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

Quelles évolutions sur cette loi ? Entre abstraction et hallucination dans le domaine du résumé de textes juridiques

Nihed Bendahman[1, 2] Karen Pinel-Sauvagnat[1] Gilles Hubert[1]
Mokhtar Boumedyen Billami[2]
(1) IRIT, 118 Route de Narbonne, 31400 Toulouse, France
(2) Berger-Levrault, 64 Rue Jean Rostand, 31670 Labège, France
{nihed.bendahman, karen.sauvagnat, gilles.hubert}@irit.fr
{nihed.bendahman, mb.billami}@berger-levrault.com

RÉSUMÉ

Résumer automatiquement des textes juridiques permettrait aux chargés de veille d'éviter une surcharge informationnelle et de gagner du temps sur une activité particulièrement chronophage. Dans cet article, nous présentons un corpus de textes juridiques en français associés à des résumés de référence produits par des experts, et cherchons à établir quels modèles génératifs de résumé sont les plus intéressants sur ces documents possédant de fortes spécificités métier. Nous étudions quatre modèles de l'état de l'art, que nous commençons à évaluer avec des métriques traditionnelles. Afin de comprendre en détail la capacité des modèles à transcrire les spécificités métiers, nous effectuons une analyse plus fine sur les entités d'intérêt. Nous évaluons notamment la couverture des résumés en termes d'entités, mais aussi l'apparition d'informations non présentes dans les documents d'origine, dites hallucinations. Les premiers résultats montrent que le contrôle des hallucinations est crucial dans les domaines de spécialité, particulièrement le juridique.

ABSTRACT

What are the evolutions of this law ? Between abstraction and hallucination in the field of legal text summarization

Automatically summarizing legal texts would allow monitoring officers to avoid information overload and to save time on a particularly time-consuming activity. In this paper, we present a corpus of French legal texts associated with reference summaries produced by experts. Using this collection, we aim at identifying which generative summarization models are the most interesting on these documents with important specificities. We study four state-of-the-art models, which we begin to evaluate with traditional metrics. In order to understand in detail the ability of this models to transcribe documents specificities, we perform a more detailed analysis on the entities of interest. In particular, we evaluate the coverage of the produced summaries in terms of entities, but also the appearance of information not present in the original documents, called hallucinations. First results show that hallucination control is crucial in specific domains such as the legal one.

MOTS-CLÉS : Résumés abstractifs, Hallucination, Evaluation, Domaine juridique.

KEYWORDS: Abstractive Summarization, Hallucination, Evaluation, Legal domain.

Actes de la 18e Conférence en Recherche d'Information et Applications (CORIA), pages 18–36.

1 Introduction

La veille juridique est une activité cruciale pour les entreprises afin de rester en phase avec l'actualité juridique. Elle leur permet d'être en permanence au fait des réglementations en cours et d'anticiper leurs évolutions à venir, afin de les appliquer au plus tôt. Cependant, avec l'inflation législative permanente, les chargés de veille éprouvent eux-mêmes une surcharge informationnelle qui complique leurs activités. Il devient très difficile pour eux d'analyser des centaines voire des milliers d'articles par jour. La synthèse de l'information qualifiée de pertinente requiert un effort considérable : identifier dans les amas de textes et de médias les éléments informationnels qui ont réellement de l'importance est particulièrement chronophage. Les chargés de veille travaillent ensuite par extraction et par abstraction de l'information pour construire des résumés dans des newsletters qui soient synthétiques et digestes.

La génération automatique de résumé représente donc une solution intéressante pour aider les chargés de veille dans leurs activités de veille juridique. Les approches de résumé peuvent être « extractives » ou « abstractives ». Les approches extractives, comme (Fabbri *et al.*, 2019; Saini *et al.*, 2019; Zhong *et al.*, 2020), retournent des extraits des textes à résumer, tandis que les approches abstractives, comme (Qi *et al.*, 2020; Zhang *et al.*, 2020a; Dou *et al.*, 2021), peuvent formuler de nouvelles phrases. Les approches abstractives visent donc à produire des résumés analogues à ce que produisent les chargés de veille.

À l'instar d'autres tâches liées au Traitement Automatique des Langues (TAL), les approches récentes de résumé automatique utilisent des modèles de langue neuronaux. Ces approches ont été essentiellement appliquées sur des collections d'actualités (« news ») (Dernoncourt *et al.*, 2018; Ma *et al.*, 2022). À notre connaissance, aucune étude de ce type d'approche n'a été réalisée dans le contexte d'informations juridiques, qui plus est en langue française.

Cet article vise donc à étudier dans quelle mesure les modèles de langue peuvent être appliqués à du résumé abstractif dans le cadre d'informations juridiques. Plus précisément, nous cherchons à répondre aux questions de recherche suivantes :

RQ1. Quelle collection de documents juridiques en français peut être utilisée pour une telle étude ?

RQ2. Quelles sont les performances atteignables par les modèles génératifs de l'état de l'art ? Quels sont les modèles qui fonctionnent le mieux ?

RQ3. Dans quelle mesure les métriques traditionnelles permettent-elles d'interpréter correctement les résultats d'évaluation dans un contexte métier ?

Pour répondre à chacune de ces questions, les contributions de cet article sont :

C1. L'identification d'une collection de test sur le juridique en langue française adaptée à l'évaluation de modèles génératifs de langue pour le résumé abstractif,

C2. La comparaison de modèles de l'état de l'art suivant les familles de métriques traditionnellement reportées (El-Kassas *et al.*, 2021; Ermakova *et al.*, 2019), c'est-à-dire ROUGE et BLEU, complétées par une similarité sémantique,

C3. L'utilisation de métriques basées sur les entités et les mots-clés relatifs à un domaine métier, notamment pour évaluer la couverture des résumés et l'apparition d'informations non présentes dans les documents d'origine, dites « hallucinations » (Akani *et al.*, 2022).

L'article est organisé comme suit. La section 2 présente une synthèse des travaux relatifs à la génération automatique de résumé abstractif et son évaluation. La section 3 présente la collection de documents juridiques en langue française permettant d'évaluer les performances des modèles génératifs de l'état de l'art. Dans la section 4, nous présentons le cadre expérimental puis détaillons et analysons les résultats suivant les métriques ROUGE, BLEU et de similarité sémantique . La section 5 détaille ensuite les résultats des évaluations suivant les entités d'intérêt. Enfin, la section 6 conclut l'article et annonce les pistes de travaux futurs.

