To enhance the dataset's quality, we've supplemented the existing human-generated references (referred to as "old references") with new references. The new references aim to better align with the semantic depth of gpt, which differs from earlier models mlsum, ilpost, and bart. This refinement ensures a more comprehensive evaluation across all models, reflecting variations in summarization styles.

3.1. Generated Summaries Analysis

In the initial examination of the dataset, distinct characteristics and common errors were observed in the summaries generated by different models. Summaries produced by mlsum tended to be excessively brief, often reduced to a single sentence, while those from ilpost and bart were also concise but typically longer than mlsum's. However, bart summaries exhibited errors such as missing accents, escaped apostrophes, and occasional substitution of Italian words with English counterparts due to the model's lack of fine-tuning on Italian data.

Notably, mlsum, ilpost, and bart summaries frequently contained semantic inaccuracies and grammatical mistakes, rendering some of them nonsensical or incomprehensible. Specifically, mlsum generated 20 such summaries out of 76, while ilpost and bart produced 16 and 6 unusable summaries, respectively. On the other hand, gpt summaries were notably longer than others but devoid of grammatical errors. Although gpt summaries occasionally incorporated information from previously summarized articles on similar topics, this inconsistency occurred infrequently (only in 4 out of 76 generations) and did not significantly impact the overall coherence and relevance.

general features

model	**nonsensical**	**one-sentence**
mlsum	20/76	74/76
ilpost	16/76	16/76
bart	6/76	6/76
gpt	4/76	0/76

Table 1: Some genereated summaries' features.

Despite this minor inconsistency, gpt summaries demonstrated superior fluency, coherence, and relevance compared to other models. This analysis led to the hypothesis that gpt generated summaries are the most suitable for the intended task.

old reference

ROUGE-2	Culture	Politics	Science	Sport
mlsum	0.2019	0.2036	0.3463	0.2940
ilpost	0.2144	0.3406	0.2549	0.2953
bart	0.4081	0.4348	0.4942	0.3699
gpt	0.2144	0.2047	0.2323	0.1566

new reference

ROUGE-2	Culture	Politics	Science	Sport
mlsum	0.1595	0.1549	0.2315	0.2135
ilpost	0.1632	0.2125	0.1950	0.2217
bart	0.3340	0.2697	0.3230	0.3342
gpt	0.2334	0.2734	0.28434	0.1847

(a) ROUGE-2 scores.

old reference

BERTScore	Culture	Politics	Science	Sport
mlsum	0.7212	0.7257	0.7873	0.7493
ilpost	0.7341	0.7810	0.7581	0.7648
bart	0.7972	0.7996	0.8296	0.7713
gpt	0.7532	0.7511	0.7770	0.7389

new reference

BERTScore	Culture	Politics	Science	Sport
mlsum	0.7116	0.7168	0.7484	0.7175
ilpost	0.7242	0.7468	0.7416	0.7368
bart	0.7800	0.7657	0.7857	0.7680
gpt	0.7746	0.7834	0.8079	0.7651

(b) BERTScore scores.

old reference

chrF	Culture	Politics	Science	Sport
mlsum	0.2365	0.2256	0.3406	0.3130
ilpost	0.2663	0.3663	0.3296	0.3557
bart	0.4542	0.4962	0.5576	0.4659
gpt	0.4541	0.4843	0.5093	0.4259

new reference

chrF	Culture	Politics	Science	Sport
mlsum	0.2110	0.1923	0.2451	0.2371
ilpost	0.2400	0.2744	0.2641	0.2816
bart	0.4022	0.3814	0.4059	0.4074
gpt	0.4849	0.5331	0.5426	0.4475

(c) chrF scores.

Figure 1: Some automatic metrics' averages over old and new references applied to each model's generated summaries, divided by set of articles.

4. Evaluation of the Summaries

We have conducted experiments to select the optimal Large Language Model (LLM) for generating summaries of Italian Web articles within ConWeb. We hypothesize that existing automatic metrics may not adequately evaluate LLM-generated summaries, and that among the considered LLMs gpt performs the best.

