Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

Étude critique du corpus CNN/DailyMail pour le résumé automatique

Fanny Bachey [1], Christophe Rodrigues[2], Aurélien Bossard[1]
(1) LIASD, EA ..., Saint-Denis, France
(2) ESILV
a.bossard@iut.univ-paris8.fr, f.bachey@iut.univ-pariS8.fr,
christophe.rodrigues@devinci.fr

RÉSUMÉ

De nombreux modèles de génération et d'évaluation sont entraînés sur des corpus sans qu'il ait été démontré qu'ils étaient appropriés pour cette tâche. C'est pourquoi nous proposons l'étude critique des données de l'un des corpus les plus utilisés dans le domaine du résumé automatique : CNN/DailyMail. Nous montrons, par une analyse théorique, puis en comparant les résumés de référence du corpus et à des résumés écrits par des humains, que les résumés de référence de CNN/DailyMail ne correspondent pas à ce que doit être un résumé, et que le corpus n'est donc pas adapté à la tâche de résumé automatique.

ABSTRACT

Critical Study of CNN/DailyMail Corpus for Automatic Summarization

Many generation and evaluation models are trained on data without proof that they are suitable for the task. We thus present a critical study of one of the most represented datasets in the summarization's state of art : CNN/DailMail. We show, by a theoretical analysis, and by comparing CNN/DailyMail reference summaries with human-written summaries, that CNN/DailyMail reference summaries do not match the requirements of summaries. So CNN/DailyMail is not suitable to the automatic summarization task.

MOTS-CLÉS : Résumé automatique - CNN/DailMail - Qualité des données - Critique.

KEYWORDS: Summarization - CNN/DailyMail - Data quality - Review.

1 Introduction

Les systèmes de résumé automatique fondés sur l'apprentissage profond nécessitent des données massives pour leur entraînement. Ce besoin en quantité est tel qu'aucun corpus de résumé automatique datant d'avant la révolution de l'apprentissage profond n'est utilisable dans ce contexte. Des solutions ont vu le jour pour palier ce problème : le scrapping pour les corpus CNN/DailyMail (Hermann *et al.*, 2015; Nallapati *et al.*, 2016a) et Gigaword (David Graff, 2003; Napoles *et al.*, 2012), ou encore utilisation de données libres semi-structurées (Chen *et al.*, 2020; Kim *et al.*, 2019). Ces corpus sont utilisés dans la très grande majorité des systèmes de résumé, pour l'apprentissage et pour l'évaluation ; il est donc nécessaire d'en réaliser une analyse critique, ce que nous nous proposons de faire dans ce papier pour le corpus de résumé automatique le plus utilisé aujourd'hui : CNN/DailyMail. Dans un

Actes des 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : articles scientifiques originaux, pages 348–359.

premier temps, nous décrivons les différents corpus existants et présentons les analyses critiques déjà menées sur ces corpus. Dans un second temps, nous présentons une première analyse critique, fondée sur des aspects définitionnels du résumé automatique et sur la construction du corpus CNN/DailyMail. Nous décrivons ensuite une analyse linguistique de ce même corpus, menée sur un nombre restreint d'exemples, avant de proposer une analyse statistique fondée sur la comparaison de résumés écrits par des humains avec les résumés de référence du corpus.

2 État de l'art

Les corpus de résumé automatique utilisables par des modèles d'apprentissage profond se sont considérablement développés ces dernières années. Parmi ceux-ci, on peut citer CNN/DailyMail (Nallapati *et al.*, 2016b), XSum (Narayan *et al.*, 2018), un corpus de nouvelles dédié au résumé « extrême », Gigaword et Newsroom (Grusky *et al.*, 2018), deux corpus de presse, arXiv (Cohan *et al.*, 2018) et PubMed (Sen *et al.*, 2008), deux corpus de résumé d'articles scientifiques et reddit-tifu (Kim *et al.*, 2019), un corpus de résumé de posts de réseau social. À noter que les trois derniers corpus cités sont très intéressants, puisqu'ils respectent des directives de création spécifiques ; leurs résumés sont donc rédigés selon des normes tacites bien définies. Cette caractéristique en fait des corpus particulièrement précieux, car ils garantissent une certaine cohérence et qualité dans les résumés produits, ce qui est essentiel pour une évaluation fiable et représentative du résumé automatique.

