Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Étude de la fidélité des entités dans les résumés par abstraction

Eunice Akani[1, 2]
(1) Aix-Marseille Univ, CNRS, LIS, Marseille, France
(2) Enedis, Marseille, France
eunice.akani@lis-lab.fr

RÉSUMÉ

L'un des problèmes majeurs dans le résumé automatique de texte par abstraction est la fidélité du résumé généré vis-à-vis du document. Les systèmes peuvent produire des informations incohérentes vis-à-vis du document. Ici, nous mettons l'accent sur ce phénomène en restant focalisé sur les entités nommées. L'objectif est de réduire les hallucinations sur celles-ci. Ainsi, nous avons généré des résumés par sampling et avons sélectionné, à l'aide d'un critère basé sur le risque d'hallucination sur les entités et les performances du modèle, ceux qui minimisent les hallucinations sur les entités. Une étude empirique du critère montre son adaptabilité pour la sélection de résumé. Nous avons proposé des heuristiques pour la détection des entités qui sont des variations ou flexions d'autres entités. Les résultats obtenus montrent que le critère réduit les hallucinations sur les entités nommées en gardant un score ROUGE comparable pour CNN/DM.

ABSTRACT

Named Entities Faithfulness in Abstractive Text Summarization.

One of the major problems in abstraction text summarization is the faithfulness of the generated summary regard to the source document. Systems may produce inconsistent information reading the document. Here, we emphasize this phenomenon by remaining focused on named entities. The goal is to reduce hallucinations on these entities. Thus, we generated summaries by sampling and selected, using a criterion based on the risk of hallucination on entities and the performance of the model, those which minimize the hallucinations on the entities. An empirical study of the criterion shows its adaptability for summary selection. We also proposed heuristics for the detection of entities which are variations or inflections of other entities. The results obtained show that the criterion reduces hallucinations on named entities by keeping a comparable ROUGE score for CNN/DM.

MOTS-CLÉS : Résumé automatique de texte, hallucination, entité nommée.

KEYWORDS: Automatic text summarization, hallucination, named entity.

1 Introduction

Le résumé automatique de texte consiste à faire une synthèse d'un document en gardant les informations pertinentes. Il existe deux types de résumé automatique : le résumé par extraction et le résumé par abstraction. Le résumé par extraction qui consiste à extraire dans un premier temps les informations importantes du document puis dans un second temps à en faire un résumé. Le résumé par abstraction consiste quant à lui à faire la synthèse d'un document en utilisant des paraphrases et des nouveaux mots. La tâche de résumé automatique a beaucoup évolué depuis l'apparition des modèles à

Actes des 16e Rencontres Jeunes Chercheurs en RI (RJCRI) et 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL), pages 21–36.

base de transformers et de modèles de langue pré-entraînés (Vaswani *et al.*, 2017; Devlin *et al.*, 2019; Lewis *et al.*, 2020). Pendant que la tâche de résumé par extraction devient de plus en plus atteignable, celle du résumé par abstraction reste un grand défi. En effet, Kryscinski *et al.*, 2019 a montré que les systèmes produisent des informations qui ne sont pas toujours fidèles au document source et cela n'est pas vérifiable par l'utilisation de mesures d'évaluation du résumé telles que ROUGE (Lin, 2004), METEOR (Banerjee & Lavie, 2005) ou encore le BERTScore (Zhang* *et al.*, 2020). 30% des résumés générés par des systèmes abstractifs contiennent des incohérences vis-à-vis du document source selon Cao *et al.*, 2018. Maynez *et al.*, 2020 a utilisé le terme « hallucination » pour qualifier ces incohérences. Elles peuvent être dues à une mauvaise association d'informations provenant du document ou à l'utilisation d'informations hors du document. Plusieurs études ont été menées pour l'évaluation factuelle des résumés générés (Durmus *et al.*, 2020; Wang *et al.*, 2020) ainsi que pour la réduction d'hallucinations (Pagnoni *et al.*, 2021; Chen *et al.*, 2021; Fan *et al.*, 2018a). Pour notre part, nous étudions les hallucinations au niveau des entités nommées, ce qui s'est avéré assez fréquent dans les résumés générés (Chen *et al.*, 2021). Plutôt que d'intervenir en amont, pendant l'entraînement du modèle, nous avons généré plusieurs résumés dans un espace afin de sélectionner le meilleur résumé suivant un critère qui pourra réduire les hallucinations et donc corriger les problèmes de fidélité du résumé généré par rapport au document source. L'idée étant de réduire le nombre d'entités hallucinées, nous avons introduit un critère de sélection de résumé basé sur le « risque » d'avoir des entités qui n'apparaissent pas dans le document source (entités hallucinées). Ce critère nous permet ainsi de choisir un résumé avec le moins d'entités hallucinées dans l'ensemble des résumés générés. Nous avons sélectionné, conjointement, les résumés ayant le meilleur score ROUGE avec la référence et avions procédé à une évaluation humaine de leurs entités pour vérifier que les entités du résumé sont correctement utilisées (leur emploi ne contredit pas le document source), et que les entités dites hors du document le sont vraiment. Cette évaluation, permet ainsi d'étudier de manière empirique le critère de sélection. Ce critère étant basé sur les entités nommées, nous avons donc mis sur place des heuristiques afin de détecter les variations d'entités nommées (ex. les nombres écrits en lettres ou en chiffres, les erreurs d'orthographe, les flexions). Nos contributions se résument donc à :

— La génération de divers résumés avec la méthode de sampling afin d'avoir un résumé qui minimise les erreurs factuelles.
— L'introduction d'un critère de sélection de résumé avec le moins d'entités hors du document.
— La mise en place d'heuristique pour la détection de variation d'entités.