2 État de l'art

2.1 Résumés abstractifs

Les approches de génération de résumé abstractif ont toujours été au cœur des recherches en traitement automatique des langues, car elles visent à produire des résumés de texte qui sont plus fluides et plus lisibles que les résumés extractifs. Tandis que les résumés extractifs sélectionnent des phrases ou des passages clés du document source, les résumés abstractifs quant à eux reformulent ce dernier de sorte à exprimer son essence.

Les premières approches neuronales de résumé abstractif sont basées sur des réseaux de neurones récurrents (RNN) (Rumelhart *et al.*, 1986) et leurs variantes LSTM (Hochreiter & Schmidhuber, 1997) et bi-LSTM (Huang *et al.*, 2015). Ces approches permettent de produire des résumés de qualité mais sont très limitées pour traiter les textes longs. Cependant, depuis l'émergence des modèles *transformers* pré-entraînés (Devlin *et al.*, 2018; Vaswani *et al.*, 2017) et les architectures séquence à séquence, les résumés abstractifs produits ont connu un progrès significatif en termes de fluidité et de lisibilité. Parmi les architectures séquence à séquence, on retrouve les modèles à base de mécanismes d'attention (Luong *et al.*, 2015) qui pondèrent les passages selon leur importance dans le document source, les réseaux pointeurs (*pointer networks*) (Vinyals *et al.*, 2015; See *et al.*, 2017) qui commencent par extraire les passages du document les plus importants et procèdent à une reformulation du reste du document par la suite, ou encore les modèles de langues pré-entraînés sur la tâche de résumé abstractif tels que Pegasus (Zhang *et al.*, 2020a) ou T5 (Raffel *et al.*, 2020).

Même si ces derniers ont montré des résultats prometteurs, la tâche reste difficile, notamment en ce qui concerne la génération de résumé dans les domaines de spécialité tels que le juridique, le médical ou les sciences (El-Kassas *et al.*, 2021). En effet, ces domaines exigent des connaissances et une terminologie spécifiques qui ne sont pas toujours bien représentées dans les modèles de langues, car ces derniers ont été pré-entraînés sur des corpus de données génériques (souvent des corpus d'actualité). Par conséquent, la production de résumés de bonne qualité dans ces domaines de spécialité, demeure encore aujourd'hui un défi majeur, notamment lorsqu'il s'agit d'autres langues, telles que le français par exemple (Zhou *et al.*, 2022).

2.2 Évaluation automatique des résumés

Pour évaluer les résumés produits par les systèmes, l'évaluation automatique la plus courante se base sur des résumés de référence (aussi appelés « Gold Standard »), généralement produits de façon manuelle. L'idée est que plus les résumés générés sont proches des résumés de référence, meilleurs ils

sont. Il s'agit donc ensuite de calculer de façon automatique une similarité entre les résumés produits et les résumés de référence.

Dans le cadre du résumé abstractif, deux grandes familles de métriques sont traditionnellement reportées (El-Kassas *et al.*, 2021; Ermakova *et al.*, 2019) :

— Les mesures ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) (Lin, 2004) se basent sur le chevauchement des mots (unigrammes, bigrammes et n-grammes) entre les deux textes à comparer. Elles sont orientées rappel : plus le résumé à évaluer contient de n-grammes du résumé de référence, meilleur est le résultat des métriques.

— Les mesures BLEU (*BiLingual Evaluation Understudy*) (Papineni *et al.*, 2002), initialement créées pour l'évaluation de la traduction automatique de texte, se basent également sur le nombre de n-grammes en commun entre les résumés à évaluer et le résumé de référence. Elles mesurent combien de n-grammes du résumé à évaluer apparaissent dans le résumé de référence, et sont donc pour leur part orientées précision.

Ces métriques, bien que simples à mettre en place, sont incapables de considérer des synonymes ou expressions sémantiquement proches. Il est donc habituel de reporter, en complément des métriques ROUGE et BLEU, des métriques basées sur la similarité sémantique des résumés. Parmi elles, nous pouvons citer BertScore (Zhang *et al.*, 2020b) ou la métrique Cos Embed utilisée dans (Dusart *et al.*, 2023), qui calculent une similarité cosinus entre des représentations générées respectivement par BERT ou Word2Vec.

L'utilisateur intéressé pourra se référer à (Ermakova *et al.*, 2019) pour la description de mesures d'évaluations complémentaires, toutes ayant leurs points forts et faibles. On relèvera cependant deux limitations principales à ces métriques :

— Aucune ne pénalise de façon spécifique la présence d'informations incohérentes (au sens où elles ne sont pas présentes dans le documents d'origine), nommées hallucinations par (Maynez *et al.*, 2020). Or, d'après (Cao *et al.*, 2018), 30 % des résumés produits par les méthodes d'état de l'art contiennent ce genre d'incohérences (*fact fabrication*). Des métriques spécifiques ont été proposées dans l'état de l'art, basées sur l'inférence, dans le cadre des systèmes de question-réponses ou encore les systèmes d'extraction d'information (Ji *et al.*, 2022).

— Aucune ne prend en compte de façon explicite les spécificités métiers des documents considérés. La mise en correspondance de n-grammes ou le calcul de similarités sémantiques peut ne pas suffire à interpréter correctement les résultats.

Afin d'analyser plus finement les résultats, nous proposons dans cet article d'utiliser et discuter des métriques complémentaires basées sur les entités d'intérêt, en évaluant leur couverture mais aussi, sur cette base, les hallucinations produites par les méthodes.

2.3 Collections de test dans le domaine juridique

Le domaine juridique est un domaine en constante évolution, centre d'intérêt croissant pour la communauté du Traitement Automatique des Langues. Cette tendance est mise en évidence par des initiatives telles que la campagne d'évaluation SemEval 2023 [1]. Cette édition propose une tâche couvrant diverses applications du TAL telles que l'analyse du discours et la détection d'entités nommées. De façon complémentaire, plusieurs corpus de textes juridiques sont désormais en libre accès, notamment le corpus belge de recherche d'articles statutaires BSARD (Louis & Spanakis, 2022).