Selon le site paperwithcode [1], le corpus CNN/DailyMail est le corpus de résumé automatique le plus utilisé dans les articles scientifiques des 5 dernières années, comme en témoigne la Figure 1 qui présente les statistiques d'utilisation des corpus. Cependant, ce corpus est conçu à l'origine pour une toute autre tâche : la compréhension de la lecture. Les résumés de référence, qui accompagnent chaque article de presse, sont composés des puces (*bullet points*) de l'article.

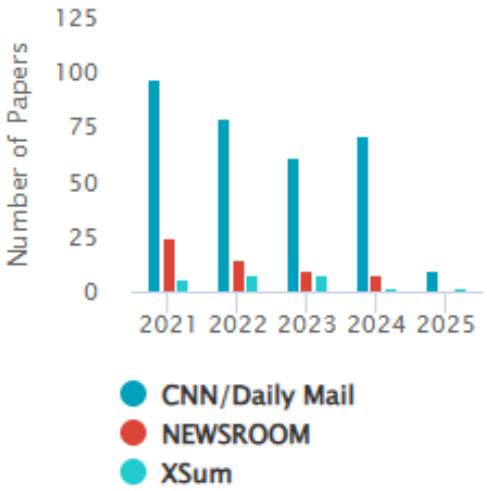

FIGURE 1 – Utilisation des corpus de résumé automatique depuis 2021. Source : paperswithcode.com

C'est donc tout naturellement qu'ont émergé des critiques au sujet de ce corpus, notamment par Gehrmann *et al.* (2023), qui montrent que les *highlights* qui composent les résumés de référence ne sont pas destinés à être utilisés comme des résumés, mais à encourager les lecteurs à lire l'article.

1. https://paperswithcode.com/dataset/cnn-daily-mail-1

Stiennon *et al.* (2020), quant à eux, montrent empiriquement, sur un échantillon de CNN/DailyMail, que les résumés de référence sont jugés en majorité moins bons que des résumés générés par des humains. Fabbri *et al.* (2021) montrent que les résumés de référence de CNN/DailyMail sont jugés par des experts et des non-experts moins cohérents, moins consistents, moins fluides et moins pertinents que des résumés générés par des systèmes de résumé, qu'ils soient abstractifs ou extractifs.

Nous proposons ici une approche différente de ces trois articles, fondée d'une part, sur la définition de résumé, d'autre part sur une analyse linguistique et sur une analyse comparative des résumés de référence de CNN/DailyMail avec des résumés humains afin d'évaluer la qualité du corpus et son utilisabilité sur la tâche de résumé automatique.

3 Critique définitionnelle

Jones (1993) et Minel (2003) ont proposé une typologie des résumés en 3 catégories : résumé indicatif, informatif et synthétique. Dans le cadre de cette analyse critique du corpus CNN/DailyMail, nous traitons des résumés informatifs, qui correspondent le plus au résumé d'actualités. Ainsi, voici une norme qui définit le résumé informatif et dont la définition nous semble coller parfaitement à la tâche.

Cette norme, la norme NF Z 44-004 (décembre 1984) n'est plus en vigueur, remplacée par la norme internationale sur les analyses (Organisation internationale de normalisation (ISO), 1976), mais elle reste pertinente pour mieux comprendre ce qu'est un résumé informatif. C'est *"une représentation abrégée du document, renseignant sur les informations quantitatives ou qualitatives apportées par l'auteur. Ce résumé doit constituer un texte autonome d'une logique rigoureuse. Il forme avec le titre du document un ensemble qui, en principe, ne doit pas être redondant. Les informations retenues pour le résumé sont généralement présentées selon leur ordre d'apparition dans le document. Cet ordre facilite l'exploitation du résumé par le lecteur habitué au plan des articles publiés dans sa spécialité. Généralement, les documents scientifiques et techniques exposent séquentiellement le but de l'étude dans l'introduction, le matériel et les méthodes utilisées, les résultats obtenus, une discussion ou une conclusion évaluant la signification et la pertinence de l'apport. Cependant, en ne négligeant aucune phase du cheminement, les diverses parties du document pourront figurer de façon inégale dans le résumé en fonction de l'importance ou de la nouveauté de l'information."* (Normes françaises, 1984).