Document :
She said she will seek a judicial review against Mark H Durkan because he took the decision to adopt the planning policy without the agreement of the full Northern Ireland Executive. Helen Jones reports for BBC Newsline.

Résumé :
Helen Jones seek a judicial review against Mark H Durkan, the Northern Ireland's environment minister reports BBC.

FIGURE 1 – Exemple d'hallucination intrinsèque (en bleu) et extrinsèque (en rouge).

2 Contexte et motivation

L'évaluation des résumés automatiques en terme de fidélité est de plus en plus importante. En effet, un résumé n'étant pas fidèle au document source n'est pas exploitable car rempli d'incohérences. Cela suscite l'attention des chercheurs en trois questions : pourquoi se focaliser sur la fidélité, comment évaluer la fidélité d'un résumé, et comment réduire les hallucinations ?

Certains travaux présentent l'insuffisance des mesures d'évaluation actuelles par catégorisation des erreurs fréquentes dans le résumé suivant des typologies d'erreurs qu'ils ont défini (Ji *et al.*, 2023). Maynez *et al.*, 2020 ont nommés les informations incohérentes vis-à-vis du document des hallucinations. Ils ont défini deux types d'hallucinations, les hallucinations intrinsèques qui sont les informations provenant du document qui ne sont pas cohérentes vis-à-vis de celui-ci et les hallucinations extrinsèques qui sont définies comme des informations du résumé qui sont hors du document source. Un exemple est donné sur la figure 1. Pagnoni *et al.*, 2021 propose une typologie plus détaillée prenant en compte la vérification de contenu, les erreurs liées au discours et la sémantique de surface. Ils ont mené une évaluation humaine à grande échelle sur des résumés de divers systèmes en utilisant le datatset CNN/Daily Mail (Hermann *et al.*, 2015; Nallapati *et al.*, 2016) et XSum (Narayan *et al.*, 2018) afin d'identifier les erreurs fréquentes dans les résumés de référence. L'analyse des hallucinations extrinsèques par Chen *et al.*, 2021 montre que, pour la majorité, cela se produit sur des entités et quantités nommées. Akani *et al.*, 2022 propose une typologie d'erreurs pour les résumés candidats et une typologie d'abstraction pour les résumés de référence. Cela leur a permis de montrer que les erreurs les plus courantes étaient les informations provenant hors du document source et les informations non inférables à partir du document source.

Pour tenter de résoudre le problème des hallucinations, certains proposent l'utilisation d'entailment textuel (Maynez *et al.*, 2020) ou encore de modèle de question-réponse (Durmus *et al.*, 2020; Wang *et al.*, 2020) pour l'évaluation de la fidélité du résumé par rapport au document source. Chen *et al.*, 2021 propose de corriger les erreurs sur les entités nommées en modifiant les entités des résumés générés par des entités du document source. Puis à l'aide des modèles qu'ils ont entraînés, ils choisissent les résumés factuellement cohérents. Cela leur permet d'éviter les hallucinations extrinsèques, mais la méthode crée des hallucinations intrinsèques dans les résumés. (Nan *et al.*, 2021) propose une méthode basée sur le filtrage des données d'entraînement et l'apprentissage multi-tâches. Fan *et al.*, 2018a propose de contrôler la génération du résumé avec une liste d'entités nommées désirées en entrée, et Narayan *et al.*, 2021 propose, pendant l'entraînement du modèle, de conditionner la génération des résumés par les entités nommées en générant d'abord la liste des entités à utiliser dans le résumé, puis le résumé lui-même.

Notre étude est proche des différentes études sur les entités nommées tant l'objectif est de réduire les hallucinations sur ceux-ci. L'étude la plus proche de la nôtre est celle de Chen *et al.*, 2021 car il s'agit de choisir un résumé parmi un ensemble de résumés créé ou généré. La différence réside dans la génération des résumés et dans la sélection du meilleur résumé. En effet, plutôt que de modifier les résumés générés après leur génération par le modèle, nous générons un ensemble de résumé par la méthode de sampling afin de couvrir un grand espace de génération multiple. La section suivante présente la mise en place de cette génération.

Dans la suite de ce document, nous appellerons hallucination, de manière générale, uniquement les informations en dehors du document source. Le terme « entités hallucinées » correspondra donc aux entités qui ne sont pas dans le document source. Aussi, il est à noter que nous utiliserons le terme « hallucination factuelle » pour parler des informations hors du document qui sont factuellement

correctes.

3 Génération de résumés par échantillonnage (sampling)

La plupart des méthodes pour la génération automatique de texte se font par l'utilisation du « beam search » pour parcourir l'espace de recherche de manière efficace. Le « beam search » conserve les num_beams ayant la probabilité élevée à chaque étape. Il minimise la possibilité de manquer des séquences cachées à forte probabilité. Cela ne permet pas d'avoir un ensemble de résumés variés dans l'espace de génération. De cette façon, la méthode de sampling est la plus adaptée pour générer plusieurs résumés. Nous avons donc généré des résumés en l'utilisant pour la sélection du prochain token à générer. Ainsi, les résumés obtenus proviennent de l'utilisation de « greedy search », « beam search », la température, du top-P et du top-K. Tandis que le « greedy search » consiste à sélectionner le token ayant la probabilité la plus élevée, l'échantillonnage avec la température consiste à remettre à l'échelle les logits avant d'appliquer la fonction softmax. Le Top-K (Fan *et al.*, 2018b) consiste à prendre les K mots suivants les plus probables et à redistribuer la probabilité entre ces K mots. Le Top-P quant à lui ou échantillonnage Nucleus (Holtzman *et al.*, 2019) consiste, sachant une probabilité p, à prendre le plus petit ensemble possible de mots suivants dont la probabilité cumulée est supérieure à la probabilité p. Il y a également une redistribution de la probabilité entre les mots de l'ensemble. Pour notre part, nous avons fait varier les différents paramètres comme suit :

