Atténuer l'impact de la qualité des références sur l'évaluation des systèmes de résumé grâce aux métriques sans référence

Théo Gigant[1, 2] Camille Guinaudeau[3] Marc Decombas[2] Frederic Dufaux[1]
(1) Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes, Gif-Sur-Yvette, France
(2) JustAI, Evreux, France
(3) Université Paris-Saclay, Japanese French Laboratory for Informatics, CNRS, Tokyo, Japon
theo.gigant@l2s.centralesupelec.fr

RÉSUMÉ

Les métriques d'évaluation sont utilisées comme des indicateurs pour évaluer les systèmes de résumé abstractif lorsque les annotations sont trop coûteuses. Pour être utiles, ces métriques doivent permettre une évaluation fine, présenter une forte corrélation avec les annotations humaines, et idéalement ne pas dépendre de la qualité des références. Cependant la plupart des métriques d'évaluation standard pour le résumé sont basées sur des références, et les métriques sans références sont faiblement corrélées à la pertinence des résumés, en particulier pour des documents longs. Dans cet article, nous introduisons une métrique sans référence qui corrèle bien avec la pertinence telle qu'évaluée par des humains, tout en étant très peu coûteuse à calculer. Nous montrons également que cette métrique peut être utilisée en complément de métriques basées sur des références afin d'améliorer leur robustesse dans des situations où la qualité des références est faible.

ABSTRACT

Mitigating the Impact of Reference Quality on Evaluation of Summarization Systems with Reference-Free Metrics

Automatic metrics are used as proxies to evaluate abstractive summarization systems when human annotations are too expensive. To be useful, these metrics should be fine-grained, show a high correlation with human annotations, and ideally be independent of reference quality ; however, most standard evaluation metrics for summarization are reference-based, and existing reference-free metrics correlate poorly with relevance, especially on summaries of longer documents. In this paper, we introduce a reference-free metric that correlates well with human evaluated relevance, while being very cheap to compute. We show that this metric can also be used alongside reference-based metrics to improve their robustness in low quality reference settings.

MOTS-CLÉS : évaluation, résumé abstractif, métrique.

KEYWORDS: evaluation, abstractive summarization, metric.

ARTICLE : **Accepté à** Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, https://aclanthology.org/2024.emnlp-main.1078/.