4.1. Automatic Evaluation

The first evaluation step has been done through the use of automatic evaluation. To evaluate the performance of summarization models, we've selected eight traditional automatic metrics: ROUGE-1, ROUGE-2, ROUGE-L, BERTScore, BLEU, BLANC, chrF, and METEOR. Some results are reported in Figure 1.

The results of applying automatic metrics to the generated summaries reveal consistent competition between the gpt and bart models. With the old reference set, bart often outperforms gpt due to the latter's high abstractiveness conflicting with the extractive nature of the old references. However, with the new abstractive references, the performance gap narrows, with both models exhibiting similar scores across various metrics. Notably, gpt sometimes struggles with sports articles, possibly due to its substitution of fan-known references with more generic terms. Metrics like chrF and METEOR consistently favor a specific model (gpt and bart, respectively) across all articles.

4.2. Human Evaluation

To assess the goodness of the generated summaries a human evaluation has been conducted. To conduct the human evaluation a crowdsourcing website, Prolific, has been used. Funding for the evaluation was provided by Doshisha University.

A form has been created for each article (76). The form has been realized following previous research [2] and with the use of the following definitions of summary's qualities:

- **Intrinsic**: the quality of the summary is measured based on the summary itself without considering the original document. The intrinsic quality is mea-

sured through *overall quality, grammaticality, non-redundancy, referential clarity, focus,* and *structure&coherence.*

- **Extrinsic**: measures the quality of the summary through the completion of some task based on the original document. The extrinsic quality is measured through *summary usefulness, source usefulness,* and *summary informativeness.*

4.2.1 Construction of the questionnaires

Each questionnaire began with an explanation of the hypothesis and intentions behind the evaluation. Crowd workers were then asked to judge intrinsic qualities using a 5-point MOS scale for each generated summary, the reference summary, and a custom-made summary. The inclusion of the poorly constructed custom summary, with instructions to rate it poorly, served to identify inattentive respondents. Notably, only the "new references" were considered during the evaluation. Following the assessment of intrinsic qualities, crowd workers evaluated extrinsic ones, including the informativeness of each summary and their comprehension of the original article, generated summaries, and reference. Comprehension questions tailored to each article were included to assess source and summary usefulness.

4.2.2 Conducting the Evaluation

The evaluation process involved obtaining five assessments for each article form from crowd workers on Prolific. To prevent biases, each crowd worker could assess a maximum of four forms. Additionally, trusted individuals were enlisted to evaluate the summaries using the same setup of Prolific, with the same restriction on the number of forms they could assess. This approach ensured a minimum of six evaluations for each article. In total, 543 responses were collected, of which 485 were deemed attentive and accepted. The evaluation involved 353 participants, including both crowd workers and acquaintances.

4.2.3 Human Evaluation Results

The human evaluation results, presented in Figures 3 and 2, illustrate the performance of different models across various quality metrics and comprehension questions.

QUESTIONS	CULTURE	POLITICS	SCIENCE	SPORT
mlsum	47/135	53/129	44/133	53/146
ilpost	25/135	49/129	47/133	38/146
bart	33/135	46/129	30/133	75/146
gpt	129/135	113/129	131/133	139/146
reference	125/135	124/129	129/133	141/146
original	130/135	124/129	132/133	143/146

Figure 2: Human evaluation's questions results.

In Figure 3, gpt, compared to other models, consistently achieves the highest scores across all quality metrics and article sets, even surpassing the reference summaries in most cases. Notably, gpt excels in informativeness, with an average score exceeding 4, while other models fall short of reaching a score of 3. Figure 2 demonstrates that gpt's summaries yield a high number of correct answers to comprehension questions, indicating its ability to convey crucial information from the original articles effectively. Despite some incorrect answers from crowd workers, gpt's performance closely aligns with the original text, highlighting its clarity and suitability for the considered task.