— Pour la température : de 0.5 à 1 avec un pas de 0.1.
$T = [0.5, 0.6, 0.7, 0.8, 0.9]$;
— Pour le Top-p : de 0.75 à 0.96 avec un pas de 0.05
$Top - p = [0.75, 0.80, 0.85, 0.90, 0.95]$;
— Pour le Top-k de 40 à 70 avec un pas de 10
$Top - k = [40, 50, 60]$

Nous avons généré ainsi 77 résumés [1] pour chaque exemple en faisant varier les différentes valeurs enumérées plus haut à chaque étape. Nous avons pris 1024 comme le nombre de tokens maximum pour l'entrée du système et 128 tokens maximum pour la sortie du système (les résumés).

Dans la suite du papier, nous introduisons un critère de sélection de résumé basé sur les entités nommées afin de réduire les hallucinations sur celles-ci.

4 Sélection de résumé par critère

Pour réduire le problème de fidélité des entités du résumé généré par rapport au document source, il est important d'étudier le risque d'hallucinations sur les entités nommées. Nous avons donc introduit « NEHR Named Entities Hallucination Risk » une mesure sur les entités nommées comme (Nan *et al.*, 2021). Cette mesure peut être utilisée pour construire un modèle de sélection de résumés minimisant les hallucinations sur les entités nommées. Certains ont proposé d'utiliser l'entailment textuel (Falke *et al.*, 2019; Maynez *et al.*, 2020) ou encore un système de question-réponse (Durmus *et al.*, 2020) afin d'évaluer la fidélité du résumé par rapport au document source. Ce sont des méthodes qui sont plutôt gourmandes en termes de ressources. Pour notre part, nous avons décidé de mettre en place une heuristique qui considère que si une entité est en dehors du document source alors elle est hallucinée.

1. 75 avec les méthodes de d'échantillonnage + la génération greedy + la génération avec le beam de taille 4

Elle rend donc le résumé non fidèle au document source. Ceci nous permet de définir le critère comme suit.

4.1 Définition

Pour un document d et un résumé s :

$$\textit{NEHR}(d, s) = (1 - \frac{|\textit{entities} \in d \wedge s|}{|entities \in s|}) \times 100 \quad (1)$$

Pour la détection des entités nommées, nous avons utilisé un système automatique de reconnaissance d'entités nommées. Ce critère n'inclut pas la référence de telle sorte qu'il puisse être utilisé en prédiction. Cependant, n'ayant pas de moyen de dire si une entité considérée comme risquée est correcte ou pas, nous avons mené une étude empirique pour savoir si NEHR est corrélé aux hallucinations dans les résumés générés. Dans la sous-section suivante, nous présenterons cette étude.

4.2 Dataset et modèle de résumé automatique

Pour notre étude, nous avons utilisé le corpus CNN/Daily Mail (Hermann *et al.*, 2015; Nallapati *et al.*, 2016) et le corpus XSum (Narayan *et al.*, 2018). CNN/Daily Mail est un corpus populaire pour la tâche de résumé automatique. Il est composé d'articles provenant des sites de CNN et Daily Mail. XSum est un corpus composé de 226 711 articles provenant de BBC de 2010 à 2017. Les différents articles traitent de plusieurs sujets notamment, l'actualité, l'éducation, le business, la météo, la technologie, la santé, la politique, le sport, etc. La particularité de ce corpus est que les résumés ont été écrits par des professionnels. Les résumés de XSum sont faits en une seule phrase. Ce corpus a été introduit comme corpus pour la tâche de résumé automatique par abstraction, car il contient 36% de nouveaux unigrams. Comme modèle, nous avons BART (Lewis *et al.*, 2020), c'est une architecture basée sur les Transformers (Vaswani *et al.*, 2017) qui est utilisée pour la tâche de résumé automatique de texte. Il existe différentes tailles du modèle. En ce qui nous concerne, nous avons utilisé BART-large qui se compose de 12 couches de Transformers aussi bien dans l'encodeur que dans le décodeur. Nous avons initialisé le modèle avec les poids du modèle se trouvant sur Hugging Face (Wolf *et al.*, 2020) aussi bien pour CNN/DM [2] que pour XSum [3].

4.3 Étude empirique du risque

Vérification de la pertinence du critère. L'étude empirique utilisée pour vérifier la pertinence du critère est la suivante :

1. Sélectionner un corpus C de résumé automatique contenant des documents sources et les résumés associés. Entraîner plusieurs systèmes de résumé automatique sur l'ensemble d'entraînement. Puis à l'aide du modèle entraîné, générer un ensemble de résumés S_d alternatifs pour le document d du jeu de données de test. Ensuite, calculer le ROUGE Score ainsi que le NEHR de chaque résumé $s \in S_d$.

2. https://huggingface.co/facebook/bart-large-cnn
3. https://huggingface.co/facebook/bart-large-xsum

2. Pour chaque document $d \in C$, sélectionner $\hat{s_d}$:

$$\hat{s_d} = \underset{s \in S_d}{\operatorname{argmax}} ROUGE(s, s_{ref})$$

3. Exécuter le système de reconnaissance automatique d'entités sur chaque résumé $\hat{s_d}$ pour y extraire les entités nommées.
4. Annoter manuellement chaque entité e détectée dans le résumé $\hat{s_d}$ suivant deux dimensions : d'abord, dans le document (in) ou en dehors du document (out) puis bien ou mal utilisé (utilisation correct/incorrect) de e dans $\hat{s_d}$.