Or, comme expliqué en §2, les résumés de référence du corpus CNN/DailyMail sont la concaténation des puces de l'article (que nous nommerons ici *highlights*, initialement conçue pour une tâche de compréhension de la lecture. Cela pose d'emblée la question de la réutilisation d'un corpus conçu pour une toute autre tâche sur le résumé automatique, et des biais que cela induit. L'amoncellement de *highlights* ne peut constituer un « texte autonome d'une logique rigoureuse », puisqu'aucune articulation logique ne peut en être même induite. Au mieux, dans le cas contestable où de tels *highlights* représenteraient de manière stricte les informations essentielles de leur article, un résumé de référence constitué de cette manière pourrait servir à entraîner un système de résumé à sélectionner et dans une moindre mesure à reformuler les informations présentes dans l'article. Se pose alors naturellement une autre question : en quoi l'emploi d'algorithmes très lourds d'apprentissage profond ou de grands modèles de langage peut bénéficier à la tâche de résumé automatique sur un corpus qui ne reflète pas, en tout cas dans sa définition, la tâche de résumé automatique ?

Pour terminer, il n'existe pas, à notre connaissance, de ligne directrice pour la rédaction des *highlights*.

Il est donc peu probable qu'un corpus constitué de cette manière soit utilisable de manière performante pour le résumé automatique.

Nous proposons, dans les sections suivantes, d'évaluer de manière empirique la qualité réelle des résumés de référence du corpus CNN/DailyMail, ainsi que des articles d'origine.

4 Analyse linguistique

Dans cette section, nous effectuons une analyse approfondie d'un article et de ses *highlights*. Cet article a été tiré au hasard du corpus CNN/DailyMail parmi les documents exploitables.

4.0.1 Analyse d'un article issu du corpus et de son highlight

Highlights : Kayahan wrote some of Turkey's best-loved pop songs . The singer was first diagnosed with cancer in 1990 . He most recently performed in February in Istanbul .

Article : (CNN)Kayahan, one of Turkey's best-loved singers and songwriters, died of cancer Friday at the age of 66. He had performed most recently in Istanbul on Valentine's Day. The performer, who was also an accomplished guitarist, was first diagnosed with cancer in 1990, the year he competed in the Eurovision Song Contest, and the year before he released the album that ignited his career. The cancer returned in 2005 and then again in 2014, Turkey's semiofficial Anadolu Agency reported. He died Friday in a hospital in Istanbul, five days after his 66th birthday. "We are in grief over losing Kayahan, who contributed to Turkish music with countless compositions and marked a generation with his songs," Prime Minister Ahmet Davutoglu tweeted. The singer, whose full name was Kayahan Acar, was born in Izmir province, in western Turkey on March 29, 1949. He grew up in Ankara, Turkey's capital, before moving to Istanbul. In 1990, he competed in the Eurovision Song Contest, finishing 17th. The following year he released an album titled "I Made a Vow," which catapulted him to prominence. Though he recorded nearly 20 albums, that one would remain his most popular. His final album was released in 2007. Other artists recorded his material throughout his career. Videos available online show a vibrant performer with a thick shock of dark hair as he accompanies himself on guitar and croons in a clear tenor. Kayahan was best known for his love songs. More recent videos show a frailer performer, seated and without a guitar, but still clearly glorying in the joy of singing a song.

FIGURE 2 – Article extrait du corpus CNN/DM

Un résumé vise à condenser l'information principale en réorganisant et reformulant les idées pour en extraire les éléments les plus pertinents de manière concise. Dans ces *highlights*, les informations sont simplement sélectionnées pour attirer l'attention sur des aspects spécifiques de la vie de Kayahan, mais sans synthétiser ou reformuler le contenu pour le résumé selon les caractéristiques définies §2. L'accent est mis sur des éléments comme sa carrière musicale, sa lutte contre le cancer et ses dernières performances, mais de façon morcelée, sans processus de condensation ou de réécriture.

Le but d'un *highlight* semble plutôt d'attirer le lecteur, notamment car il ne fait que souligner quelques faits, sans indiquer de liens explicites. Il n'y a pas de subordonnées, de relations complexes entre les

éléments. Or, on attend d'un résumé qu'il soit cohérent, c'est-à-dire qu'il contienne des parties qui présentent entre elles des rapports logiques, ce qui n'est pas le cas dans les *highlights*. Les phrases sont déclaratives et permettent une compréhension immédiate mais partielle.

