NLP: Conception et évaluation d'un système de résumé automatique abstractif francophone

Vincent Guigue

Equipe : MLIA, en collaboration avec l'entreprise QWAM-CI.
Encadrement : Vincent Guigue (vincent.guigue@lip6.fr)
Profil: Etudiant en master 2 ou école d'ingénieur avec des compétences avancées en programmation, deep learning et idéalement NLP.
Candidature Lettre de motivation & CV à `vincent.guigue@lip6.fr`. Les lettres de recommandation sont un atout. Les entretiens auront lieu dès que possible pour un démarrage en février 2022
Thèse: Ce sujet est prévu pour une poursuite en thèse, dans le cadre d'un projet industriel monté avec l'entreprise QWAM.

1 Contexte

L'évolution récente et rapide des modèles de langues génératifs pré-entrainés ouvre des perspectives en résumé automatique, avec des applications dans différents champs thématiques: compte rendus de réunion, rapport techniques, suivi de projet etc...

Historiquement, la tâche de résumé automatique est considérée comme extrêment dure et les premières approches étaient fondées sur l'extraction de phrases [GMCK00, Mih04]. L'idée de base était de cerner les principales thématiques présentes dans le ou les documents puis de sélectionner des phrases représentantes dans chaque domaine.

Bien que pertinente, cette approche souffre d'un manque de lisibilité et de continuité. Ces approches sont aussi particulièrement difficiles à évaluer. Se baser sur l'apparition des mots par rapport à une vérité terrain, avec une métrique orientée précision [PRWZ02] ou rappel ROUGE [LH03] pose à la fois la question du choix des mots dans la vérité terrain et des synonymies dans les propositions émise [HLZF06].

L'émergence des modèles de langues pré-entrainés tels que ELMo [PNI+18], BERT [DCLT18], GPT [RN18] ou T5 [RSR+20] ouvre des perspectives en génération de texte et en évaluation. Cependant, ces approches soulèvent également de nouvelles problématiques. Au niveau de l'évaluation, les progrès sont de deux natures: d'une part, il est possible d'évaluer la correspondance des phrases dans l'espace latent, ce qui élimine une partie du problème de synonymie [ZKW+19]. Mais l'avancée la plus importante est liée aux progrès en détection d'entités nommées [TGG, TGSG20]. En effet, de nouvelles métriques très prometteuses émergent au niveau des entités voire même du Question Answering [SLPS19, SDL+20].

En parallèle, ces modèles imposent de nouvelles contraintes comme le fait de lutter contre les hallucinations qui sont susceptibles d'apparaître dans les textes générés [RRS+21].

2 Sujet de stage

Le stagiaire prendra en main les outils de l'état de l'art en NLP et en particulier les modèles BERT/BART [LLG+19] et T5 [RSR+20]. Il étudiera attentivement les dernières propositions faites sur la langue française comme CamemBERT [MMS+19] ou FlauBERT [LVF+19].

L'enjeu du stage est de vérifier l'efficacité des modèles et des métriques de résumé automatique sur les corpus francophones. Un tel travail s'accompagnera évidemment de propositions et de développements pour pallier les lacunes qui seront identifiées.

References

[DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[GMCK00] Jade Goldstein, Vibhu O Mittal, Jaime G Carbonell, and Mark Kantrowitz. Multi-document summarization by sentence extraction. In *NAACL-ANLP 2000 Workshop: Automatic Summarization*, 2000.

[HLZF06] Eduard H Hovy, Chin-Yew Lin, Liang Zhou, and Junichi Fukumoto. Automated summarization evaluation with basic elements. In *LREC*, volume 6, pages 604–611. Citeseer, 2006.

[LH03] Chin-Yew Lin and Eduard Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 150–157, 2003.

[LLG+19] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[LVF+19] Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. Flaubert: Unsupervised language model pre-training for french. *arXiv preprint arXiv:1912.05372*, 2019.

[Mih04] Rada Mihalcea. Graph-based ranking algorithms for sentence extraction, applied to text summarization. In *Proceedings of the ACL interactive poster and demonstration sessions*, pages 170–173, 2004.

[MMS+19] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot. Camembert: a tasty french language model. *arXiv preprint arXiv:1911.03894*, 2019.

[PNI+18] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237, 2018.

[PRWZ02] 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, 2002.

[RN18] Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018.

[RRS+21] Clément Rebuffel, Marco Roberti, Laure Soulier, Geoffrey Scoutheeten, Rossella Cancelliere, and Patrick Gallinari. Controlling hallucinations at word level in data-to-text generation. *arXiv preprint arXiv:2102.02810*, 2021.

[RSR+20] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020.

[SDL+20] Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. Mlsum: The multilingual summarization corpus. *arXiv preprint arXiv:2004.14900*, 2020.

[SLPS19] Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. Answers unite! unsupervised metrics for reinforced summarization models. *arXiv preprint arXiv:1909.01610*, 2019.

[TGG] Bruno Taillé, Vincent Guigue, and Patrick Gallinari. Contextualized embeddings in named-entity recognition: An empirical study on generalization. *Advances in Information Retrieval*, 12036:383.

[TGSG20] Bruno Taillé, Vincent Guigue, Geoffrey Scoutheeten, and Patrick Gallinari. Let's stop incorrect comparisons in end-to-end relation extraction! *arXiv preprint arXiv:2009.10684*, 2020.

[ZKW+19] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.