Abstraction ou hallucination ? État des lieux et évaluation du risque pour les modèles de génération de résumés automatiques de type séquence-à-séquence

Eunice Akani

Abstraction ou hallucination ? État des lieux et évaluation du risque pour les modèles de génération de résumés automatiques de type séquence-à-séquence

Eunice Akani
Aix-Marseille Univ, CNRS, LIS, Marseille, France
Enedis, Marseille, France
eunice.akani@lis-lab.fr

RÉSUMÉ

Nous présentons un résumé étendu de l'article (Akani *et al.*, 2022) présenté à la conférence TALN 2022.

ABSTRACT

Abstraction or Hallucination ? Status and Risk assessment for sequence-to-sequence Automatic Text Summarization Models..

We present an extended abstract of the paper (Akani *et al.*, 2022) that was presented at the 2022 TALN conference.

MOTS-CLÉS : Résumé automatique de texte, hallucination, mesure d'évaluation.

KEYWORDS: Automatic text summarization, hallucination, evaluation metric.

1 Résumé étendu

Le résumé automatique de texte par abstraction consiste à générer automatiquement une synthèse d'un document en capturant ses informations importantes. Cette tâche a connu un regain d'intérêt grâce à l'arrivée des modèles Transformers et des modèles de langue pré-entraînés (Vaswani *et al.*, 2017; Devlin *et al.*, 2019). Malgré ces avancées, la tâche reste difficile surtout quand il s'agit d'un résumé par abstraction — résumé un texte en utilisant de nouveaux mots ou des paraphrases. Cao *et al.* (2018) ont montré que 30% des résumés produits par une sélection de systèmes de résumé automatique par abstraction contiennent des informations incohérentes vis-à-vis du document source qui ne sont pas capturées par les métriques d'évaluation habituelles telle que ROUGE (Lin, 2004). Maynez *et al.* (2020) ont qualifié ces incohérences d'hallucinations. Ils définissent une hallucination comme étant une information se trouvant dans le résumé que l'on ne peut déduire à partir du document source. Sur cette base, deux types d'hallucinations sont évoquées : les hallucinations *intrinsèques*, définies comme des mots ou groupes de mots du résumé qui sont tirées du document mais qui ne sont pas déductibles, et les hallucinations *extrinsèques*, qui introduisent des informations hors document. Alors que les hallucinations intrinsèques peuvent générer des contre-sens et être considérées comme des erreurs, les hallucinations extrinsèques sont souvent des abstractions ou introductions d'informations dont on ne peut directement vérifier la factualité qu'à partir d'une connaissance générale dépassant le document à résumer. Une illustration de ces phénomènes est donnée dans l'exemple de la table 1. En analysant

DOCUMENT : Il tourna sept films de la saga, dont "L'espion qui m'aimait", "Rien que pour vos yeux" et "Octopussy". Outre sa carrière cinématographique, Roger Moore s'était illustré au début de sa carrière dans plusieurs séries télévisées comme "Ivanhoé", "Le Saint" ou "Amicalement vôtre", où il partageait l'affiche avec Tony Curtis. Avec plus de cinquante films à son actif, Roger Moore avait quelque peu délaissé le grand écran ces dernières années. Ses dernières apparitions se sont faites essentiellement dans des téléfilms ou des séries. En 2003, il est fait chevalier commandeur de l'Ordre de l'Empire britannique et obtient également, en 2008, le titre de Commandeur des Arts et des Lettres décerné par la France. Très sensible à la cause animale, il soutenait activement l'association PETA. Après trois divorces, Roger Moore était marié depuis 2002 à une richissime danoise, Kristina Tholstrup.	**REF :** Roger Moore [s'est éteint]. [L'acteur britannique] connu pour [son élégance en toutes circonstances et son humour] avait endossé [le costume de James Bond] [de 1973 à 1985].
	PTGEN : "roger moore, association peta, [est décédé] [dimanche] [à l'âge de 85 ans], [a annoncé sa famille à l' afp] .l'association peta a également [fait part de son passage] à roger moore, "" ivanhoé""."
	C2C : [L'acteur britannique] Roger Moore, connu notamment pour son rôle dans "L'espion qui m'aimait", [est décédé] [mardi] [à l'âge de 87 ans], [a annoncé sa famille à la télévision].
	BARTHEZ : L'acteur et [réalisateur américain] Roger Moore, [décédé] [à l'âge de 95 ans], est connu pour son rôle dans "L'espiosn qui m'aimait", "Rien que pour vos yeux" et "Octopussy".
	MT5 : Le [cinéaste] [britannique] Roger Moore [est décédé] [mardi] [à l'âge de 77 ans], [a annoncé son avocat] Tony Curtis.

TABLE 1 – Exemple de sorties des différents systèmes (PTGEN (See *et al.*, 2017), C2C (Martin *et al.*, 2020), BARTHEZ (Kamal Eddine *et al.*, 2021) et MT5 (Xue *et al.*, 2021)) utilisés sur un document du corpus « Orange-Sum Abstract » (Kamal Eddine *et al.*, 2021). La référence est annotée suivant la typologie des abstractions tandis que les résumés candidats sont annotés suivant la typologie des erreurs. Le code couleur est le suivant : AbsNInf, HorsDoc, NonInf et le gris pour les entités hors du document. Le résumé généré par PTGEN est grammaticalement incorrect.

L'acteur britannique Roger Moore, connu notamment pour son rôle dans "L'espion qui m'aimait", est décédé mardi à l'âge de 87 ans a annoncé sa famille à la télévision.