CULTURE	overall	grammar	non-red.	clarity	focus	structure	inform.
mlsum	3.0388	3.4361	4.0213	3.2055	3.6103	3.2143	2.5978
ilpost	2.6466	3.1704	3.5476	2.9586	3.0238	2.7807	2.43995
bart	3.3233	3.2381	3.7932	3.5376	3.6666	3.3446	2.948
gpt	4.3032	4.3809	4.2619	4.2895	4.1930	4.3133	4.1581
reference	4.1040	4.0276	4.0752	4.1241	3.9348	4.1103	3.9578
POLITICS	overall	grammar	non-red.	clarity	focus	structure	inform.
mlsum	2.9449	3.3621	3.9474	3.3559	3.6291	3.3170	2.7998
ilpost	2.8822	3.2306	3.6804	3.3308	3.3496	3.1366	2.7415
bart	2.9210	3.0301	3.8258	3.1153	3.3609	2.9975	2.6729
gpt	4.1767	4.2644	3.9586	4.1428	4.0050	4.1366	4.1842
reference	3.7958	3.9261	3.9248	3.8947	3.7619	3.8346	3.9408
SCIENCE	overall	grammar	non-red.	clarity	focus	structure	inform.
mlsum	3.0165	3.4044	3.8251	3.1491	3.6898	3.2493	2.6434
ilpost	2.5602	3.1951	3.4408	2.9461	3.2827	2.8022	2.5160
bart	3.1766	3.1395	3.8019	3.3763	3.5992	3.2681	2.7943
gpt	4.1981	4.2556	4.0502	4.1388	4.0063	4.1121	4.1777
reference	3.7364	3.8796	3.6468	3.9044	3.6753	3.8220	3.8858
SPORT	overall	grammar	non-red.	clarity	focus	structure	inform.
mlsum	3.1136	3.4674	4.0435	3.3488	3.7337	3.4298	2.8251
ilpost	2.3754	2.7895v	3.5022	2.7433	3.0857	2.5849	2.5347
bart	3.0456	3.1310	3.7289	3.3155	3.3738	3.0917	2.9696
gpt	4.0334	4.1610	3.9430	4.0577	3.9927	4.0222	4.0240
reference	3.6591	3.6386	3.7877	3.6312	3.6008	3.6053	4.0984

Figure 3: Quality averages' scores from Human Evaluation.

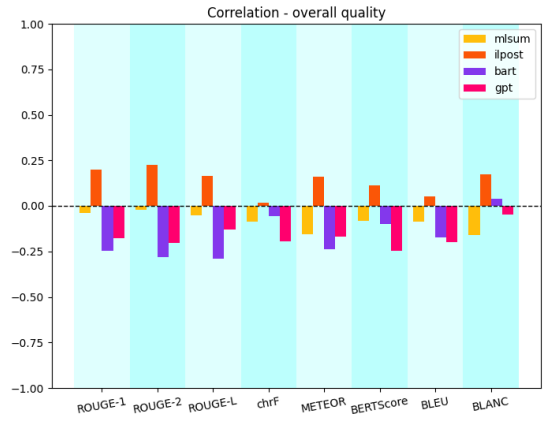
5. Correlation Analysis

From the evaluations carried out, it's worth noting the discrepancy between the automatic evaluation, which did not select a specific model as the best one, and human evaluation, which highly preferred the performances of gpt. This highlights the limitations of traditional evaluation metrics in assessing the effectiveness of LLM-generated summaries accurately. We decided to conduct further analysis to better understand these differences.

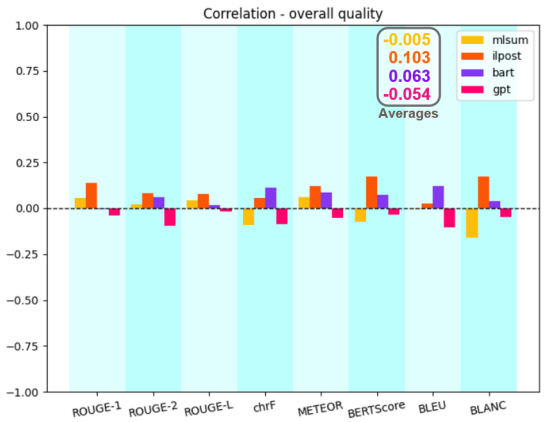
To the results from automatic metrics and human evaluation, we have applied **Spearman correlation**, a statistical method used to identify potential relationships between datasets. Spearman correlation coefficients range from -1 to +1, with zero indicating no correlation, positive values indicating a positive correlation, and negative values indicating a negative correlation. Correlations of -1 or +1 imply an exact monotonic relationship. A high correlation would indicate alignment between metrics and human judgments, while a weaker correlation would prompt reconsideration of the chosen metrics or highlight potential limitations in capturing certain aspects of summary quality.