1. https://semeval.github.io/SemEval2023/tasks.html

En ce qui concerne la génération automatique de résumés, bien qu'il existe plusieurs corpus génériques de résumés d'actualité tels que CNN/DM (Nallapati *et al.*, 2016), NYT (Sandhaus, 2008) ou OrangeSum en français (Eddine *et al.*, 2020), ces derniers ne portent pas sur des textes juridiques. Parallèlement, des jeux de données juridiques français pour le résumé ont été développés, tels que le corpus CASS de comptes rendus de la cour de cassation (Bouscarrat *et al.*, 2019), ou encore récemment le corpus EUR-Lex-Sum (Aumiller *et al.*, 2022), qui est un corpus d'articles juridiques provenant de la plateforme de loi de l'Union Européenne. Cependant, à notre connaissance, il n'existe à ce jour aucun corpus en français combinant des problématiques liées à l'actualité et à l'utilisation de vocabulaire métier juridique.

3 Une collection de test pour le résumé de textes juridiques en français

Afin d'évaluer les performances des modèles de génération de résumé abstractif, nous avons identifié une collection de documents de l'actualité juridique française [2]. Cette collection sera notre corpus de travail pour cet article.

La collection est constituée de 8 485 documents de veille juridique et réglementaire pour les collectivités territoriales et les administrations publiques. Chaque document comporte (a) un titre, (b) un texte (corps/contenu), (c) un ensemble de méta-données associées, notamment la thématique du document, et (d) un résumé produit manuellement. Un exemple de document de la collection est présenté dans le tableau 1. Toutes ces informations sont rédigées par des spécialistes du domaine dont l'objectif est de maintenir l'actualité juridique à jour. Ces spécialistes se focalisent sur la nouveauté et l'évolution dans les lois, en commençant souvent par la présentation du contexte de la loi, le code et les différents aléas de cette dernière, suivis de la présentation du changement qui a eu lieu. Chaque résumé reprend le contexte général du contenu du document et l'actualité décrite dans ce dernier.

Si on considère notre tâche de résumé automatique, nous pouvons utiliser cette collection en considérant le contenu de chaque document comme le *document source* et chaque résumé produit manuellement par un éditeur spécialiste en la matière comme le *résumé de référence* (Gold Standard). L'objectif des systèmes que nous évaluons sera donc de produire pour chaque document source un résumé (*résumé généré*) le plus proche possible du résumé de référence. En moyenne, les documents sources et les résumés de référence comportent respectivement 485 et 81 mots. Aucune différence notable n'est observée entre les différentes thématiques.

La collection aborde sept thématiques juridiques. Chaque document est associé à une seule thématique, attribuée par les spécialistes. Le tableau 2 décrit brièvement chacune d'elles en illustrant avec un exemple de titre d'article, et fournit la répartition des documents selon ces différentes thématiques.

2. Cette collection, non diffusable à ce stade, est la propriété de notre partenaire industriel.

Titre : La Cour de cassation communique.
Contenu : Comme le précise sa préface, élaborée en collaboration avec le service de documentation, des études et du rapport et avec le service de communication, cet état des lieux doit paraître chaque mois (à l'exception des mois de juillet et août). Son objectif est de faire connaître l'activité de la chambre criminelle de la Cour de cassation à un public plus large que celui des magistrats, des avocats et des professeurs de droit. Car « la Cour de cassation tranche, en particulier dans le domaine pénal, des questions diverses et difficiles qui, par l'enjeu qui s'y attache, intéressent l'ensemble des citoyens ». Accessible sur le site de la Cour de cassation, elle est envoyée gratuitement par voie électronique à toute personne qui en fait la demande. Les arrêts sont classés par thématiques, par exemple pour ce premier numéro, Audience Blanchiment, Détention provisoire, etc.
Thématique : Justice
Résumé : La Cour de cassation vient de publier le premier numéro d'une sélection commentée de ses arrêts rendus par la chambre criminelle (no 1, juin 2020).

TABLE 1 – Exemple de document juridique avec ses titre, contenu, thématique et résumé.

4 Expérimentations et résultats

Nous présentons dans cette section le protocole d'expérimentations que nous avons utilisé, les modèles que nous avons choisis pour la génération automatique du résumé abstractif ainsi que les résultats obtenus. Il est à noter que les expérimentations que nous avons effectuées n'avaient pas pour objectif d'améliorer les performances des modèles utilisés en termes de génération de résumé, mais plutôt d'évaluer leur capacité de compréhension des textes de nature juridique.

4.1 Protocole expérimental

Pour notre étude, nous avons sélectionné des modèles de génération de résumé abstractif pour lesquels des variantes pré-entraînées en français étaient disponibles. Nous avons récupéré l'ensemble de ces modèles à partir de la plateforme *HuggingFace* [3]. Ces modèles sont tous basés sur une architecture de séquence à séquence, qui se compose d'une partie encodeur qui reçoit le texte source du document en entrée et produit une représentation vectorielle. Cette représentation est ensuite transmise au décodeur, qui a pour rôle de générer le texte du résumé en sortie.

3. https://huggingface.co/

Thématique	Description	Nombre de documents
Commande Publique	Couvre des sujets tels que les règles de passation des marchés publics, la réglementation des délais de paiement, les évolutions récentes du droit de la commande publique, etc. **Exemple :** *« Accord-cadre de travaux à bons de commande : quid du règlement des prestations ? »*.	3 398
Comptabilité et Finances locales	Aborde des sujets tels que la réglementation des budgets des collectivités territoriales, la gestion des dépenses et des recettes, les évolutions récentes du droit des finances locales, etc. **Exemple :** *« Sur quelles perspectives économiques préparer son budget 2023 ? »*.	577
Élections et Démocratie participative	Concerne les lois, règlements et jurisprudences relatifs aux élections en France (élections nationales, régionales, départementales, municipales). **Exemple :** *« Quel est l'office du juge lorsqu'il est saisi d'un compte de campagne ? »*.	303
État civil et Cimetières	Couvre des sujets tels que la tenue des registres d'état civil (naissance, mariage, décès), la délivrance des actes d'état civil, les règles relatives à la gestion et à l'entretien des cimetières, etc. **Exemple :** *« Quid du retour de la France au sein de la Commission internationale de l'état civil ? La réponse est non ! »*.	1 314
Justice	Concerne les lois, règlements et jurisprudences relatifs au système judiciaire français, ainsi qu'aux droits et obligations des acteurs de la justice, tels que les magistrats, les avocats, les victimes, etc. **Exemple :** *« 3 questions sur 10 ans de partenariat entre le Conseil National des Greffiers des Tribunaux de Commerce (CNGTC) et l'Ecole Nationale de la Magistrature (ENM) ! »*.	803
Ressources Humaines territoriales	Aborde des sujets tels que le recrutement, la formation, l'évaluation et la promotion des agents publics, les règles relatives à la discipline et à la sanction des agents, ainsi que les évolutions récentes du droit de la fonction publique territoriale. **Exemple :** *« Il faut remplir les conditions pour requalifier des vacations en CDI »*.	250
Urbanisme	Concerne les lois, règlements et jurisprudences relatifs à l'aménagement du territoire, ainsi que les évolutions récentes du droit de l'urbanisme. **Exemple :** *« Dispositifs de végétalisation de constructions : possibilité de déroger aux règles du PLU »*.	1 840