5 Analyse statistique

Dans cette section, nous analysons statistiquement le corpus CNN/DailyMail. Dans un premier temps, nous cherchons à savoir s'il y a une cohérence dans la rédaction des *highlights*, d'après un critère de surface : leur taille. Dans un second temps, nous comparons les résumés de référence du corpus CNN/DailyMail – l'aggrégat des *highlights* – à des résumés réalisés par des humains.

5.1 Taux de compression

Les résumés sont généralement limités en nombre de mots ou en pourcentage de compression du document d'origine. Nous cherchons à savoir si les résumés résultants de l'aggrégation de *highlights* répondent ou non à des lignes directrices en termes de taille ou de taux de compression, soit le pourcentage de réduction du texte par rapport au document d'origine. Dans le cas où cela ne serait pas le cas, l'instabilité du corpus aurait des conséquences néfastes en termes de qualité d'apprentissage et d'inférence, les systèmes apprenant sur des données non cohérentes.

La figure 3 présente la répartition du nombre de mots par résumé de référence au sein du corpus CNN/DailyMail. La figure 4 présente la répartition du taux de compression.

Comme le montre la figure 3, la distribution du nombre de mots par résumé de référence est asymétrique, avec une concentration entre 35 et 70 mots. Cependant, on constate une large variabilité, avec des valeurs allant du simple au double. Cette dispersion pourrait introduire une difficulté pour les modèles à générer des résumés de longueur uniforme.

Le nombre de mots par résumé de référence a une *variance* de 507.01 . Cela signifie que la longueur des résumés de référence varie considérablement autour de la moyenne. En d'autres termes, certains résumés de référence sont bien plus longs ou plus courts que la moyenne de manière assez marquée. La variance élevée indique une grande dispersion des longueurs des résumés de référence. L'écart-type de 22.52 nous montre que la plupart des résumés de référence diffère en moyenne de 22 à 23 mots de la longueur moyenne.

Dans la figure 4, on observe une distribution fortement asymétrique, avec une majorité de résumés de référence ayant un taux de compression compris entre 5 et 10 %, mais une partie non négligeable de valeurs s'étendant jusqu'à 25 %. Ce pourcentage élevé peut s'expliquer par le fait que la longueur des articles est très variable. En effet, il n'y a pas de ligne directrice précise pour la production de ces articles de presse, ni, ce que nous cherchions à montrer ici, pour la production des *highlights*. Cette instabilité du taux de compression complique l'utilisation de ce corpus pour l'entraînement de modèles de résumé, car elle empêche une compression standardisée entre les documents.

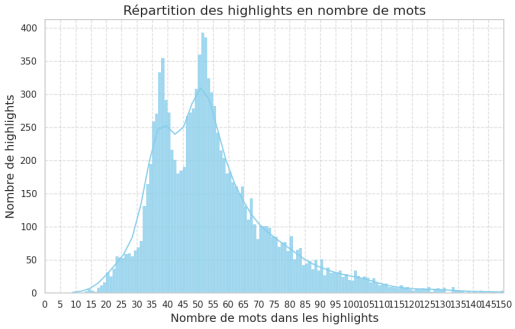

FIGURE 3 – Répartition du nombre de mots par résumé de référence dans le corpus de test de CNN/DM.

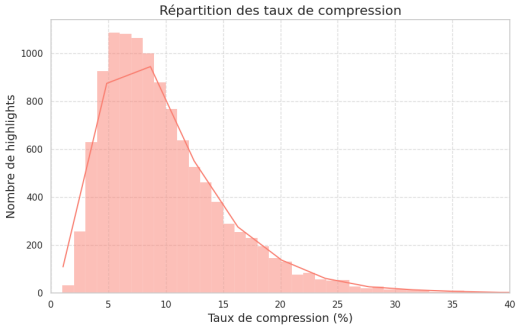

FIGURE 4 – Taux de compression des résumés de référence dans le corpus de test de CNN/DM.

5.2 Résumé manuel vs résumés de référence du corpus

Nous avons étudié, au-delà de la qualité présumée des résumés de référence du corpus (qui sont l'aggrégation des *highlights* d'un article), leur qualité en les comparant à des résumés manuels à l'aide de métriques d'évaluation et d'évaluation manuelle. Nous avons pour cela choisi des échantillons de test, puis fait résumer par des volontaires chacun des articles des échantillons, et comparé les résumés manuels avec les résumés de référence.

