Explicitly interpretable attention mechanisms for automatic summarization

Benoit Favre & Thierry Artières
{firstname.lastname}@lis-lab.fr

November 2019

1 Context

Automatic summarization is a natural language processing task which consists in generating a shorter version of a document or set of documents on a specific topic. It is one of the most challenging tasks of the domain because it requires building a full understanding of the input documents, placing the detected facts in the context of background knowledge, evaluating the relevance of each fact, and generating a fluent and concise linguistic representation as output.

While previous summarization generation techniques where mostly extractive, that is based on copy and rearrangement of the input, recent approaches heavily rely on deep learning, by extending the encoder-decoder approach with mechanisms which balance reusing the input with generating new words [11], or using attention mechanisms to capture redundancy [10]. The domain is flourishing with novel approaches[1] that propose better modeling of the problem.

Yet, evaluating the quality of a summarization systems is challenging because there is no notion of gold standard and typical evaluation metrics cannot be applied to text generation tasks. Over the years, researchers have resorted to several methods for evaluation, such as ROUGE [6] which automatically compares system outputs to a set of hand-written summaries, or Pyramid [7] which adds an additional layer of manual alignment to facts for more relevant results. Despite the fact that recent approaches, such as [9, 3], leverage latest development in machine learning, there is a growing suspicion in the adequacy of the whole evaluation setup for the summarization task [5].

The goal of this project is to devise interpretable summarization evaluation metrics, which shall produce reliable and accurate predictions of human evaluation metrics [2]. The approach will be evaluated on data from the TAC evaluation campaigns, in particular those used in the AESOP track [8] for assessing evaluation metrics. In that task, given a summarization task and automatic summarization system outputs, the *evaluator* must predict rankings from human-generated evaluation scores.

[1] A comprehensive list of summarization techniques can be found at https://github.com/mathsyouth/awesome-text-summarization

This project is developed in the context of a collaboration with NIT Silchar in India, and the trainee will have the opportunity to collaborate with colleagues from that project.

2 Learning strategies for enforcing interpretability

We will explore the use of **attention mechanism** [1, 13] for interpretability. An attention mechanism is a model component that is learned from the data to sequentially focus the attention of the prediction model on a part of its input. Although attention seems an appealing idea for interpretability, [12] pointed out that it is not fully justified at least for textual data, which we believe comes from the fact these mechanisms are not actually trained with any interpretability-based criterion.

We will explore how to drive attention mechanism learning to enhance their interpretability capacity. As the usual datasets do not integrate useful supervision for this, we will have to devise strategies to reach our goals, based on e.g. adversarial learning [4] or on new innovative strategies to develop.

The approach will be applied on an automatic summarization task, evaluated through manual annotation of important factors in machine-produced summaries.

References

[1] A. Brown, A. Tuor, B. Hutchinson, and N. Nichols. Recurrent neural network attention mechanisms for interpretable system log anomaly detection. In *Wshp on ML for CS*, 2018.

[2] Hoa Trang Dang and Karolina Owczarzak. Overview of the tac 2008 update summarization task. In *TAC*, 2008.

[3] Yanjun Gao, Chen Sun, and Rebecca J Passonneau. Automated pyramid summarization evaluation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 404–418, 2019.

[4] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680, 2014.

[5] Wojciech Kryściński, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. Neural text summarization: A critical evaluation. *arXiv preprint arXiv:1908.08960*, 2019.

[6] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[7] Ani Nenkova and Rebecca Passonneau. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the human language technology conference of the north american chapter of the association for computational linguistics: Hlt-naacl 2004*, pages 145–152, 2004.

[8] Karolina Owczarzak, John M Conroy, Hoa Trang Dang, and Ani Nenkova. An assessment of the accuracy of automatic evaluation in summarization. In *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*, pages 1–9. Association for Computational Linguistics, 2012.

[9] Stefan Ruseti, Mihai Dascalu, Amy M Johnson, Danielle S McNamara, Renu Balyan, Kathryn S McCarthy, and Stefan Trausan-Matu. Scoring summaries using recurrent neural networks. In *International Conference on Intelligent Tutoring Systems*, pages 191–201. Springer, 2018.

[10] Alexander M Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*, 2015.

[11] Abigail See, Peter J Liu, and Christopher D Manning. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*, 2017.

[12] Sofia Serrano and Noah A. Smith. Is attention interpretable? In *ACL*, 2019.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *NeurIPS*, 2017.