Cette étude nous permettra de savoir si le critère du risque est corrélé ou pas avec la fidélité du résumé et si le pourcentage d'entités incorrectes est plus élevé pour les entités en dehors du document source que pour les entités dans le document source. Nous avons sélectionné parmi les 77 résumés générés dans la section 3 les résumés qui ont un ROUGE score (Lin, 2004) maximal avec la référence pour ne pas être dépendant du système de résumé automatique et ainsi d'avoir un meilleur ROUGE score que tous les systèmes de l'état de l'art actuel.

Pour l'extraction des entités nommées, nous avons utilisé le système FLERT [4] (Schweter & Akbik, 2020) qui a été entraîné sur OntoNotes, un large corpus pour la tâche qui a 18 tags différents. Les 18 tags sont : nombre cardinal, date, événement, installation, entité géopolitique, langue, loi, lieu, argent, groupes, nombre ordinal, organisation, pourcentage, personne, produit, quantité, temps et œuvre d'art.

	summary	ROUGE	NEHR
CNN	ROUGE max	57.45 / 32.59 /41.63	4.6
	ROUGE min	30.04 / 09.33 /19.47	6.0
	Logit	41.99 / 18.96 / 28.01	5.6
XSUM	ROUGE max	60.14 / 35.68 / 51.20	45.91
	ROUGE min	27.43 / 07.51 / 21.46	47.39
	Logit	40.26 / 16.79 / 31.29	47.55

TABLE 1 – ROUGE (R-1/R-2/R-L) et NEHR sur les différents résumés produits par notre méthode d'échantillonnage. ROUGE Max, ROUGE min et Logit correspondent respectivement à la sélection du résumé avec le score ROUGE maximal par rapport à la référence la référence, du résumé avec le score ROUGE minimal avec la référence et du résumé ayant le logit maximal produit par le modèle.

Le tableau 1 montre les résultats obtenus en calculant le ROUGE Score ainsi que le critère NEHR sur CNN/DM et XSum. On voit dans le tableau 1 une variation considérable du ROUGE Score qui ne traduit pas une variation du critère NEHR. Cela peut s'expliquer par le fait que le ROUGE score n'est pas un bon indicateur de la fidélité du résumé par rapport au document source. Par contre on a une variation légère du NEHR. Ainsi, nous avons décidé d'annoter les résumés sélectionnés qui ont un ROUGE score maximal par rapport à la référence afin de savoir si le critère NEHR est un bon indicateur de la fidélité du résumé par rapport au document source.

Annotation des entités du meilleur résumé en terme de ROUGE score. Nous avons effectué une évaluation humaine pour savoir si les entités détectées comme risquées par le système d'entités nommées le sont vraiment et si les entités sont utilisées dans le bon contexte ou non.

4. https://huggingface.co/flair/ner-english-ontonotes

Ainsi, nous avons choisi de manière arbitraire 50 exemples du jeu de test de CNN/DM pour les annotations. Pour chaque résumé, nous avons choisi de manière aléatoire le même nombre d'entités hors du document que d'entités dans le document pour l'évaluation manuelle. Nous avons obtenu 145 entités hors du document source et 145 entités dans le document source soit un total de 290 entités à annoter.

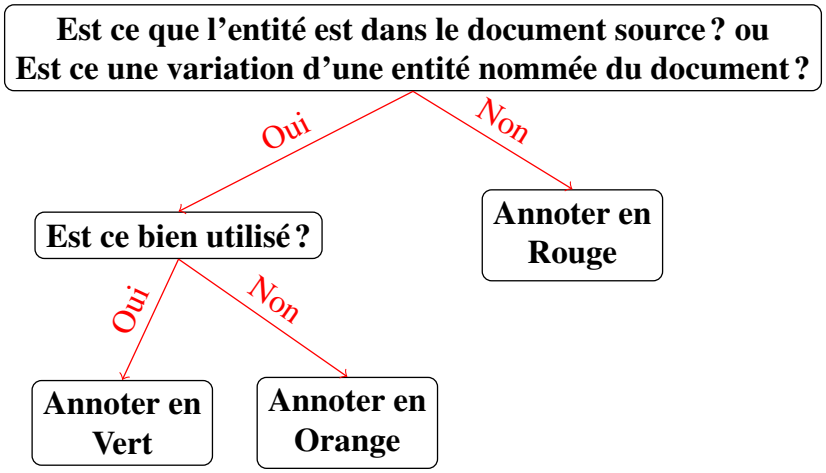

FIGURE 2 – Phase d'annotation des entités nommées. Vert : Les entités nommées qui sont dans le document ou des variations d'entités dans le document qui sont bien utilisées (ne contredisent pas le document) ; Orange : Les entités qui sont dans le document ou des variations d'entités du document qui sont mal utilisé ; Rouge : Les entités hors du document source.

Nous avons ensuite demandé aux annotateurs d'annoter les entités comme étant correctement ou mal utilisées lorsqu'elles sont dans le document source sinon si elles sont en dehors du document source, il s'agissait de vérifier qu'elles ne soient pas des variations d'entités dans le document. En présence de variation, il fallait annoter leur utilisation dans le résumé. Les variations sont les entités du document source qui ont des erreurs de frappe dans le résumé ou encore des dates écrites d'une autre manière. Un exemple de variation peut être *England* pour *Britain* ou encore *trentre trois* pour *33* ou *Balloteli* à la place de *Balotelli*... On definit une entité nommée du résumé comme étant bien utilisée si son emploi ne contredit pas le document source. La figure 2 nous présente la procédure d'annotation qui a été donnée à chaque annotateur. En reprenant l'exemple de la figure 1, on peut annoter les entités nommées du résumé en suivant la procédure de la figure 2. On obtient ainsi cette annotation sur la partie gauche de la figure 3. En vert, nous avons les entités qui ne contredisent pas le document source comme « Mark H Durkan » et « BBC ». En effet, c'est BBC qui a rapporté la news et c'est contre Mark H Durkan que le contrôle judiciaire est demandé mais ce n'est pas Helen Jones qui a demandé ce contrôle. Ainsi, l'entité « Helen Jones » est mal utilisée dans le résumé ; elle est donc mise en orange. On met l'entité « the Northern Ireland's environment minister » en rouge car elle ne provient pas du document. En effet, il n'est pas marqué dans le document qu'il s'agit du ministre de l'environnement ou pas.

