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, Benoit Favre, Frederic Bechet

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[1, 2] Frédéric Bechet[1] Benoit Favre[1]
(1) Aix-Marseille Univ, CNRS, LIS, Marseille, France
(2) Enedis, Marseille, France
prenom.nom@lis-lab.fr

RÉSUMÉ

La génération de texte a récemment connu un très fort intérêt au vu des avancées notables dans le domaine des modèles de langage neuronaux. Malgré ces avancées, cette tâche reste difficile quand il s'agit d'un résumé automatique de texte par abstraction. Certains systèmes de résumés génèrent des textes qui ne sont pas forcément fidèles au document source. C'est sur cette thématique que porte notre étude. Nous présentons une typologie d'erreurs pour les résumés automatique et ainsi qu'une caractérisation du phénomène de l'abstraction pour les résumés de référence afin de mieux comprendre l'ampleur de ces différents phénomènes sur les entités nommées. Nous proposons également une mesure d'évaluation du risque d'erreur lorsqu'un système tente de faire des abstractions sur les entités nommées d'un document.

ABSTRACT

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

Text generation has recently received more interest due to the significant advances in the field of neural language models. Despite these advances, the task remains difficult when it comes to abstractive text summarization. Some text summarization systems generate texts that are not necessarily faithful to the source document. This is the focus of our study. We present a typology of errors for automatic summaries, and characterization of the abstraction phenomenon for reference summaries to understand the extent of these different phenomena on named entities. We also propose a metric to evaluate the risk of errors when a system attempts to abstract named entities from a document.

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

KEYWORDS: Automatic text summarization, hallucination, evaluation metric.

1 Introduction

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 utilisant de nouveaux mots ou des paraphrases. Cao *et al.*,

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 les entités hors du document. Le résumé généré par PTGEN est grammaticalement incorrect.

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 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 ». Cela n'est pas mentionné dans le document ; c'est par conséquent 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

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 » (Kamal Eddine *et al.*, 2021) généré par le système CAMEMBERT2CAMEMBERT (C2C) (Martin *et al.*, 2020). En bleu les hallucinations extrinsèques : informations tirées du document mais non inférable à partir de celui-ci. Et en rouge les hallucinations extrinsèques : informations n'étant pas mentionné dans le document.

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.

Dans cet article, nous analysons les sorties de systèmes de résumé de l'état de l'art sur un corpus français afin de savoir à quel point les modèles sous-jacents sont sujets aux hallucinations. Nous proposons tout d'abord une typologie des erreurs et abstractions contenues dans les résumés. Puis, nous nous 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 permet de comparer les productions de systèmes de génération RNN et transformers 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.

2 Typologie d'hallucination dans le résumé automatique

Le concept d'hallucination trouvé dans la littérature anglophone fait appel à la fois à la notion d'erreur et à celle d'abstraction, ce qui rend peu intuitives les analyses que l'on peut en tirer. Nous proposons de l'affiner sous la forme d'une typologie double.

Typologies des erreurs Plusieurs études proposent une typologie des erreurs du résumé automatique. Alors que Maynez *et al.*, 2020 se sont concentrés sur les types d'hallucination intrinsèque et extrinsèque, Pagnoni *et al.*, 2021 décrivent une typologie d'erreurs plus détaillée prenant en compte la vérification du contenu, les erreurs liées au discours et à la sémantique de surface. Une étude plus ancienne définit un bon résumé comme étant un texte grammatical, non redondant, dont les références sont claires, et qui est structuré et cohérent (Dang, 2005). Nous nous appuyons sur ces typologies pour proposer une liste d'erreurs adaptée à l'annotation des résumés produits par les systèmes par abstraction :

- **hors du document (HorsDoc)** : informations qui ne proviennent pas du document source ;
- **agrammaticalité (Agram)** : concerne les phrases qui ne sont pas correctement formées ;
- **erreur de référence (RefE)** : lorsqu'une proposition a un sujet erroné ou remplacé par un autre. Il concerne également les pronoms non résolus (erreurs de coréférence) ;
- **contresens (CtrSens)** : toutes les informations qui contredisent le document source ;
- **non inférable (NonInf)** : informations non inférerable à partir du document ;

— **autres** : autres erreurs.