Choix des échantillons Le corpus CNN/DailyMail est trop important pour réaliser une analyse manuelle sur l'ensemble de ses paires article/résumé. Nous avons donc restreint l'analyse manuelle à des échantillons de test. Nous avons voulu sélectionner deux types de paires article/résumé : des paires « standard » pour lesquelles la difficulté de la production de résumés est dans la moyenne, et des paires pour lesquelles la production de résumés semble difficile. Pour ce faire, nous avons lancé neuf systèmes de résumés sur la partie test de CNN/DailyMail, parmi lesquels des systèmes extractifs et génératifs et présentés en Tableau 1.

Modèle	**Paradigme**	**Référence**
BART	Génératif	Lewis (2019)
Icsisumm	Extractif	Gillick & Favre (2009)
Gensim TextRank	Extractif	Barrios *et al.* (2016); Mihalcea & Tarau (2004)
Centrum	Extractif	Puduppully *et al.* (2022); Radev *et al.* (2000)
Pegasus	Génératif	Zhang *et al.* (2020)
Genetic	Extractif	Bossard & Rodrigues (2011)
BRIO	Génératif	Liu *et al.* (2022)
LexRank	Extractif	Erkan & Radev (2004)
T5	Génératif	Raffel *et al.* (2020)

TABLE 1 – Systèmes de résumé utilisés

Nous avons ensuite évalué les résumés ainsi générés avec la méthode ROUGE (Lin, 2004) et créé deux échantillons : un en choisissant au hasard des paires article/résumé parmi celles qui présentent un fort écart-type de scores ROUGE donc qui présentent des difficultés pour certains systèmes et

— Lisez attentivement et dans son intégralité le document d'origine pour en comprendre le contenu et la structure principale.
— Identifiez les informations essentielles :
  — informations principales et/ou arguments centraux
  — éléments de contexte nécessaires à la compréhension
  — points clés de la structure du document
— Essayez de respecter au maximum l'ordre d'apparition des informations extraites dans votre restitution.
— Rédigez un résumé concis. Ce résumé devra contenir 50 mots maximum.
— Veillez à éviter les répétitions et les détails non essentiels. Assurez-vous également que le résumé est compréhensible sans le document d'origine.
— N'utilisez pas un vocabulaire trop élaboré afin que le résumé soit lisible et compréhensible par le plus grand nombre.
— Relisez et corrigez votre texte.

FIGURE 5 – Consignes aux annotateurs pour la rédaction de résumés

non pour d'autres, et l'autre parmi celles qui présentent un écart-type de scores ROUGE autour de la moyenne, donc plus représentatif du corpus.

Directives pour l'écriture de résumés Les résumeurs humains ont rédigé leur résumé selon la consigne présentée en Figure 5. Nous avons choisi de demander des résumés de 50 mots, qui correspondent à la moyenne du nombre de mots des résumés de référence.

Attribution des résumés aux annotateurs Nous avons réparti les résumés entre cinq annotateurs, en faisant en sorte d'avoir trois résumés humains par article et de maximiser les croisements entre annotateurs afin de réduire les biais liés à l'attribution de documents.

Mesures de qualité des résumés de référence Les résumés de référence de chacun des deux échantillons sont évalués en comparant leurs scores ROUGE vis-à-vis des résumés humains, et les scores ROUGE des résumés humains entre eux.

Résultats Le Tableau 2 présente la moyenne des scores ROUGE-1, 2 et L entre résumés humains et entre résumés humains et résumés de référence sur l'échantillon 1, soit celui sur lesquels les systèmes de résumé automatiques présentaient la plus grande déviation à la moyenne. Le Tableau 3 présente quant à lui les mêmes données, mais sur l'échantillon 2, soit celui sur lesquels les systèmes de résumé automatique présentaient une déviation à la moyenne standard.