Pour une évaluation plus poussée, nous avons décidé d'annoter les 145 entités hors du document pour évaluer leur factualité. Ainsi, les entités ayant été marquées comme hors du document source ont été annotées comme *variation*, *acceptable* ou *hallucination*. *Variation* correspond aux entités détectées comme hors du document mais qui sont en réalité des variation d'entités du document, *acceptable* correspond aux entités qui sont hors du document, mais factuellement correctes et *hallucination* aux entités qui ne sont pas factuellement correctes. Si l'on reprend l'exemple précédent de la figure 3,

FIGURE 3 – Exemple d'annotation des entités nommées suivant le guide de la figure 2.

l'entité **the Northern Ireland's environment minister** n'est pas dans le document source. Mais en se basant sur des connaissances générales, on sait que Mark H Durkan était le ministre de l'environnement de l'Irlande du Nord. Ainsi, on peut la marquer comme acceptable.

Cela nous permet d'avoir deux étages d'annotation, le premier pour savoir le type d'entités (si une entité est dans ou hors du document source) et le second pour savoir si elle est bien utilisée ou non. Nous avons collecté les annotations manuelles de 3 annotateurs. De plus, nous avons calculé l'accord inter-annotateur pour les annotations. Le tableau 2 présente le coefficient Cohen Kappa entre les 3 annotateurs. Le coefficient étant supérieur à 0.63 dans les 3 cas, cela est un indicateur d'un accord important entre les annotateurs.

A1 - A2	A1 - A3	A2 - A3
0.6833	0.6468	0.6302

TABLE 2 – Accord inter-annotateurs Cohen kappa (Cohen, 1960) entre les 3 annotateurs.

Les résultats obtenus pour un annotateur choisi aléatoirement sont consignés dans les tableaux 3a et 3b. Pour les entités hors du document, nous avons permis aux annotateurs de se référer à leur connaissance générale, mais également à Internet pour le second étage d'annotation.

Le tableau 3a montre que pour les entités en dehors du document, 90% sont bien utilisées (ne contredisent pas le document source). Cela signifie que quand une entité du résumé généré est dans le document source, la plupart du temps, elle est bien utilisée. Pour ce qui est des entités en dehors du document source, nous avons 71% qui sont bien utilisées. Le tableau 3b contient la distribution des entités hors du document selon les 3 labels présentés plus haut (*variation, acceptable et hallucination*). Parmi les entités hors du document, 59% sont des variations d'entités du document, 17% sont acceptables (inférables) et 29% sont des hallucinations (n'étant pas dans le document source). Pour ce qui est des variations et des entités acceptables, environ 90% sont bien utilisées dans les résumés (ligne %correct). Ce qui est aligné aux résultats obtenus quand une entité est dans le document source. Ainsi, la proportion qui vient des hallucinations est de 30%. Donc, en minimisant le nombre d'entités en dehors du document source, nous réduisons le risque d'hallucination sur les entités nommées et cela permet d'augmenter la fidélité du résumé par rapport au document source.

Dans la section suivante, nous avons utilisé le critère du risque pour sélectionner des résumés. Cela nous permettra de voir l'impact du critère sur la qualité des résumés.

	%correct
in-doc	90.3
out-doc	71.0

(a) % d'entités correctement utilisées dans les résumés générés selon l'annotation manuelle.

	variation	accept.	hallucination
distribution	59.3%	11.7%	29%
%correct	90	88.0	0

(b) Distribution des entités hors du document selon les 3 partitions avec le pourcentage d'entités correctement utilisées pour chaque ensemble.

TABLE 3 – Résultats des annotations des entités nommées du résumé avec le score ROUGE maximal avec la référence.

4.4 Sélection de résumé en utilisant le NEHR

L'idée est d'évaluer l'impact de l'utilisation du critère NEHR pour la sélection de résumés parmi plusieurs résumés pendant l'inférence. Ainsi, dans cette première partie, nous présentons le processus de sélection de résumés automatiques basé sur le NEHR mis au point. Pour évaluer l'efficacité de ce critère, nous avons évalué les résumés sélectionnés à l'aide des mesures d'évaluation habituelles pour le résumé telles que le ROUGE Score et le BERT Score (Lin, 2004; Zhang* *et al.*, 2020) mais également en suivant la valeur du NEHR (équation 1).

Description Les résumés sélectionnés en utilisant le critère NEHR ont été comparés avec deux baselines (le résumé avec le plus grand logit et le meilleur résumé avec un beam de taille 4). Cette comparaison a été faite en termes de ROUGE score et BERT score.