The correlation has been calculated for each quality (intrinsic and extrinsic) with each applied metric, considering both old and new reference sets. In Figure 4 is shown the correlation between the overall quality and the automatic metrics. The analysis reveals that correlation scores in general do not exceed 0.25, and some are even negative. Previous research anticipated a human correlation of at least 0.30 for quality summaries, but notably, summaries favored by human evaluation, particularly those generated by GPT, did not receive significantly higher scores in automatic metrics. Despite expectations based on previous research, the correlation between human evaluation and automatic metrics is low, suggesting that automatic metrics do not accurately reflect the summary quality. This finding supports the thesis that traditional automatic metrics are unsuitable for summarization evaluation.

As for the overall quality, which is the quality that should enclose all the other ones, summaries from ilpost exhibit the highest correlation with automatic metrics' scores obtained with the old references, while more abstractive models like BART and GPT display negative behavior. The extractiveness of the old reference summaries tends to clash with human judgments on this end. However, with the new reference sets, BART and ilpost show higher correlations, while GPT exhibits the lowest correlation. Overall, the correlation patterns within the new references indicate a lower range of values, with previously high correlations decreasing and negative correlations increasing.



(a) Taking into account the **old reference**.



(b) Taking into account the **new reference**.

Figure 4: Correlation between the overall quality and the metrics.

6. Conclusions

The analysis carried out showed clear results. Our findings can be summarized as follows:

- Evaluation with traditional metrics presents an uncertain picture regarding

which model generates superior summaries for Italian web articles.
- Human evaluation unequivocally favors GPT-3.5, consistently awarding it higher scores compared to other models.
- Correlation analysis reveals weak correlations with automatic metrics, particularly concerning GPT-3.5 summaries, which align closely with user preferences according to human evaluation.

This results highlights a significant discrepancy between human and automatic evaluations, underscoring the inadequacy of traditional automatic metrics in summarization evaluation, especially for our task which involved Italian articles. Human judgment indicates that GPT-3.5 output summaries are more comprehensible, easier to follow, and more accurate. GPT-3.5 emerges as the most suitable model for ConWeb among those considered.

6.1. Lesson Learned

Based on the experiments conducted and the resulting outcomes, certain criteria emerge which a metric should adhere to in order to be deemed suitable for summary evaluation:

- It's crucial for metrics to consider semantics beyond mere textual overlap, acknowledging synonyms, antonyms, inflectional variants, and paraphrases, as language allows for expressing the same idea in diverse ways.
- Regular updates to metrics are essential to adapt to evolving language variations and idioms.
- Enhanced evaluation accuracy can be achieved by incorporating multiple reference summaries or directly referencing the original article.

Advancements in generative Large Language Models (LLMs) offer a promising alternative. Utilizing LLMs for evaluating other LLMs bypasses the need for reference summaries, streamlining the evaluation process and reducing costs and time expenditure, especially in mitigating issues like crowd worker inattention experienced during human evaluations.

As this research concludes, new approaches to summary evaluation are emerging, particularly to address the challenges of assessing abstractive summaries. The new context-based approaches [1] eliminate the need for reference summaries and evaluate generated texts based on contextual understanding.

6.2. Future Works

Future studies could apply these new context-based evaluation approaches to the previously assessed Italian articles. It would be valuable to determine if the high correlation seen in other languages can be replicated in Italian. Moreover, these methods allow for faster fine-tuning of large language models without the need for human reference generation. By scraping Italian web pages, researchers can quickly gather articles for model training. The resulting summaries can then be evaluated alongside scores from context-based metrics, potentially improving summarization quality evaluation in Italian.

References

[1] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. Gptscore: Evaluate as you desire, 2023.

[2] Neslihan Iskender, Tim Polzehl, and Sebastian Möller. Reliability of human evaluation for text summarization: Lessons learned and challenges ahead. In Anya Belz, Shubham Agarwal, Yvette Graham, Ehud Reiter, and Anastasia Shimorina, editors, *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 86–96, Online, April 2021. Association for Computational Linguistics.

[3] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[4] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2018.

[5] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. Benchmarking large language models for news summarization, 2023.