Métrique	Sujet 1	Sujet 2	Moyenne	σ
ROUGE 1	humains	highlights	0.226	0.041
ROUGE 1	humains	humains	0.252	0.037
ROUGE 2	humains	highlights	0.070	0.022
ROUGE 2	humains	humains	0.068	0.024
ROUGE L	humains	highlights	0.226	0.040
ROUGE L	humains	humains	0.209	0.038

TABLE 2 – Moyenne des scores ROUGE2 CNN/DM sur l'échantillon 1

Les scores inter-humains sont beaucoup plus élevés sur que les scores humains/référence sur l'échantillon 2, tout en présentant un écart-type plus faible. En revanche, sur l'échantillon 1, les scores inter-humains sont assez semblables aux scores humains/référence

Métrique	**Sujet 1**	**Sujet 2**	**Moyenne**	σ
ROUGE 1	humains	highlights	0.182	0.051
ROUGE 1	humains	humains	0.227	0.030
ROUGE 2	humains	highlights	0.042	0.027
ROUGE 2	humains	humains	0.066	0.022
ROUGE L	humains	highlights	0.173	0.050
ROUGE L	humains	humains	0.212	0.029

TABLE 3 – Moyenne des scores ROUGE2 CNN/DM sur l'échantillon 2

6 Discussion

La comparaison des résumés humains entre eux et des résumés humains contre les résumés de référence montre que sur l'échantillon 2, celui qui est a priori le plus représentatif du corpus, les résumés humains sont bien plus proches entre eux qu'ils ne le sont des résumés de référence, ce qui tend à montrer que les résumés de référence ne sont pas construits comme un résumé écrit par un humain avec des consignes de rédaction précises.

En revanche, sur l'échantillon 1, celui sur lequel les écarts de score des résumés automatiques sont les plus importants, il n'y a pas de différence notable entre les résumés humains et les résumés de référence. Cela peut être dû à la difficulté de la tâche sur cet échantillon, en raison de la façon dont il a été construit. En effet, les résumeurs humains nous ont fait part de plusieurs problèmes :

— Contenu insuffisant dans le document d'origine ;
— Contenu trop hétérogène dans le document d'origine ;
— Contenu sans faits dans le document d'origine.

Ces retours d'expérience soulignent un problème que nous n'avions pas envisagé au début de l'étude : les documents d'origine eux-mêmes peuvent ne pas être propices à la tâche de résumé automatique.

Les différentes analyses menées sur le corpus CNN/DailyMail tendent donc à corroborer les études de Fabbri *et al.* (2021); Stiennon *et al.* (2020) et Gehrmann *et al.* (2023), à savoir que le corpus CNN/DailyMail, bien qu'il soit majoritaire dans les travaux de recherche en résumé automatique, n'est pas un corpus adapté à la tâche.

De plus, la lecture de certains articles et de leur résumé de référence montre, comme l'illustre la Figure 6 que certains résumés de référence comportent des informations qui sont absentes de l'article et non déductible des informations qu'il contient. Si ce problème était trop fréquent, il serait un biais important, que ce soit pour l'apprentissage ou l'évaluation de résumé.

7 Conclusion

Dans cet article, nous avons proposé une analyse critique du corpus de résumé automatique le plus utilisé de nos jours, CNN/DailyMail. Que ce soit l'analyse théorique, fondée sur les définitions