Pour la sélection de résumés, nous avons proposé un critère basé sur le critère NEHR mais également sur la performance du modèle. D'abord, nous sélectionnons les résumés qui ont un NEHR minimal. Puis, le résumé sélectionné est celui qui a le plus grand logit. Supposons H l'ensemble des résumés obtenus par échantillonnage du modèle, V l'ensemble des résumés avec un risque minimal, $P(\cdot|model)$ est la probabilité donnée par le modèle a un résumé et $\hat{s}$ est le résumé sélectionné :

$$V = \left\{ x \in H | risk(x) = \min_{s' \in H} NEHR(s') \right\} \tag{2}$$

$$\hat{s} = \underset{s \in V}{\operatorname{argmax}} P(s|model) \tag{3}$$

Mésure automatique Nous avons utilisé le ROUGE score et le BERTScore (Lin, 2004; Zhang* *et al.*, 2020) pour évaluer le résumé. Les résultats sont consignés dans le tableau 4. Ce tableau montre que BART-Large avec un *beam = 4* donne le meilleur résultat selon les différentes métriques aussi bien pour CNN/DM que pour XSum. Notre approche minimise bien le risque d'hallucination en ayant des scores ROUGE et BERTScore équivalents aux autres approches pour CNN/DM. Cependant pour XSum, on perd en ROUGE Score.

Le NEHR dépend de la capacité à retrouver les entités dans le document source. Ainsi, il est donc important de détecter efficacement les entités nommées qui peuvent être des variations d'entités nommées du document. La section suivante présente les heuristiques introduites pour la détection de variation.

		R-1	R-2	R-L	BERTScore	NEHR
CNN	BEAM 4	**43.74**	**20.84**	**30.44**	**32.00 / 88.52**	1.53
	BEST LOGIT	41.99	18.96	28.01	29.89 / 88.17	5.60
	MIN NEHR (ours)	42.31	19.21	28.41	30.36 / 88.25	**0.02**
XSUM	BEAM 4	**45.32**	**22.20**	**37.10**	**51.56 / 91.82**	39.84
	BEST LOGIT	40.26	16.79	31.29	45.23 / 90.76	47.55
	MIN NEHR (OUR)	40.05	16.41	31.33	45.79 / 90.85	**18.78**

TABLE 4 – Evaluation des résumés de CNN/DM et XSum en terme de (R-1, R-2, R-L), BERTScore et NEHR. Les deux valeurs du BERTScore correspondent au score avec et sans le paramètre rescale. *NEHR* est le pourcentage d'entités en dehors du document source. BEAM 4 correspond à la génération avec beam égale à 4. BEST LOGIT correspond à la sélection des résumés avec le meilleur logit. Et, MIN NEHR est notre méthode de sélection de résumé décrite dans la section 4.4.

5 Détection de variation

59% des entités considérées comme hors du document étant des variations d'entités dans le document source (voir tableau 3b). Nous avons mis en place une détection automatique de variations afin d'avoir une meilleure précision sur les entités qui sont hors du document.

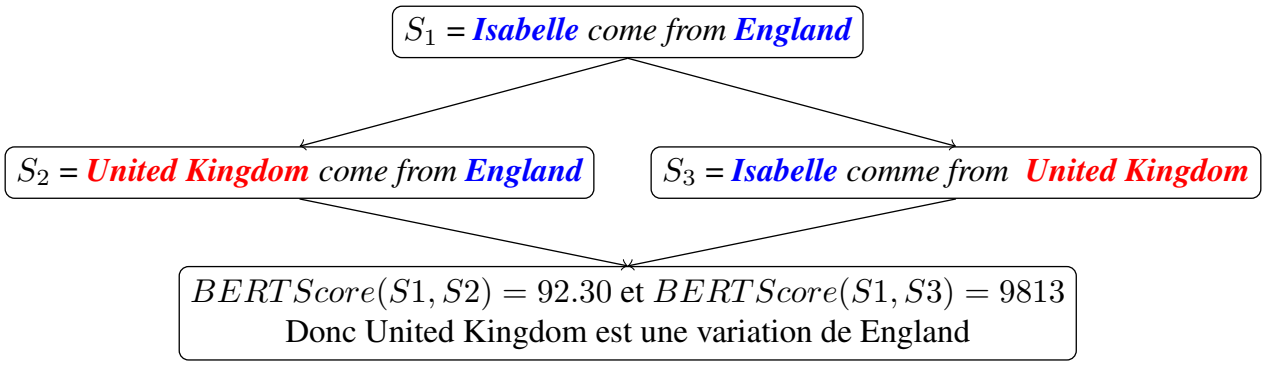

FIGURE 4 – BERTScore pour la détection de variation

Heuristiques de détection de variation Nous avons introduit plusieurs heuristiques pour la détection des variations :

— *Processing* : Consiste à supprimer dans le texte les caractères indésirables. Dans le cas d'un corpus anglais vérifié si ce n'est pas l'indicatif de la possession qui est gênant.
— *Stemming* : Pour la transformation des flexions en leur radical ou racine.
— *Distance de levenshtein* [5] : Elle permet la correction orthographique des mots. Pour un mot et un texte donné, il s'agit de retrouver dans le texte, des mots qui sont proches du mot en entrée lorsqu'on applique une suppression de caractère, un remplacement, une transposition ou encore une insertion. Ainsi, on regarde la distance entre les mots.
— *Alpha2Digit* [6] : On utilise un package python qui convertit les nombres écrits en lettre dans un texte en chiffre.

5. https://norvig.com/spell-correct.html
6. https://pypi.org/project/text2num/

— *SplitWord* : L'idée est de traiter les entités qui viennent d'un groupe de mots. De cette façon, on regarde de manière individuelle si chaque mot du groupe appartient au document source.
— *BERTScoreCheck* : Le BERTScore (Zhang* *et al.*, 2020) étant basé sur des représentations contextuelles, l'idée est de retrouver une entité du document source proche de l'entité qu'on veut tester. Le BERTScore étant une mesure entre des phrases, nous avons mis au point un moyen de l'utiliser pour la détection de variation. La figure 4 donne un exemple. Supposons que la phrase $S1$ provient du document, et « United Kingdom » une variation d'une entité du document. Ici, on la remplace par chaque entité pour créer de nouvelles phrases puis on calcule le BERTScore des phrases obtenues vis-à-vis de la phrase de départ $S1$. Le score le plus élevé entre $S1$ et $S2$ revient à dire que United Kingdom est une variation de l'entité « England ».