Typologie des abstractions A la différence de Pagnoni *et al.*, 2021, nous avons décidé d'annoter les résumés de référence afin de comprendre les différents types d'abstraction. En effet Durmus *et al.*, 2020 a montré que l'abstractivité d'un modèle dépend des données (plus les données d'entraînement sont abstractives et plus les modèles tenteront d'atteindre le même niveau d'abstractivité). Ainsi, nous avons identifié 3 catégories d'abstractions :

— les abstractions inférables à partir du document source (**AbsDoc**) : ce sont des abstractions qui peuvent être déduites en lisant le document source ;
— les abstractions inférables à partir des connaissances générales de l'annotateur (**AbsInf**) : ce sont des abstractions que l'annotateur peut déduire grâce à son bagage intellectuel et sa culture générale ;
— les abstractions non inférables (**AbsNInf**) : ce sont des abstractions qui peuvent être déduites ni depuis le document, ni par l'annotateur.

Le tableau 1 présente un exemple contenant quelques erreurs des différents systèmes ainsi que quelques abstractions du résumé de référence.

3 Contexte expérimental

Dans cette étude nous nous sommes intéressés au corpus *Orange-Sum* car il est l'un des seuls corpus disponibles pour la langue française de taille suffisante pour apprendre un modèle de génération de résumés par abstraction. Enfin nous avons considéré 4 modèles de génération.

Données Pour nos expériences, nous avons utilisé le corpus de résumé automatique en français Orange-Sum (Kamal Eddine *et al.*, 2021). C'est un corpus d'articles d'actualité provenant du site « Orange Actu » de février 2011 à septembre 2020. Il comprend plusieurs catégories (la France, le monde, la politique, l'automobile et la société) et possède deux versions :

— « Orange-Sum Title » : dans cette version ce sont les titres des articles qui sont considérés comme leurs résumés ;
— « Orange-Sum Abstract » : dans cette version ce sont les *chapeaux* ou les *accroches* des articles qui sont pris comme *résumés*.

Comme on peut le voir, dans le corpus « Orange-Sum Title » il n'y a pas de résumés fait manuellement mais plutôt des *pseudo-résumés*. L'avantage est bien évidemment la taille du corpus disponible ; l'inconvénient est le fait qu'il n'y a qu'un seul résumé par document et que ce *pseudo-résumé* ne peut pas être considéré comme un véritable résumé du document, comme nous le verrons dans l'analyse du corpus.

Dans notre cas, nous avons utilisé « Orange-Sum Abstract » qui était beaucoup plus adapté à nos objectifs de part la taille des *pseudo-résumés*. Il est constitué de 21 401 documents et résumés pour l'entraînement, 1500 pour la validation et le test. Les documents sources et les résumés font respectivement en moyenne 350 mots / 12,06 phrases et 32,12 mots / 1,43 phrases.

Systèmes de résumé automatique Les systèmes suivants sont considérés dans cette étude :

— **PtGen** (See *et al.*, 2017) : modèle séquence à séquence basé sur un RNN avec un mécanisme d'attention, un pointeur/générateur et un mécanisme de couverture.

— **CAMEMBERT2CAMEMBERT (CTC)** (Martin *et al.*, 2020; Scialom *et al.*, 2020) : transformer seq2seq initialisé à partir des poids de CamemBERT-base et pré-entraîné sur la tâche de résumé automatique sur la partie française de MLSUM.
— **BARTHEZ** (Kamal Eddine *et al.*, 2021) : Modèle français basé sur l'architecture BART (Lewis *et al.*, 2020), possédant 6 couches bidirectionnels pour l'encodeur et le décodeur à base de transfomers.
— **MT5** (Xue *et al.*, 2021) : variante multilingue de T5 (Raffel *et al.*, 2020). Il a été initialisé avec les poids de mT5 entraîné sur XLSUM (Hasan *et al.*, 2021), corpus de résumé multilingue.

Un finetuning des modèles à base de transformers sur Orange-Sum a été fait pour nos expériences sauf celui de BARTHEZ dont les poids entraînés par Kamal Eddine *et al.*, 2021 [1] ont été utilisés en inférence.