Article : 'You guys ready to play some golf ?' asked the starter. And, yes he was. That's exactly what Tiger Woods was ready for. Some golf. He smiled, slapped his playing partner Rory McIlroy on the back. Damn right he was ready. Not to win maybe. Not yet. By the summer, though, who knows ? There have been enough sightings of the old Tiger at Augusta this week not to give up on him yet. The problem with new Tiger, though, is that as desperate as he appears to march on, obstacles line his path, like tree roots. In some cases, exactly like tree roots. At the ninth, playing out of the Augusta pines, Woods found such a root at full pelt on his follow through. He dropped his club like a red-hot poker, face creased with pain. Woods claimed that a bone popped out in his right arm, right there, but he reset it. There was a degree of scepticism about the extent of that injury, but either way it continued to bother him for several holes and was rotten luck. There is, however, a well-worn method of avoiding such misfortune. Hit it on the fairway. Woods did not find a solitary preferred landing area until the 13th, when he was rewarded with an eagle. Up to that point, he spent so much time in the trees he should have been fitted with a lumberjack shirt. At the tenth he was in so deep they were thinking of sending Lassie for help. Tiger Woods looks over his second shot on the seventh hole as the Augusta crowd look on . Woods (left) speaks to playing partner Rory McIlroy as the pair wait to tee off on the first tee . Woods reacts in pain after hitting a shot out of the pine straw on the ninth hole . Still, it could have been worse. This was the Masters that Tiger was advised to miss. There were some sage voices out there, saying his old friend would turn on him, that he might be embarrassed by his favourite homeland course. His game was broken, he was broken, it was argued. This was no country for an old man with faltering short irons. Some folk wanted their happy memories preserved. These four days, however, have revealed a healthier reality. Woods didn't come back too soon, he just took his leave too late. Looking at the tournaments he played in January and February, it was plain he didn't want to be there. He should have called time out then. Had he done so, had he started his recovery earlier, who knows where his game would have been by the time he reached Augusta ? Were it not for the phenomenon of Jordan Spieth's first 54 holes, he would have been in contention here starting the final day. Woods, like McIlroy, began six under par but with Spieth ten shots ahead, it was idealistic to even contemplate a charge. Woods would have needed a score approaching the course record, and Spieth a collapse – and he is not well-placed for miracles yet. Another major ? Well, you wouldn't bet against that one. Woods tees off on the tenth but in the end he did not manage to challenge runaway winner Jordan Spieth . Woods plays out of the bunker by the sixteenth green at Augusta National . It could be argued that the enthusiasm for Tiger's performance this weekend shows how far he has fallen. After all, two of his four rounds – the first and last – were over par, and in the head to head on Sunday McIlroy beat him by seven shots. This was far from tournament winning golf. Then again McIlroy is the world No 1, Woods 110 places below and has barely played this year. He has reshaped his swing which continues to be a work in progress. He would not say when he will play next after this. 'It won't be for a while,' Woods confirmed. 'I like what I'm doing, so I'm going to go back and work on that.' Where this will leave him for the Majors short-term it is hard to say. He finished in the top 20 here, his best return since 2013 and many rated this as the best of several recent comebacks, one that inspired speculation about his readiness for St Andrews in July. He loves the Old Course almost as much as he loves Augusta.[...]

Résumé de référence : Tiger Woods finishes tied for seventeenth in the 2015 Masters at Augusta. Woods could only score 73 on Sunday, seven shots less than Rory McIlroy . **There has been enough to suggest that Woods is not finished yet**. Jordan Spieth won this year's Masters after finishing on 18 under par.

FIGURE 6 – Paire article(coupé)/résumé de référence issué du corpus CNN/DailyMail. En gras dans le résumé : les informations absentes de l'article

du résumé, l'analyse de surface portant sur la cohérence du corpus ou l'analyse comparative des résumés de référence du corpus avec des résumés humains, tout pointe vers la conclusion suivante, vers laquelle pointaient déjà les études de Fabbri *et al.* (2021); Stiennon *et al.* (2020) et Gehrmann *et al.* (2023) sur ce même corpus : CNN/DailyMail n'est pas un corpus adapté à la tâche de résumé automatique.

De plus, le corpus CNN/DailyMail semble comporter des biais importants liés à la qualité des données des documents d'origine, et qui, s'ils étaient trop fréquents, pourraient nuire considérablement à la fois à l'apprentissage des modèles de résumé et à leur évaluation.

Une analyse manuelle de certains résumés de référence a montré que les résumés de référence du corpus CNN/DailyMail comportaient parfois des informations introuvables et non déductibles de l'article d'origine. Nous souhaitons pousser plus loin l'étude du corpus en abordant ce problème sous l'angle de la détection d'hallucinations.

Références

BARRIOS F., LÓPEZ F., ARGERICH L. & WACHENCHAUZER R. (2016). Variations of the similarity function of textrank for automated summarization. *arXiv preprint arXiv :1602.03606*.

BOSSARD A. & RODRIGUES C. (2011). Combining a multi-document update summarization system–cbseas–with a genetic algorithm. In *Combinations of Intelligent Methods and Applications : Proceedings of the 2nd International Workshop, CIMA 2010, France, October 2010*, p. 71–87 : Springer.

CHEN Y., POLAJNAR T., BATCHELOR C. & TEUFEL S. (2020). A corpus of very short scientific summaries. In R. FERNÁNDEZ & T. LINZEN, Éds., *Proceedings of the 24th Conference on Computational Natural Language Learning*, p. 153–164, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.conll-1.12.

COHAN A., DERNONCOURT F., KIM D. S., BUI T., KIM S., CHANG W. & GOHARIAN N. (2018). A discourse-aware attention model for abstractive summarization of long documents. In M. WALKER, H. JI & A. STENT, Éds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 2 (Short Papers)*, p. 615–621, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-2097.