TABLE 2 – Répartition par thématique des documents juridiques de la collection de test.

4.1.1 Modèles génératifs évalués

Ci-après, nous décrivons les 4 modèles que nous avons utilisés pour le fine-tuning (voir la sous-section suivante 4.1.2 pour les paramètres utilisés) :

– **BART (Lewis *et al.*, 2020)** : il s'agit d'un modèle de langue avec une architecture séquence à

séquence, où l'encodeur est bidirectionnel et le décodeur est auto-régressif.

– BARThez (Eddine *et al.*, 2020) : il s'agit d'un modèle *transformer* français dédié à la génération de résumé, inspiré du modèle BART, possédant 6 couches bidirectionnelles pour l'encodeur ainsi que 6 couches pour le décodeur.

– Bert2Bert (Chen *et al.*, 2022) : il s'agit d'une architecture séquence à séquence ayant comme encoder et decoder une variante du modèle BERT pré-entraîné sur le français.

– T5 (Raffel *et al.*, 2020) : il s'agit d'un modèle de langue ayant une architecture auto-regressive de décodage. T5 a été entraîné sur plusieurs tâches du traitement automatique du langage naturel : la traduction automatique, la réponse aux questions, la génération de résumé, etc.

4.1.2 Fine-tuning des modèles de langue

Comme chacun des modèles utilisés a été pré-entraîné avec un corpus de données générique, généralement issu du web, nous avons décidé de procéder à une étape d'ajustement (*fine-tuning*) des modèles sur les données de spécialité que nous avons utilisées. Les modèles BART et T5 ayant été initialement pré-entrainés sur des corpus anglais, nous avons utilisé des variantes de ces derniers entrainées sur du texte français [4] [5].

La configuration par défaut de chaque modèle a été conservée, nous avons uniquement modifié les hyperparamètres suivants :

— Nombre d'epochs : 10,
— Batch size : 6,
— Nombre de tokens du texte donné en entrée : 512 tokens,
— Nombre de tokens du résumé généré : 100 tokens.

Nous avons utilisé la technique k-*fold* (ou validation croisée) pour le fine-tuning avec $k = 5$. Par conséquent, chaque modèle a été ajusté 5 fois avec à chaque fois une répartition $\frac{4}{5}$ de la collection utilisés pour l'entraînement et $\frac{1}{5}$ de la collection utilisé pour le test. Chaque partie ($\frac{1}{5}$ de la collection) contient 1 697 paires (document source – résumé de référence). Le pourcentage de distribution des thématiques, quant à lui, a été conservé sur la constitution de chacune des parts, c'est-à-dire, à égalité entre chaque part.

4.2 Résultats

Afin d'évaluer les performances des modèles que nous avons ajustés (ou « fine-tunés »), nous avons utilisé deux types distincts de métriques d'évaluation. D'une part, nous avons employé les métriques ROUGE (ROUGE-1, ROUGE-2 et ROUGE-L) (Lin, 2004) et BLEU (Papineni *et al.*, 2002), couramment utilisées dans la tâche de génération de résumés. Ces métriques permettent d'évaluer les résumés produits en termes de correspondance des n-grammes. D'autre part, nous avons utilisé un score de similarité sémantique *CosSim* pour évaluer la correspondance sémantique entre les résumés produits et les résumés de référence. Ce dernier est obtenu en calculant le cosinus entre les représentations des résumés de référence et des résumés générés. Notre collection étant en français, nous avons utilisé le modèle de langue CamemBERT (Martin *et al.*, 2020) pour le calcul

4. airKlizz/bart-large-multi-fr-wiki-news
5. plguillou/t5-base-fr-sum-cnndm

de représentations. Les résultats présentés dans le tableau 3 révèlent que le modèle Bert2Bert se distingue des autres en termes de performances, avec les meilleurs scores ROUGE, BLEU et CosSim. Les modèles BART et BARTHez ont des résultats relativement similaires et, à l'opposé, le modèle T5 montre les plus faibles résultats.

Modèle	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	CosSim
BART	0,34	0,13	0,20	0,29	0,92
BARThez	0,34	0,15	0,22	0,27	0,92
Bert2Bert	**0,39**	**0,19**	**0,25**	**0,36**	**0,93**
T5	0,30	0,08	0,17	0,30	0,89

TABLE 3 – Résultats des différents modèles sur les mesures ROUGE-1, ROUGE-2, ROUGE-L, BLEU (nous reportons la F-mesure) ainsi que le score de similarité sémantique.

Évaluation par thématique

Nous avons ensuite analysé les différents modèles selon chaque thématique de la collection, afin de vérifier si les comportements étaient similaires entre les thématiques. Les résultats de cette analyse sont affichés dans la figure 1 pour la métrique ROUGE-L. Il est à noter que les résultats pour les autres métriques (ROUGE-1, ROUGE-2, BLEU et CosSim) sont comparables et ne permettent pas de dégager de conclusions complémentaires. Ils n'ont donc pas été reportés ici.