4 Expériences et Résultats

Les expériences suivantes n'ont pas pour objectif d'améliorer les performances des différents modèles, mais d'étudier leurs sorties ainsi que mesurer le risque d'erreur lorsqu'ils tentent de faire des abstractions. Il reste tout de même intéressant d'évaluer leurs performances en terme de ROUGE et de BERTScore. Cela permet notamment de voir l'informativité des résumés générés.

Mesures d'évaluation Pour évaluer les systèmes, nous avons utilisé ROUGE (Lin, 2004) qui permet de tester les capacités du modèles à générer des ngrams proches du résumé de référence et BERTScore (Zhang* *et al.*, 2020), une mesure de similarité à base de plongements contextuels appliquée à la séquence de tokens des résumés de référence et candidats. La table 2 présente les résultats obtenus sur le jeu de test d'Orange-Sum. Les modèles à base de transformers ont des performances équivalentes tandis que celui à base de RNN a les scores les plus bas. Nos résultats sont similaires à ceux présentés par Kamal Eddine *et al.*, 2021.

Models	R-1	R-2	R-L	BERTScore
PTGEN (RNN)	28.16	9.55	19.12	19.80 / 69.95
C2C	**31.83**	**13.22**	22.87	27.04 / 72.66
BARTHEZ	31.81	13.20	**23.07**	**28.63 / 73.25**
MT5	30.77	12.64	22.49	27.59 / 72.86

TABLE 2 – ROUGE (ROUGE-1, ROUGE-2, ROUGE-L) et BERTScore F1 pour les différents modèles. Les deux valeurs du BERTScore correspondent au BERTScore avec et sans le paramètre rescale.

Procédure d'annotation Nous avons utilisé les 100 premiers exemples du jeu de test sur lesquels nous avons généré des résumés à l'aide des 4 modèles soit 400 résumés annotés. En considérant la référence comme information complémentaire au document source, nous avons suivi la procédure d'annotation suivante pour les résumés automatiques et les résumés références :

1. https://huggingface.co/moussaKam/barthez-orangesum-abstract

— Lecture du document ainsi que les différents résumés générés par les systèmes, surlignage et annotation des informations non fidèles à la source et à la référence suivant la typologie des erreurs définie au chapitre 2.
— Annotation des références suivant la typologie d'abstraction pour comprendre le type d'abstraction montrées en exemple aux systèmes lors de l'apprentissage.

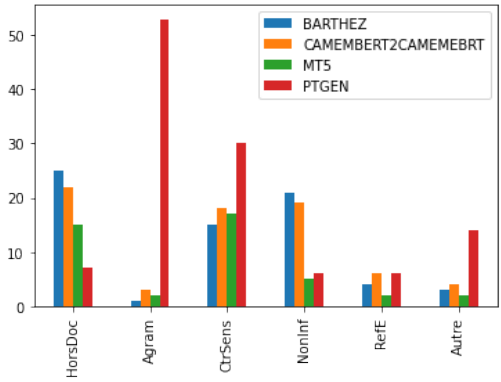

FIGURE 2 – Les différents types d'erreurs dans chaque système.

Models	% de résumés avec au moins une erreur	#nb moyen d'erreurs par résumé
PTGEN	88	1.29
C2C	44	1.26
BARTHEZ	43	1.27
MT5	**30**	**1.12**

TABLE 3 – Statistiques des erreurs de chaque système.

Résultats obtenus La figure 2 et le tableau 3 montrent les résultats obtenus. On remarque que PTGEN génère moins d'informations hors document *(HorsDoc)* comparé aux systèmes à base de Transformers mais en même temps il produit le plus de résumés erronés à cause des nombreux contresens *(CtrSens)* auxquels il est sujet. BARTHEZ et C2C ont des erreurs fréquentes similaires ; il s'agit d'informations hors du document source *(HorsDoc)* et non directement inférables *(NonInf)*. Cela montre leur tentative à faire des abstractions. Au vu des résultats de l'annotation, MT5 est le système qui commet le moins d'erreurs car seulement 30% des résumés associés contiennent au moins une erreur.