Pour créer les règles de détection de variation, nous avons conçu un corpus de variation à partir de CNN/DM et avons sélectionné 91 variations. Le tableau 5 montre le pourcentage de variations détectées par chaque heuristique puis par la combinaison de différentes heuristiques. Cette combinaison se fait en utilisant successivement chaque heuristique pour la détection des variations dans l'ordre suivant : Processing Processing + Stemming + Alpha2digit > Levenshtein > SplitWord > BertScoreCheck.

Méthodes	% **de variations détectées**
Processing	14
Levenshtein	24
SplitWord	30
BertScoreCheck	**39**
Processing + Stemming + Alpha2digit	44
Levenshtein + SplitWord	42
BertScoreCheck + Processing + Stemming + Alphadigit	53
Levenshtein + SplitWord + Processing + Stemming	**59**
All method combined	**60**

TABLE 5 – Pourcentage de variations détectées pour chaque heuristique.

Nous avons ensuite créé un autre corpus contenant 43 variations et 26 entités hors du document source pour savoir si les heuristiques ne détectent pas des entités comme des variations pourtant en réalité elles n'appartiennent pas au document source. On constate que, sans l'utilisation du BERT score (5), on est à 59% du pourcentage d'entités détectées pourtant avec nous sommes à 60%. Pour vérifier ce résultat, nous avons utilisé la combinaison de méthodes pour détecter les entités dans ou en dehors du document source avec ou sans l'utilisation de l'heuristique BERTScoreCheck.

Le tableau 6 présente les résultats. On peut voir que l'utilisation du BERTScore introduit des erreurs dans la détection car certaines entités qui sont hors du document ont été qualifiées comme étant dans le document (voir le score de précision sur les variations). Pour le coût de la méthode, le gain obtenu n'est pas celui attendu. Ainsi, nous avons utilisé uniquement la combinaison des autres heuristiques basées sur les règles.

Ayant obtenu un système pour la détection de variations, nous avons sélectionné à nouveau les résumés générés par la méthode de sampling avec le NEHR comme critère de sélection en prenant en compte le fait que les variations d'entité dans le document sont des entités du document. Le tableau 7 présente les résultats obtenus. Aux différentes méthodes présentées dans le tableau 4, nous avons

	Sans BertScore			Avec BertScore		
	précision	rappel	f1 score	précision	rappel	f1 score
Variation	**1.00**	0.60	**0.75**	**0.90**	0.65	**0.76**
Hors du document	0.60	**1.00**	**0.75**	0.61	**0.88**	0.72

TABLE 6 – Précision, Rappel et F1 score pour la détection de variation et d'entités hors document

également utilisé une méthode basée sur de l'entailment. C'est-à-dire le résumé sélectionné est celui qui à la similarité la plus élevée avec le document source.

		R-1	R-2	R-L	NEHR	%HallDoc
CNN/DM	BEAM 4	**43.74**	**20.84**	**30.44**	0.5	3.86
	BEST LOGIT	41.99	18.96	28.01	2.6	20.57
	ENTAILMENT	43.61	19.69	29.26	1.62	12.92
	MIN NEHR + VAR	42.19	19.12	28.24	**0.003**	**0.035**
XSUM	BEAM 4	**45.32**	**22.20**	**37.10**	27.67	52.48
	BEST LOGIT	40.26	16.79	31.29	31.05	61.24
	ENTAILMENT	40.92	17.14	31.96	27.08	54.98
	MIN NEHR + VAR	40.16	16.54	31.31	**6.92**	**21.49**

TABLE 7 – Évaluation du résumé généré sur les jeux de données CNN/DM et XSum. ROUGE (R-1, R-2, R-L) pour les différents critères de sélection. *NEHR* est le pourcentage d'entités en dehors du document source. Pour BEAM 4, BEST LOGIT, ENTAILMENT la valeur de *NEHR* est avant et après détection de variation BEST LOGIT correspond au moment où nous sélectionnons le meilleur logit de toutes synthèses générées. Et MIN NEHR + VAR est notre méthode de sélection de résumé proposée après l'utilisation de la détection de variation . %HallDoc correspond au pourcentage d'exemples avec au moins une entité hors du document.

Bien que les heuristiques de détection de variations ont été testées sur CNN/DM, on voit l'efficacité sur XSUM en comparant les résultats obtenus dans le tableau 4 aux résultats du tableau 7. Pour les deux datasets, MIN NEHR + VAR donne de meilleurs résultats et réduit bien la quantité d'entités hors du document source. Au niveau du ROUGE, nous avons des scores comparables pour CNN/DM mais un écart en ce qui concerne XSUM. Pour comprendre l'impact du critère NEHR sur les différents datasets, nous avons calculé le pourcentage d'exemples avec au moins une entité hors du document (colonne %HallDoc). Pour XSUM on divise de plus de 2 fois le nombre d'exemples avec des entités hors du document. De même, nous avons calculé le pourcentage d'exemples avec au moins une entité hors du document pour les résumés de référence de chaque corpus afin de comprendre pourquoi le pourcentage d'exemple avec au moins une entité hors du document est élevé. Les résultats sont consignés dans le tableau 8. Ces résultats nous montrent que plus de la moitié des références du corpus XSum ont au moins une entité hors du document soit 2 fois plus que le corpus CNN/DM. Cela montre l'abstractivité du corpus XSum.