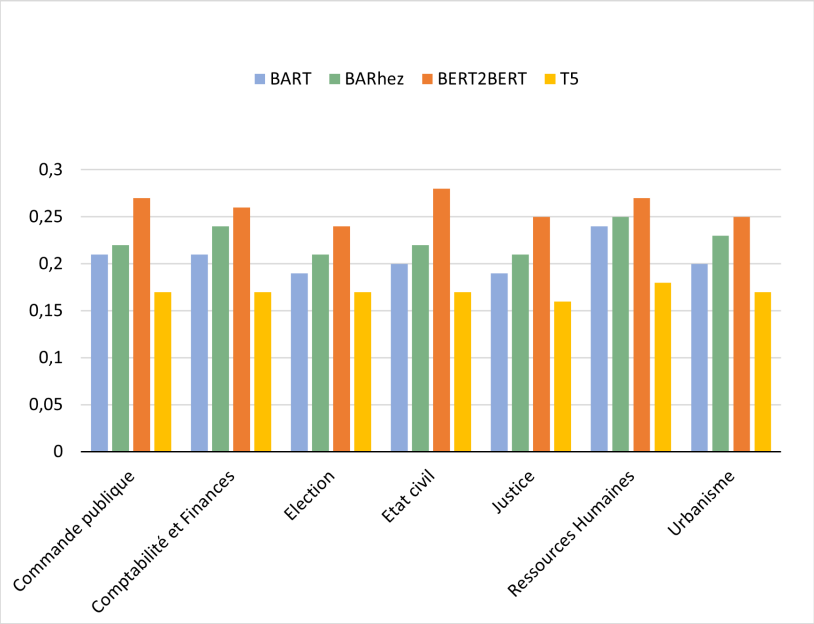

FIGURE 1 – Évaluation des performances des modèles par thématique selon la métrique ROUGE-L.

Nous pouvons distinguer deux observations des résultats de la figure 1. Premièrement, les résultats obtenus pour l'ensemble du corpus sont cohérents avec ceux obtenus pour chacune des thématiques. En effet, le modèle Bert2Bert surpasse les autres modèles sur l'ensemble des domaines, tandis que le modèle T5 obtient les scores les plus bas. Deuxièmement, bien que les documents soient présents en proportion différente dans les thématiques, les scores obtenus d'une thématique à l'autre

sont relativement similaires. Par exemple, le modèle Bert2Bert obtient un score de 0,27 pour les thématiques *Commande Publique* et *Ressouces Humaines territoriales*, bien que leurs occurrences soient très différentes (3398 et 250 respectivement). Ceci peut s'expliquer par le fait que les documents sont rédigés de manière uniforme et similaire, quelle que soit la thématique abordée.

5 Étude détaillée des résultats se basant sur les entités d'intérêt

Les domaines de spécialité tels que le juridique, la santé ou encore les sciences de manière générale sont des domaines particulièrement sensibles, où chaque concept utilisé, chaque nom propre ou même adjectif a une signification concise. Par conséquent, la véracité de l'information véhiculée dans ces domaines tient une place très importante. Les scores ROUGE et BLEU que nous avons obtenus nous permettent d'avoir une première indication sur la qualité des résumés générés par les modèles de langue. Cependant, ils ne nous permettent pas d'évaluer la *couverture* des résumés en terme de vocabulaire d'intérêt (vocabulaire métier et entités nommées), pas plus qu'ils ne permettent d'évaluer les *incohérences* par rapport au document source.

Dans cette section, nous proposons donc de nous focaliser sur une analyse plus poussée des résumés générés. Nous définissons le concept d'*entité d'intérêt*, c'est-à-dire une entité liée au domaine juridique ou une « simple » entité nommée. Dans la suite, nous choisissons d'évaluer la couverture des résumés et leurs incohérences à partir des entités d'intérêt détectées dans le document source et les résumés générés.

Les incohérences des documents sont appelées *hallucinations* dans la littérature. Une hallucination est définie comme une information qui ne peut pas être déduite à partir du document source (Maynez *et al.*, 2020). On distingue deux types d'hallucinations : les hallucinations *intrinsèques* qui correspondent à des mots ou groupes de mots présents dans le document source mais mal utilisés dans le résumé généré (un code de loi mentionné à mauvais escient par exemple), alors que les hallucinations *extrinsèques* correspondent à des mots ou groupes de mots non présents dans le document source. À ce stade, il est important de noter que les hallucinations peuvent être correctes (*factuelles*) : elles peuvent se baser sur une connaissance générale acquise en dehors du document. Les hallucinations intrinsèques étant compliquées à identifier, nous nous focalisons dans cet article sur les hallucinations extrinsèques afin d'évaluer les incohérences.

Une illustration des hallucinations est proposée dans l'exemple du tableau 4. L'exemple donné concerne un article juridique sur l'attribution des prénoms aux enfants à la naissance, qui énumère les différentes lois sur le sujet par ordre chronologique et explique les particularités et nouveautés de chacune d'elles. Dans le tableau, les entités d'intérêt sont en gras. Elles sont en rouge lorsqu'elles correspondent à des hallucinations extrinsèques. Cette notion d'hallucination extrinsèque a son pendant dans le résumé de référence. Les experts se sont en effet ici basés sur leur connaissances préalables et ont introduit une entité que nous considérons comme *abstraction* (en violet).

Dans la suite de cette section, nous détaillons la typologie des entités d'intérêt considérées, puis évaluons nos résultats en termes de couverture et d'hallucination.