En ce qui concerne les résumés de référence, 65% d'entre eux contiennent au moins une abstraction annotée (soit en moyenne 1,17 abstractions par résumé). Les statistiques détaillées des abstractions sont consignées dans la table 4 (%abstraction représente le pourcentage d'apparition de chaque type d'abstraction). Ces résultats montrent que les résumés d'Orange-Sum contiennent souvent des informations complémentaires au document source mais non inférables à partir de celui-ci. On peut ainsi se poser la question de la pertinence de l'heuristique choisie pour constituer ce corpus qui consistait à considérer que les *chapeaux* ou *accroches* d'articles constituaient des résumés de ces mêmes articles.

	%abstraction
ABSDOC	26.5
ABSINF	13.68
ABSNINF	59.83

TABLE 4 – Pourcentage d'apparition des différents types d'abstraction dans les 100 résumés de référence annotés.

Études sur les entités nommées Notre étude se penche également sur les entités, nous avons donc étudié le *risque* d'hallucination des modèles lorsqu'ils tentent de faire des abstractions sur les entités nommées. Pour ce faire, nous avons évalué manuellement les 100 premiers résumés du jeu de test à partir des annotations précédentes. L'évaluation manuelle consiste à classer le type d'erreur sur les entités lorsqu'il y en a (soit hors du document ou mal utilisée). Ensuite nous avons calculé le nombre d'erreurs sur les entités mais également le nombre d'entités hors du document potentiellement incorrectes (le risque d'erreur).

En effet, lorsqu'une entité n'est pas dans le document source, il est difficile d'affirmer qu'elle est correcte ou non sans faire appel à une source extérieure. Ainsi, la source complémentaire que nous avons utilisé pour affirmer qu'une entité n'est pas correcte est le résumé de référence ; C'est à dire que lorsqu'une entité n'est pas dans le document source mais qu'elle apparaît dans le résumé de référence, elle est considérée comme correcte.

Ces résultats sont confirmés par une évaluation automatique sur la totalité du jeu de test (1500) de l'intersection entre les entités nommées des résumés produits par les systèmes et celles du document source et du résumé de référence. Pour ce faire, à l'aide d'un système de reconnaissance d'entités nommées (Akbik *et al.*, 2018), nous avons pu extraire les entités de type *PER* (nom de personne), *LOC* (nom de lieu), *ORG* (nom d'organisation) et *MISC* (autre nom). Comme les entités de type date, quantité, heure, pourcentage, et valeur numérique ne sont pas reconnues par le système, nous les avons extraites à l'aide d'expressions régulières écrites pour l'occasion. Ensuite nous avons calculé le nombre d'entités hors du document et celles qui sont potentiellement erronées. Les résultats sont consignés dans le tableau 5.

	Manuel (100 résumés)		**Auto (1500 résumés)**	
Models	$\neg Doc$	$\neg Doc \cap \neg Ref$	$\neg Doc$	$\neg Doc \cap \neg Ref$
PTGEN	**5.4**	100	**6.6**	97.2
C2C	16.53	90.47	9.7	87.5
BARTHEZ	21.12	**74.07**	15.4	**80.8**
MT5	13.39	93.33	8.6	88.7

TABLE 5 – Statistiques sur les prédictions des entités nommées avec analyse manuelle sur les 100 résumés et analyse automatique sur les 1500. $\neg Doc$ désigne le pourcentage d'entités en dehors du document. $\neg Doc \cap \neg Ref$ désigne, parmi les entités hors du document, le pourcentage d'entités qui ne sont pas dans le résumé de référence.

La deuxième et la dernière colonnes contiennent respectivement les analyses manuelles sur les 100 résumés et automatiques sur les 1500 résumés. On mesure le *risque* d'hallucination en calculant le nombre d'entités prédites qui ne sont pas dans le document à résumer ($\neg Doc$), et le % d'entre eux qui sont potentiellement des hallucinations car non présents dans les résumés de référence ($\neg Doc \cap \neg Ref$). Pour les analyses manuelles, il est a noté que les entités ont été identifiées par l'annotateur.

Les résultats obtenus par l'évaluation manuelle est du même ordre que ceux obtenus automatiquement. Le tableau 5 nous montre qu'à chaque fois que PTGEN essaie de faire des abstractions en utilisant des entités hors du document, il génère des entités potentiellement erronées car dans la majorité des cas, les entités prédites hors du document ne sont pas dans la référence. BARTHEZ est le système qui prend le plus de risque car il prédit le plus grand nombre d'entités hors du document mais c'est le système qui a le plus grand BERTScore et est parmi les meilleurs en terme de ROUGE.