FIGURE 1 – Exemple des types d'hallucination définit par Maynez *et al.* (2020). Résumé d'un article de « Orange-Sum Abstract » généré par le système C2C. En bleu les hallucinations intrinsèques : informations tirées du document mais non inférables à partir de celui-ci. Et en rouge les hallucinations extrinsèques : informations n'étant pas mentionnées dans le document.

plus en détail le système C2C (Figure 1), on remarque que le résumé produit mentionne que « l'acteur est britannique », ce qui n'est pas précisé dans le document. En effet, il est bien mentionné qu'il a été fait chevalier commandeur de l'Ordre de l'Empire britannique mais cela ne signifie pas qu'il est britannique ; c'est donc une hallucination intrinsèque car plusieurs éléments du document sont associés sans être pour autant déductibles à partir de celui-ci. Aussi, le système précise que « l'acteur est décédé mardi à l'âge de 87 ans ». N'étant pas dans le document c'est donc une hallucination extrinsèque. Pour tenter de limiter la production d'hallucinations, Maynez *et al.* (2020) ont utilisé le « textual entailment », c'est-à-dire la capacité d'un texte à inférer un autre texte, comme mesure de sélection d'un résumé le plus aligné à la source possible. Ces travaux ont montré que le « textual entailment » n'était pas une mesure suffisante pour garantir la fidélité du résumé par rapport au document source. Durmus *et al.* (2020) ont proposé d'exploiter un système de question-réponse pour évaluer la factualité d'un résumé, en générant des questions à partir des phrases du résumé et en vérifiant que leur réponse dans le document est identique à celle à l'origine de la question. Bien que plusieurs études aient été publiées sur les hallucinations dans le cadre de résumés automatiques en anglais, il n'y en a aucune sur le français.

L'étude porte sur l'analyse de sorties de systèmes de résumé de l'état de l'art sur un corpus français, Orange-Sum (Kamal Eddine *et al.*, 2021), afin de savoir à quel point les modèles sous-jacents sont sujets aux hallucinations. Ainsi, nous avons d'abord introduit une typologie des erreurs et abstractions contenues dans les résumés. Puis, nous nous sommes concentrons sur les hallucinations extrinsèques, et en particulier sur les entités nommées qui peuvent apparaître dans un résumé hypothèse mais pas dans le document source. Ceci nous a permis de comparer les productions de systèmes de génération RNN [1] et transformers [2] à l'aide d'une analyse manuelle de leurs erreurs, et de proposer une mesure d'évaluation du risque potentiel d'erreurs (le *risque d'hallucination*) lorsqu'un modèle essaie de faire des abstractions sur les entités.

L'analyse des sorties des différents systèmes a montré que malgré des scores ROUGE très intéressants, ces systèmes sont encore affectés par de nombreuses erreurs liées à la factualité des informations qu'ils présentent. Notre mesure d'évaluation du risque d'hallucination sur les entités nommées a permis de constater que les systèmes séquence à séquence à base de transformers prennent énormément de risque en essayant de prédire des entités hors du document. Aussi, l'analyse du corpus Orange-Sum montre les limites de ce corpus. En effet, ces résumés de référence contiennent des informations non présentes dans les documents.

Les typologies ainsi que les détails des résultats obtenus sont disponibles dans l'article original Akani *et al.* (2022).

Références

AKANI E., FAVRE B. & BECHET F. (2022). Abstraction ou hallucination ? état des lieux et évaluation du risque pour les modèles de génération de résumés automatiques de type séquence-à-séquence. In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale / Travaux originaux*, p. 1–10, Avignon, France : Association pour le Traitement Automatique des Langues. Abstraction or Hallucination ? Status and Risk assessment for sequence-to-sequence Automatic.

CAO Z., WEI F., LI W. & LI S. (2018). Faithful to the original : Fact aware neural abstractive summarization. *ArXiv*, **abs/1711.04434**.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

DURMUS E., HE H. & DIAB M. (2020). FEQA : A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 5055–5070, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.454.

KAMAL EDDINE M., TIXIER A. & VAZIRGIANNIS M. (2021). BARThez : a skilled pretrained French sequence-to-sequence model. In *Proceedings of the 2021 Conference on Empirical Methods*

1. Modèle à base de pointeur-générateur (See *et al.*, 2017)
2. CamemBERT2CamemBERT (Martin *et al.*, 2020; Scialom *et al.*, 2020), Barthez (Kamal Eddine *et al.*, 2021) et mT5 (Xue *et al.*, 2021)

in Natural Language Processing, p. 9369–9390, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.740.

LIN C.-Y. (2004). ROUGE : A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, p. 74–81, Barcelona, Spain : Association for Computational Linguistics.

MARTIN L., MULLER B., SUÁREZ P. J. O., DUPONT Y., ROMARY L., DE LA CLERGERIE É. V., SEDDAH D. & SAGOT B. (2020). Camembert : a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

MAYNEZ J., NARAYAN S., BOHNET B. & MCDONALD R. (2020). On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 1906–1919, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.173.

SCIALOM T., DRAY P.-A., LAMPRIER S., PIWOWARSKI B. & STAIANO J. (2020). Mlsum : The multilingual summarization corpus. *arXiv preprint arXiv :2004.14900*.

SEE A., LIU P. & MANNING C. (2017). Get to the point : Summarization with pointer-generator networks. In *Association for Computational Linguistics*.

VASWANI A., SHAZEER N. M., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. & POLOSUKHIN I. (2017). Attention is all you need. *ArXiv*, **abs/1706.03762**.

XUE L., CONSTANT N., ROBERTS A., KALE M., AL-RFOU R., SIDDHANT A., BARUA A. & RAFFEL C. (2021). mT5 : A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 483–498, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.41.