Document source : Le choix des prénoms, liberté totale... ou presque. La **loi du 1 avril 1803** avait fixé les règles concernant le choix des prénoms, alors limité aux seuls en usage dans les différents calendriers ainsi qu'à ceux des personnages notoirement illustres dans l'Histoire. Ainsi, l'officier d'état civil n'avait alors qu'un simple rôle de vérification sans autre choix que de refuser tout prénom non conforme à cette prescription légale. Toutefois, la **Cour de cassation** avait admis dès 1981 qu'il n'y avait pas lieu d'exiger que le calendrier invoqué émane d'une autorité officielle. Poursuivant dans cette voie, la **loi n° 93-22 du 8 janvier 1993** a profondément modifié cette architecture en disposant, par l'**article 57 du Code civil**, que « [...]. L'officier d'état civil porte immédiatement dans l'acte de naissance les prénoms choisis. [...] Lorsque ces prénoms ou l'un d'eux, seul ou associé aux autres prénoms ou au nom, lui paraissent contraires à l'intérêt de l'enfant, l'officier de l'état civil en avise sans délai le procureur de la République. Celui-ci peut saisir le juge aux affaires familiales ». Ainsi, d'un système rigide instauré dès la période révolutionnaire, cette loi donne une liberté de principe et quasi-totale aux parents. Désormais, nul besoin de choisir un prénom existant, bien au contraire : il suffit de l'inventer... ou alors de donner à un prénom traditionnel une orthographe moins conventionnelle... L'unique limite à cette liberté est l'intérêt de l'enfant : le choix des prénoms ou leur association ne doit pas lui nuire. La loi donne alors à l'officier d'état civil un rôle central puisque c'est lui seul qui doit déterminer s'il convient ou non d'alerter le **Parquet**.	**Résumé de référence :** Jusqu'en 1993, la **loi du 1er avril 1803** encadrait strictement le choix des prénoms de l'enfant et cantonnait donc les parents au strict choix de prénoms déjà existants. Également, le principe d'immutabilité de l'état des personnes a longtemps limité drastiquement les cas de modification ou de changement de ces mêmes prénoms. C'est ainsi qu'une liberté nouvelle est apparue avec la loi du 8 janvier 1993 dans le domaine du choix et la **loi no 2016-1547 du 18 novembre 2016**, dite « Justice du XXIe siècle », élargit encore cette liberté à la sphère de la modification et du changement de prénom.
	Résumé généré par BART : La **loi n° 93-22 du 8 janvier 1993** portant réforme de la procédure à l'état civil a modifié les règles concernant le choix des prénoms des enfants par les parents. Les parents ne sont plus obligés de choisir un prénom existant pour leur enfant. La **Cour de cassation** s'est prononcée quant au calendrier en affirmant qu'il n'y avait pas lieu d'exiger que ce dernier émane d'une autorité officielle.
	Résumé généré par BARThez : La **loi n 93-22 du 8 janvier 1993** a profondément modifié cette architecture en disposant, par l'**article 71 du Code civil**, que « les prénoms de l'enfant sont choisis par ses père et mère .... L'officier d'état civil porte immédiatement dans l'acte de naissance les prénoms choisis par les parents ou l'un d'eux, seul ou associé aux autres prénoms ou au nom ».
	Résumé généré par Bert2Bert : La **loi n° 93-22 du 18 janvier 1992** a introduit dans le Code civil une disposition spécifique relative à l'attribution du nom de famille à un enfant. En effet, à partir de cette date, les parents ont la quasi-liberté de choisir les prénoms qu'ils désirent attribuer à leurs enfants. L'officier d'état civil est le seul acteur à pouvoir révoquer cette liberté donnée aux parents s'il juge que le prénom choisi nuit à leur enfant.
	Résumé généré par T5 : L'**article 57 du Code civil** prévoit que « le choix des prénoms est effectué par les parents sous le contrôle de l'officier d'état civil dans les actes de naissance, sans distinction de nom ou de prénom. Il n'y a donc pas lieu d'exiger que le choix des prénoms soit associé aux autres prénoms ou au nom de l'enfant choisi [...] ».

TABLE 4 – Exemple d'un document source, du résumé de référence associé et des résumés générés par les différentes approches. Les entités d'intérêt sont repérées en gras. Dans les résumés générés, les hallucinations extrinsèques sont en rouge. L'abstraction du résumé de référence est en violet.

5.1 Typologie des entités d'intérêt considérées

Afin d'identifier les différentes entités d'intérêt présentes dans notre collection, nous avons effectué une analyse manuelle de plusieurs échantillons. Nous avons observé une forte présence d'entités nommées de type personne, organisation ou localisation ainsi qu'un ensemble d'entités liées au domaine juridique. Nous avons extrait les entités nommées à l'aide d'un modèle CamemBERT (Martin *et al.*, 2020). Les entités juridiques quant à elles ont été extraites avec des expressions régulières en ne considérant que les types d'entités juridiques les plus courants et les patrons syntaxiques les plus stables afin d'éviter des biais de résultats. De façon détaillée :

- Nous considérons 4 types d'entités juridiques :
  - Loi
  - Article de loi
  - Proposition de loi
  - Décret
- Ces entités peuvent être exprimées selon 3 patrons syntaxiques différents :
  - Entité N° Numéro : ce patron syntaxique correspond à une entité nommée suivie d'un « N° » (abréviation de « numéro ») et d'un numéro identifiant l'entité en question. Exemple : « Loi n° 2016 - 1547 ».
  - Entité Code - Numéro : ce patron syntaxique correspond à une entité nommée suivie d'un code identifiant l'entité, suivi d'un tiret et d'un numéro. Exemple : « Article L. 2122-18 ».
  - Entité du Date : ce patron syntaxique correspond à une entité nommée suivie de l'article « du » suivi d'une date. Exemple : « Décret du 2 juin 2021 ».

Il convient de noter que ces patrons syntaxiques peuvent parfois être combinés dans une même expression. Par exemple, on peut rencontrer des expressions telles que « Loi n°5125 du 21 janvier 2023 » qui utilisent à la fois les patrons 1 et 3. Par ailleurs, nous avons choisi de ne considérer les dates que lorsqu'elles sont associées aux entités, car leur forme est très variable, et nous n'avons pas pris en compte les entités liées au temps (horaires, périodes de la journée, etc). Enfin, dans le cas où une entité est identifiée à la fois comme entité nommée et comme entité juridique, nous la considérerons exclusivement comme faisant partie de l'ensemble des entités juridiques. En effet, nous avons observé que certaines lois portent des noms de personnes ou de lieux.

5.2 Métriques considérées

Sachant les entités d'intérêt présentées dans la section précédente, nous proposons d'évaluer deux types de métriques : la couverture et le taux d'hallucination / abstraction.

Soient $\mathcal{N}(d), \mathcal{N}(r), \mathcal{N}(g)$, les nombres d'entités d'intérêt présentes respectivement dans le document source d , dans le résumé de référence r (gold standard) et dans le résumé généré g.

Une première catégorie de métrique concerne la couverture des résumés :

- le taux de couverture c_g des résumés générés :

$$c_g = \frac{\mathcal{N}(g \cap d)}{\mathcal{N}(d)} \tag{1}$$

où $\mathcal{N}(g \cap d)$ est le nombre d'entités de d trouvées dans le résumé généré.

— le taux de couverture c_r des résumés de référence :

$$c_r = \frac{\mathcal{N}(r \cap d)}{\mathcal{N}(d)} \tag{2}$$

où $\mathcal{N}(r \cap d)$ est le nombre d'entités de d trouvées dans le résumé de référence rédigé par les experts. c_r peut être vu comme un maximum atteignable par les différents modèles.

