Evaluation of the Machine Translation of Scientific Documents

François Yvon - Rachel Bawden

April 2023

Subject

Evaluation in Machine Translation MT evaluation is a crucial component of model development and remains a challenging subject (Kocmi et al., 2021 ; Freitag et al., 2022). Human evaluation remains the best judge of quality and research in manual evaluation has evolved over time, particularly with the progress made in MT and the availability of crowdworking platforms (Graham et al., 2013 ; Freitag et al., 2021). The development of automatic metrics, which seek to replicate human judgments of translation quality, is a major area of study. Many metrics exist, from simple ones that rely on lexical overlap such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005) and ChrF (Popović, 2015) to those relying on more modern techniques (e.g. pre-trained neural language models) such as BERTscore (Zhang et al., 2020), BARTScore (Yuan et al., 2021), BLEURT (Sellam et al., 2020) and COMET (Rei et al., 2020). As the quality of MT systems improves, it is becoming increasingly important to have accurate metrics. In response to the controversial debate concerning MT reaching human parity (Wu et al., 2016 ; Hassan et al., 2018), analysis has shown that there is still significant margin for improvement in MT if the human evaluation setup is improved (Toral et al., 2018 ; Laübli et al., 2018), for example by using expert annotators and taking into account document context, and therefore automatic metrics must succeed in distinguishing between high quality systems. Recent developments in automatic evaluation have shown that, like in human evaluation (Freitag et al., 2021), fine-grained error analysis can also be helpful (Lu et al., 2022, 2023). Recent progress in large generative language models (e.g. GPT (Brown et al., 2020) and PaLM (Chowdhery et al., 2022) can also provide new possibilities, for example in the automatic generation of metric training data (Mohtashami et al., 2023) and as evaluation metrics themselves (Kocmi and Federmann, 2023 ; Lu et al., 2023).

Evaluation of Scientific Documents

The evaluation of the MT of scientific documents poses specific challenges beyond the general problems faced, one of them being the heavy use of domain-specific terms, which, if translated incorrectly, severely impact the quality of the translation. Evaluation metrics should therefore also be sensitive to specific challenges faced by the evaluation of scientific documents : (i) the correct translation of terms, (ii) the coherent translation of terms within a document (with respect to term variants, use of acronyms, etc.) and (iii) the capacity to maintain a logical argument between sentences and sections. Previous work has suggested providing complementary measures to evaluate these specific aspects, for example correct term translation (Alam et al., 2021) and lexical cohesion (Wong and Kit, 2012 ; Gong et al., 2015). These notably require developing evaluation metrics that take into account document-level context, rather than evaluating sentences in isolation (Jiang et al., 2022 ; Vernikos et al., 2022).

Directions to be Explored

This PhD will explore the question of evaluation of machine translation for scientific documents, both in terms of manual and automatic methods. This will require :

— gaining a solid grasp of the current methods available, and how well they perform on state-of-the-art machine translation systems, including on large language models such as GPT (Brown et al., 2020).
— identifying and quantifying the difficulties faced by evaluation metrics, particularly for scientific documents (e.g. terminological issues, abbreviations, references, etc.)
— developing new methods of automatic evaluation to handle the particularities of these kinds of text, integrating document context and other aspects that are important for academic texts (coherence and readability).
— evaluating the methods developed against human judgments of MTquality, being sure that the methods of human judgment collection are appropriate and of the highest possible standard.

There are several possible directions for new methods of automatic evaluation, although those explored in practice will of course follow developments in the field :

— Question-based metrics : Inspired by the use of question-based metrics to evaluate text generation tasks (Scialom et al., 2021) [1] this direction looks at how terminologies, relation extraction and information extraction can be used as a means of evaluation of translation quality. For example, (i) can the same relations be found between a reference (human-produced)

1. Question-based metrics involve using automatically generated questions and evaluating an output on the ability to answer the question given the output.

translation and an automatically produced one ? (ii) can terms be matched in similar parts of the document ? (iii) how coherent is the use of terms within a document ? and (iv) can the same information be extracted from an MT output and the source or reference text ?

— Generation of synthetic data : In line with recent work on the use of large language models to generate additional training data (Mohtashami et al., 2023), this direction will look at how to exploit existing NLP models to benefit evaluation metrics, particularly when it comes to error analysis. The challenges faced will be covering all ranges of phenomena and not just the most frequent, handling the confidence or uncertainty of the model in its predictions and being able to maintain a good performance in non-English languages.
— Generation of error analyses and explanations of problems/errors : Along the lines of work in both human evaluation (Freitag et al., 2021) and automatic metrics (Lu et al., 2023), it will be important to envisage new paradigms of evaluation, involving both finer grained categories of errors and potentially also the generation of explanations that can provide users with information about where errors may lie, for example to aid post-edition of of machine translated outputs.

Context

This PhD will be financed by the ANR project MaTOS Machine Translation for Open Science which aims to develop new methods of automatically translating and evaluating scientific documents. The project focuses on translation between English and French, for which resources are readily available and translations are of a reasonable quality and coherence. The PhD will be co-supervised by Rachel Bawden (Inria, ALMAnaCH project-team) and François Yvon (CNRS).