DAVID GRAFF C. C. (2003). English gigaword. DOI : https ://doi.org/10.35111/0z6y-q265.

ERKAN G. & RADEV D. R. (2004). Lexrank : Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, **22**, 457–479.

FABBRI A. R., KRYŚCIŃSKI W., MCCANN B., XIONG C., SOCHER R. & RADEV D. (2021). Summeval : Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, **9**, 391–409. DOI : 10.1162/tacl_a_00373.

GEHRMANN S., CLARK E. & SELLAM T. (2023). Repairing the cracked foundation : A survey of obstacles in evaluation practices for generated text. *J. Artif. Int. Res.*, **77**. DOI : 10.1613/jair.1.13715.

GILLICK D. & FAVRE B. (2009). A scalable global model for summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing*, p. 10–18.

GRUSKY M., NAAMAN M. & ARTZI Y. (2018). Newsroom : A dataset of 1.3 million summaries with diverse extractive strategies. In M. WALKER, H. JI & A. STENT, Éds., *Proceedings of the*

2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers), p. 708–719, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-1065.

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

JONES K. S. (1993). What might be in a summary ? *Information retrieval*, **93**(1), 9–26.

KIM B., KIM H. & KIM G. (2019). Abstractive summarization of Reddit posts with multi-level memory networks. In J. BURSTEIN, C. DORAN & T. SOLORIO, Éds., *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. 2519–2531, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1260.

LEWIS M. (2019). Bart : Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv :1910.13461*.

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.

LIU Y., LIU P., RADEV D. & NEUBIG G. (2022). Brio : Bringing order to abstractive summarization. *arXiv preprint arXiv :2203.16804*.

MIHALCEA R. & TARAU P. (2004). TextRank : Bringing order into text. In D. LIN & D. WU, Éds., *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, p. 404–411, Barcelona, Spain : Association for Computational Linguistics.

MINEL J.-L. (2003). Filtrage sémantique. du résumé à la fouille de textes. *Hermes, paris*.

NALLAPATI R., ZHOU B., DOS SANTOS C., GU̇LÇEHRE Ç. & XIANG B. (2016a). Abstractive text summarization using sequence-to-sequence RNNs and beyond. In S. RIEZLER & Y. GOLDBERG, Éds., *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.

NALLAPATI R., ZHOU B., DOS SANTOS C., GU̇LÇEHRE Ç. & XIANG B. (2016b). Abstractive text summarization using sequence-to-sequence RNNs and beyond. In S. RIEZLER & Y. GOLDBERG, Éds., *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.

NAPOLES C., GORMLEY M. & VAN DURME B. (2012). Annotated Gigaword. In J. FAN, R. HOFFMAN, A. KALYANPUR, S. RIEDEL, F. SUCHANEK & P. P. TALUKDAR, Éds., *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX)*, p. 95–100, Montréal, Canada : Association for Computational Linguistics.

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 E. RILOFF, D. CHIANG, J. HOCKENMAIER & J. TSUJII, Éds., *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.

ORGANISATION INTERNATIONALE DE NORMALISATION (ISO) (1976). *ISO 214 :1976(fr) Documentation — Analyse pour les publications et la documentation*. ISO.

PUDUPPULLY R., JAIN P., CHEN N. F. & STEEDMAN M. (2022). Multi-document summarization with centroid-based pretraining. *arXiv preprint arXiv :2208.01006*.

RADEV D. R., JING H. & BUDZIKOWSKA M. (2000). Centroid-based summarization of multiple documents : sentence extraction, utility-based evaluation, and user studies. In *NAACL-ANLP 2000 Workshop : Automatic Summarization*.

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.

SEN P., NAMATA G., BILGIC M., GETOOR L., GALLAGHER B. & ELIASSI-RAD T. (2008). Collective classification in network data. *AI Mag.*, **29**(3), 93–106. DOI : 10.1609/aimag.v29i3.2157.

STIENNON N., OUYANG L., WU J., ZIEGLER D. M., LOWE R., VOSS C., RADFORD A., AMODEI D. & CHRISTIANO P. F. (2020). Learning to summarize from human feedback. *CoRR*, **abs/2009.01325**.

ZHANG J., ZHAO Y., SALEH M. & LIU P. (2020). Pegasus : Pre-training with extracted gap-sentences for abstractive summarization. In *International conference on machine learning*, p. 11328–11339 : PMLR.