Les métriques de couverture se basent toutes sur le document source à résumer ($\mathcal{N}(d)$). Elles diffèrent en cela des métriques proposées par (Nan *et al.*, 2021) qui comparent les résumés générés aux résumés de référence ($\mathcal{N}(r)$). Nous n'avons pas fait ce choix afin d'avoir des métriques généralisables dans le cas où les résumés de référence n'existeraient pas. D'autre part, comparer les entités à celles du document source permet non seulement d'évaluer les résumés générés, mais également les résumés de référence, ce qui n'est à notre connaissance pas fait dans la littérature.

Une deuxième catégorie de métrique est liée à l'apparition d'entités dans les résumés générés/de références, entités qui n'étaient pas présentes dans les documents sources. Nous définissons :

— le taux d'hallucination (extrinsèque) h :

$$h = \frac{\mathcal{N}(\neg g)}{\mathcal{N}(g)} \tag{3}$$

où $\mathcal{N}(\neg g)$ est le nombre d'entités hallucinées dans g. Le taux traduit le pourcentage d'entités hallucinées dans le résumé généré.

— le taux d'abstraction a :

$$a = \frac{\mathcal{N}(\neg r)}{\mathcal{N}(r)} \tag{4}$$

où $\mathcal{N}(\neg r)$ est le nombre d'entités abstraites dans r, c'est-à-dire, le pourcentage d'entités de r qui ne font pas partie des entités de d. Ces abstractions peuvent être comparées aux hallucinations extrinsèques des résumés générés dans le sens où elles ne portent pas sur des connaissances présentes dans d. Elles proviennent des experts qui ont réalisés les résumés de référence : ces derniers peuvent en effet se servir de leurs connaissances *a priori* pour rédiger les résumés. Il est cependant à noter que mêmes si comparables à des hallucinations extrinsèques, les abstractions sont factuelles, c'est-à-dire qu'on peut les considérer comme vraies, au contraire de certaines hallucinations extrinsèques.

À ces métriques, afin d'avoir une vision plus globale, nous ajoutons la proportion de résumés générés touchés par des hallucinations :

$$p_h = \frac{|G_h|}{|G|} \tag{5}$$

et la proportion de résumés de référence concernés par des abstractions :

$$p_a = \frac{|R_a|}{|R|} \tag{6}$$

où G est l'ensemble des résumés générés g, $G_h \in G$ est l'ensemble des résumés générés g contenant au moins une hallucination extrinsèque, R est l'ensemble des résumés de référence et R_a l'ensemble des résumés de référence contenant au moins une abstraction.

5.3 Résultats

Avant d'examiner les résultats en termes de couverture et hallucination, nous avons étudié la répartition et le nombre des entités d'intérêt dans les documents sources et les différents résumés. Cette analyse est illustrée dans la figure 2. Une première observation est que les entités juridiques sont, de façon non surprenante, de loin les plus présentes, à la fois dans les documents source et dans les résumés. Concernant les modèles, T5 se comporte très différemment des autres modèles : il est capable de générer beaucoup plus d'entités (et même beaucoup plus d'entités que le résumé de référence). Nous constatons enfin que les modèles ont d'une manière générale du mal à générer les entités de type personne (tous les modèles ont un nombre d'entités générées inférieur à celui du résumé de référence).

(a) (b)

FIGURE 2 – Comparaison du nombre d'entités présentes : (a) dans les documents sources, (b) dans les résumés de référence et résumés générés par type d'entité.

Le tableau 5 présente les résultats de couverture et hallucination des résumés de référence. Nous pouvons constater que les résumés de référence couvrent faiblement les entités présentes dans le document source (couverture à 0,09), ce qui n'est pas surprenant étant donnée la taille des résumés de référence. D'autre part, 28% des entités considérées sont abstraites. Ces résultats ne remettent cependant pas en question la factualité (véracité) des résumés, étant donné qu'ils ont été rédigés par des spécialistes du domaine.

	Couverture c_r	Abstraction a	Proportion p_a
Résumé de référence	0,09	0,28	0,36

TABLE 5 – Taux de couverture, d'abstraction et proportion des résumés concernés par les abstractions des résumés de référence.

Concernant les résumés générés automatiquement, les résultats du tableau 6 montrent des différences significatives dans les taux obtenus par les différents modèles. Bien que le modèle Bert2Bert ait obtenu les meilleurs scores ROUGE et BLEU, il obtient le taux de couverture le plus faible, le taux d'hallucination de loin le plus élevé (>60%) ainsi que la proportion de résumés hallucinés la plus élevée. Le modèle T5 présente quant à lui la meilleure couverture des entités du document source,

surpassant même celle des résumés de référence. Enfin, le modèle BART présente les taux les plus bas d'hallucination et de proportion d'hallucination.

Modèle	Couverture c_g	Hallucination h	Proportion p_h
BART	0,13	**0,19**	**0,25**
BARThez	0,09	0,21	0,29
Bert2Bert	0,03	0,61	0,72
T5	**0,18**	0,22	0,47

TABLE 6 – Taux de couverture, d'hallucination et proportion des résumés concernés par les hallucinations des résumés générés par les différents modèles.

Nous avons également regardé les résultats sous l'angle de la thématique abordée par les documents. Comme dans la section 4.2, nous n'avons pas observé de différence entre les différentes thématiques et ne reportons donc pas les résultats ici.

Une analyse complémentaire a concerné l'étude des taux de couverture et d'hallucination en fonction des entités concernées (personne, organisation, localisation et juridique). Les hallucinations concernent principalement les entités juridiques, probablement en raison de leur forte présence dans la collection de données. Les entités de type personne, organisation et localisation sont hallucinées de manière relativement similaire.

Enfin, afin d'examiner plus en détail les hallucinations, nous avons calculé un pourcentage d'intersection entre les entités hallucinées des résumés générés et les entités abstraites des résumés de référence. Les résultats sont présentés dans le tableau 7. Ils donnent une indication sur la factualité des hallucinations. Une fois encore, Bert2Bert obtient les résultats les moins convaincants, en contradiction avec les résultats des métriques traditionnelles. Ces analyses doivent cependant être poussées : sans remise en contexte des entités hallucinées, on ne peut pas déduire leur exacte factualité. Elles peuvent en effet être utilisées en provoquant des contre-sens ou de façon erronée.