Main activities

The main activities of the PhD will include :

— keeping up-to-date with related work on the topic, and producing a report on the state of research in the field in the context of the ANR project
— carrying out research on the topic outlined above, both in the development of new ideas, positioning with respect to related work and validation of the methodology via experiments and analysis
— the presentation of work both internally to colleagues and externally in the form of conference/journal/workshop papers and in the final PhD thesis
— interacting and exchanging with colleagues on NLP topics

Expertise required and qualities sought

The position is a 3-year funded PhD position of starting on the 1st September 2023 at the earliest. Candidates should have a Master 2 or equivalent (e.g. engineering school) in computer science (speciality artificial intelligence, machine learning or natural language processing). They should have a good level in programming (python), experience with neural networks and an interest in natural language processing. A good written and spoken level of English is required, and knowledge of French is preferred.

We are looking for highly motivated candidates with a strong background in NLP, machine learning and an interest in linguistics and language. Ideally, candidates should be able to show initiative, creativity and have a good eye for analysis of data and results. In your application (which can be in English or in French), please include :

- CV
- Letter of motivation
- Letters of recommendation
- Optionally an example of your previous written work (if possible related to NLP), for example a master's thesis, research paper, etc.

General information and how to apply

Theme/Domaine : Langue, parole et audio Town/city : Paris Inria centre : Centre Inria de Paris Estimated starting date : 2023-09-01 Duration of contract : 3 years Deadline to apply : 2023-05-22

Please apply via the Inria recrutement website : https://recrutement.inria.fr/public/classic/en/offres/2023-06180

Benefits package

- Subsidised meals
- Partial reimbursement of public transport costs
- Leave : 7 weeks of annual leave + 10 extra days off due to RTT (statutory reduction in working hours) + possibility of exceptional leave (sick children, moving home, etc.)
- Possibility of teleworking
- Flexible organization of working hours (after 12 months of employment)
- Professional equipment available (videoconferencing, loan of computer equipment, etc.)
- Social, cultural and sports events and activities
- Access to vocational training
- Social security coverage

References

— Md Mahfuz Ibn Alam, Antonios Anastasopoulos, Laurent Besacier, James Cross, Matthias Gallé, Philipp Koehn, and Vassilina Nikoulina. On the evaluation of machine translation for terminology consistency, CoRR, abs/2106.11891, 2021.

— Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 65–72, Ann Arbor, Michigan, 2005.

— Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are Few-Shot learners. In Advances in Neural Information Processing System, pages 1877–1901. Curran Associates, Inc., 2020.

— Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. CoRR, abs/2204.02311, 2022.

— Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. Experts, errors, and context: A large-scale study of human evaluation for machine translation. Transactions of the Association for Computational Linguistics, 9 :1460–1474, 2021.

— Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In Proceedings of the Seventh

Conference on Machine Translation (WMT), pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid), 2022.

— Zhengxian Gong, Min Zhang, and Guodong Zhou. Document-level machine translation evaluation with gist consistency and text cohesion. In Proceedings of the Second Workshop on Discourse in Machine Translation, pages 33–40, Lisbon, Portugal, 2015. Association for Computational Linguistics.

— Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. Continuous measurement scales in human evaluation of machine translation. In Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, pages 33–41, Sofia, Bulgaria, 2013.

— Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. Achieving human parity on automatic chinese to english news translation. CoRR, abs/1803.05567, 2018.

— Yuchen Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. BlonDe: An automatic evaluation metric for document-level machine translation. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, pages 1550–1565, Seattle, United States, 2022.

— Tom Kocmi and Christian Federmann. Large language models are state-of-the-art evaluators of translation quality. CoRR, abs/2302.14520, 2023.

— Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In Proceedings of the Sixth Conference on Machine Translation, pages 478–494, Online, 2021.

— Samuel Läubli, Rico Sennrich, and Martin Volk. Has machine translation achieved human parity? a case for document-level evaluation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4791–4796, Brussels, Belgium, 2018.

— Qingyu Lu, Liang Ding, Liping Xie, Kanjian Zhang, Derek F. Wong, and Dacheng Tao. Toward human-like evaluation for natural language generation with error analysis. CoRR, abs/2212.10179, 2022.

— Qingyu Lu, Baopu Qiu, Liang Ding, Liping Xie, and Dacheng Tao. Error analysis prompting enables Human-Like translation evaluation in large language models: A case study on ChatGPT. CoRR, abs/2303.13809, 2023.

— Amirkeivan Mohtashami, Mauro Verzetti, and Paul K. Rubenstein. Learning translation quality evaluation on low resource languages from large language models. CoRR, abs/2302.03491, 2023.

— Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA, 2002.
— Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 392–395, Lisbon, Portugal, 2015.
— Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2685–2702, Online, 2020.
— Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. QuestEval: Summarization asks for fact-based evaluation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6594–6604, Online and Punta Cana, Dominican Republic, 2021.
— Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7881–7892, Online, 2020.
— Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In Proceedings of the Third Conference on Machine Translation : Research Papers, pages 113–123, Brussels, Belgium, 2018.
— Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. Embarrassingly easy Document-Level MT metrics: How to convert any pretrained metric into a Document-Level metric. In Proceedings of the Seventh Conference on Machine Translation (WMT), pages 118–128, Abu Dhabi, United Arab Emirates (Hybrid), 2022.
— Billy T. M. Wong and Chunyu Kit. Extending machine translation evaluation metrics with lexical cohesion to document level. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 1060–1068, Jeju Island, Korea, 2012.
— Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. CoRR, abs/1609.08144, 2016.
— Weizhe Yuan, Graham Neubig, and Pengfei Liu. BARTScore: Evaluating generated text as text generation. In Curran Associates, Inc., editor, Advances in Neural Information Processing Systems, pages 27263–27277,

Online, 2021.
— Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In 8th International Conference on Learning Representations. Proceedings of the 8th International Conference on Learning Representations, Addis Ababa, Ethiopia, 2020.