5 Discussion

Les différents résultats obtenus ont permis de confirmer que l'abstractivité des données agit sur les modèles et les encourage à utiliser des informations hors du document, ce qui confirme l'assertion de Durmus *et al.*, 2020. Les résultats obtenus par l'annotation des résumés de référence portent l'attention sur l'adaptation du corpus pour la de tâche de résumé automatique par abstraction. En effet, il s'agit d'un corpus de nouvelles dont les résumés contiennent des informations qui ne sont pas présentes dans le document. Cela s'explique du fait que les références ont été produites à partir du chapeau de l'article qui est écrit afin de pousser le lecteur à s'intéresser à l'article et est parfois une information complémentaire à l'article plutôt qu'un résumé en tant que tel.

Limites Notre mesure d'évaluation du risque d'hallucination nous 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. Les entités étant détectées à l'aide de systèmes de reconnaissance automatique et de règles, la mesure dépend de la qualité de ces systèmes. L'annotation manuelle des erreurs ayant été faite sur seulement 100 résumés, il serait intéressant d'augmenter le nombre de résumés annotés et de déterminer un score d'accord inter-annotateurs. Néanmoins, les résultats obtenus en évaluant automatiquement les différents modèles à l'aide de notre mesure vont dans le même sens que ceux obtenus manuellement.

6 Conclusion

Nous avons proposé une typologie des erreurs et abstractions en résumé automatique, et avons annoté un corpus de sorties de systèmes à l'aide de cette typologie. 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. Cette étude ouvre plusieurs axes de recherches comme le fait de minimiser le risque d'hallucination de chaque modèle en les pénalisant lorsqu'ils génèrent des entités nommées hors du document source ou encore de faire un modèle de sélection des meilleurs résumés suivant le risque d'hallucination. Ces travaux ont aussi montré les limites du corpus Orange-Sum pour l'apprentissage de systèmes, car ces résumés de référence contiennent des informations non présentes dans les documents. Ainsi, il peut être intéressant d'analyser le comportement des modèles entraînés sur un jeu de données dont les résumés de référence contenant des informations hors document ont été supprimés.

Remerciements

Remerciement à Romain Gemignani, responsable innovation à Enedis pour son soutien et ses conseils pendant la réalisation de ces travaux.

Ces travaux ont bénéficié d'un accès aux moyens de calcul de l'IDRIS au travers de l'allocation de ressources 2021-AD011012525 attribuée par GENCI.

Références

AKBIK A., BLYTHE D. & VOLLGRAF R. (2018). Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, p. 1638–1649.

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

DANG H. T. (2005). Overview of duc 2005. In *In Proceedings of the Document Understanding Conf. Wksp. 2005 (DUC 2005) at the Human Language Technology Conf./Conf. on Empirical Methods in Natural Language Processing (HLT/EMNLP*.

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.

HASAN T., BHATTACHARJEE A., ISLAM M. S., MUBASSHIR K., LI Y.-F., KANG Y.-B., RAHMAN M. S. & SHAHRIYAR R. (2021). XL-sum : Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics : ACL-IJCNLP 2021*, p. 4693–4703, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.findings-acl.413.

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 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.

LEWIS M., LIU Y., GOYAL N., GHAZVININEJAD M., MOHAMED A., LEVY O., STOYANOV V. & ZETTLEMOYER L. (2020). BART : Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7871–7880, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.703.

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.

PAGNONI A., BALACHANDRAN V. & TSVETKOV Y. (2021). Understanding factuality in abstractive summarization with FRANK : A benchmark for factuality metrics. In *Proceedings of the 2021*

Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, p. 4812–4829, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.383.

RAFFEL C., SHAZEER N., ROBERTS A., LEE K., NARANG S., MATENA M., ZHOU Y., LI W. & LIU P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, **21**(140), 1–67.

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.

ZHANG* T., KISHORE* V., WU* F., WEINBERGER K. Q. & ARTZI Y. (2020). Bertscore : Evaluating text generation with bert. In *International Conference on Learning Representations*.