Les résultats obtenus dans cette partie montrent que les systèmes tentent de reproduire le caractère abstrait du corpus XSum qui a 63% des exemples du jeu de test qui ont une entité en dehors du document. On constate qu'en utilisant notre critère de sélection de résumé, ce pourcentage est réduit à 21% ce qui peut augmenter la fidélité des entités nommées au regard du document source.

Dataset	% HallDoc	#lines
Cnn/DM	29.33	11490
Xsum	63.65	11333

TABLE 8 – Le pourcentage de résumé de référence avec au moins une entité hors du document source

6 Conclusion et Discussion

En somme, nous avons proposé d'utiliser la méthode de sampling pour générer des résumés couvrant un large espace de recherche afin de sélectionner des résumés ayant le moins d'hallucination au niveau des entités. Cette problématique est très importante du fait que pour une utilisation industrielle, il est important que les entités nommées ne soient pas hallucinées. Notre étude empirique du risque nous a montré que les entités dans le document sont à 90% du temps bien utilisées. Cette étude nous a permis de comprendre que plusieurs entités taguées comme hors du document sont à 59% des variations, et de nous pencher sur la détection de variations au moyen d'heuristique. L'utilisation du critère NEHR combiné aux prédictions du modèle pour la sélection du résumé réduit le risque d'hallucination sur les entités nommées.

Limites Cette étude présente plusieurs limites notamment le fait que le NEHR soit dépendant de la qualité du système de reconnaissance automatique d'entités nommées. Nous avons fait l'annotation humaine des entités uniquement sur 50 paires de données de résumé-document et quelques entités sélectionnées dans le résumé. Les résumés obtenus par l'utilisation du critère NEHR combiné à la prédiction du modèle n'ont pas été évalués manuellement. Il peut être intéressant d'évaluer les différentes sorties.

Remerciements

Remerciement à Frédéric Bechet et Benoit Favre pour l'encadrement et Romain Gemignani, responsable innovation à Enedis.

Ces travaux ont bénéficié d'un accès aux moyens de calcul de l'IDRIS au travers de l'allocation de ressources 2021-AD011012525R2 attribuée par GENCI.

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 (abstraction or hallucination ? status and risk assessment for sequence-to-sequence automatic). In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, p. 2–11, Avignon, France : ATALA.

BANERJEE S. & LAVIE A. (2005). 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, p. 65–72, Ann Arbor, Michigan : Association for Computational Linguistics.

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

CHEN S., ZHANG F., SONE K. & ROTH D. (2021). Improving faithfulness in abstractive summarization with contrast candidate generation and selection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 5935–5941, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.475.

COHEN J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**(1), 37–46. DOI : 10.1177/001316446002000104.

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.

FALKE T., RIBEIRO L. F. R., UTAMA P. A., DAGAN I. & GUREVYCH I. (2019). Ranking generated summaries by correctness : An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 2214–2220, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1213.

FAN A., GRANGIER D. & AULI M. (2018a). Controllable abstractive summarization. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, p. 45–54, Melbourne, Australia : Association for Computational Linguistics. DOI : 10.18653/v1/W18-2706.

FAN A., LEWIS M. & DAUPHIN Y. (2018b). Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 889–898, Melbourne, Australia : Association for Computational Linguistics. DOI : 10.18653/v1/P18-1082.

HERMANN K. M., KOČISKÝ T., GREFENSTETTE E., ESPEHOLT L., KAY W., SULEYMAN M. & BLUNSOM P. (2015). Teaching machines to read and comprehend. DOI : 10.48550/ARXIV.1506.03340.

HOLTZMAN A., BUYS J., DU L., FORBES M. & CHOI Y. (2019). The curious case of neural text degeneration. DOI : 10.48550/ARXIV.1904.09751.

JI Z., LEE N., FRIESKE R., YU T., SU D., XU Y., ISHII E., BANG Y. J., MADOTTO A. & FUNG P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, **55**(12), 1–38. DOI : 10.1145/3571730.

KRYSCINSKI W., KESKAR N. S., MCCANN B., XIONG C. & SOCHER R. (2019). Neural text summarization : A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 540–551, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1051.

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.

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.

NALLAPATI R., ZHOU B., DOS SANTOS C., GULCEHRE C. & XIANG B. (2016). Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, p. 280–290, Berlin, Germany : Association for Computational Linguistics. DOI : 10.18653/v1/K16-1028.

NAN F., NALLAPATI R., WANG Z., NOGUEIRA DOS SANTOS C., ZHU H., ZHANG D., MCKEOWN K. & XIANG B. (2021). Entity-level factual consistency of abstractive text summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume*, p. 2727–2733, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.eacl-main.235.

NARAYAN S., COHEN S. B. & LAPATA M. (2018). Don't give me the details, just the summary ! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 1797–1807, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/D18-1206.

NARAYAN S., ZHAO Y., MAYNEZ J., SIMÕES G., NIKOLAEV V. & MCDONALD R. (2021). Planning with learned entity prompts for abstractive summarization. *Transactions of the Association for Computational Linguistics*, **9**, 1475–1492. DOI : 10.1162/tacl_a_00438.

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.

SCHWETER S. & AKBIK A. (2020). Flert : Document-level features for named entity recognition.

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

WANG A., CHO K. & LEWIS M. (2020). Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 5008–5020, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.450.

WOLF T., DEBUT L., SANH V., CHAUMOND J., DELANGUE C., MOI A., CISTAC P., RAULT T., LOUF R., FUNTOWICZ M., DAVISON J., SHLEIFER S., VON PLATEN P., MA C., JERNITE Y., PLU J., XU C., LE SCAO T., GUGGER S., DRAME M., LHOEST Q. & RUSH A. (2020). Transformers : State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 38–45, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-demos.6.

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