BART	BARThez	Bert2Bert	T5
30%	22%	20%	27%

TABLE 7 – Pourcentage d'entités hallucinées faisant partie des entités abstraites

Tous ces résultats confirment dans leur ensemble qu'une simple analyse sur les métriques ROUGE et BLEU n'est pas suffisante dans un contexte métier. Le modèle Bert2Bert qui semblait être le plus performant sur les métriques classiques s'avère être celui qui génère le plus d'hallucinations "non contrôlées". Par conséquent, nous envisageons de poursuivre une étude plus détaillée des modèles T5 et Bart.

6 Conclusion

Dans cet article, nous avons identifié une collection d'articles d'actualité juridique. Nous avons ajusté (*fine-tuné*) sur cette collection 4 modèles de langue pour le résumé abstractif. Nous les avons évalué avec les métriques classiques du résumé automatique. Nous avons également mené une analyse détaillée des résumés générés à l'aide de la détection des entités nommées et des entités du domaine juridique. Cette analyse a montré que les scores ROUGE et BLEU ne sont pas suffisants pour évaluer des résumés abstractifs métiers. Cette observation souligne l'importance de prendre en considération des critères supplémentaires d'évaluation des résumés tels que la pertinence et la fidélité des informations produites, qui sont cruciaux dans les domaines de spécialité, tels que le juridique.

Cette étude ouvre sur plusieurs perspectives. À court terme, nous souhaitons poursuivre notre évaluation des hallucinations : (i) en détectant les hallucinations intrinsèques, (ii) en analysant la factualité des hallucinations dans leur ensemble, et (iii) en ajoutant d'autres métriques dédiées à l'évaluation des résumés, telles que les métriques basées sur les modèles questions-réponses (comme QuestEval (Scialom *et al.*, 2021) ou QAGS (Wang *et al.*, 2020)), les métriques basées sur la détection des faits (comme FactCC (Goodrich *et al.*, 2019)) ainsi que les métriques basées sur l'implication textuelle (comme PARENT (Dhingra *et al.*, 2019)).

A plus long terme, les modèles de générations peuvent être améliorés selon deux axes : (i) la limitation des hallucinations, dont une piste réside dans la suppression des abstractions dans les résumés de référence (Nan *et al.*, 2021), et (ii) le contrôle de ces dernières, en apprenant aux modèles à halluciner des informations factuelles (véridiques). Ces deux perspectives pourraient permettre d'obtenir des résultats plus précis et fiables dans la génération de résumés dans le domaine juridique, domaine métier dans lequel la véracité de l'information est cruciale.

Références

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

AUMILLER D., CHOUHAN A. & GERTZ M. (2022). Eur-lex-sum : A multi-and cross-lingual dataset for long-form summarization in the legal domain. *arXiv preprint arXiv :2210.13448*.

BOUSCARRAT L., BONNEFOY A., PEEL T. & PEREIRA C. (2019). Strass : A light and effective method for extractive summarization based on sentence embeddings. *arXiv preprint arXiv :1907.07323*.

CAO Z., WEI F., LI W. & LI S. (2018). Faithful to the original : Fact-aware neural abstractive summarization. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18 : AAAI Press.

CHEN C., YIN Y., SHANG L., JIANG X., QIN Y., WANG F., WANG Z., CHEN X., LIU Z. & LIU Q. (2022). bert2BERT : Towards reusable pretrained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 2134–2148, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.151.

DERNONCOURT F., GHASSEMI M. & CHANG W. (2018). A repository of corpora for summarization. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan : European Language Resources Association (ELRA).

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*.

DHINGRA B., FARUQUI M., PARIKH A., CHANG M.-W., DAS D. & COHEN W. W. (2019). Handling divergent reference texts when evaluating table-to-text generation. *arXiv preprint arXiv :1906.01081*.

DOU Z., LIU P., HAYASHI H., JIANG Z. & NEUBIG G. (2021). Gsum : A general framework for guided neural abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, NAACL-HLT 2021*, p. 4830–4842 : Association for Computational Linguistics.

DUSART A., PINEL-SAUVAGNAT K. & HUBERT G. (2023). Tssubert : How to sum up multiple years of reading in a few tweets. *ACM Trans. Inf. Syst.* DOI : 10.1145/3581786.

EDDINE M. K., TIXIER A. J.-P. & VAZIRGIANNIS M. (2020). Barthez : a skilled pretrained french sequence-to-sequence model. *arXiv preprint arXiv :2010.12321*.

EL-KASSAS W. S., SALAMA C. R., RAFEA A. A. & MOHAMED H. K. (2021). Automatic text summarization : A comprehensive survey. *Expert Systems with Applications*, **165**, 113679.

ERMAKOVA L., COSSU J. & MOTHE J. (2019). A survey on evaluation of summarization methods. *Inf. Process. Manag.*, **56**(5), 1794–1814.

FABBRI A., LI I., SHE T., LI S. & RADEV D. (2019). Multi-news : A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 1074–1084 : Association for Computational Linguistics.

GOODRICH B., RAO V., LIU P. J. & SALEH M. (2019). Assessing the factual accuracy of generated text. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, p. 166–175.

HOCHREITER S. & SCHMIDHUBER J. (1997). Long short-term memory. *Neural computation*, **9**(8), 1735–1780.

HUANG Z., XU W. & YU K. (2015). Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv :1508.01991*.

JI Z., LEE N., FRIESKE R., YU T., SU D., XU Y., ISHII E., BANG Y., MADOTTO A. & FUNG P. (2022). Survey of hallucination in natural language generation. *ACM Comput. Surv.* Just Accepted, DOI : 10.1145/3571730.

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 (ACL'20)*, p. 7871–7880.

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

LOUIS A. & SPANAKIS G. (2022). A statutory article retrieval dataset in French. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 6789–6803, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.468.

LUONG T., PHAM H. & MANNING C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, p. 1412–1421, Lisbon, Portugal : Association for Computational Linguistics. DOI : 10.18653/v1/D15-1166.

MA C., ZHANG W. E., GUO M., WANG H. & SHENG Q. Z. (2022). Multi-document summarization via deep learning techniques : A survey. *ACM Comput. Surv.*, **55**(5). DOI : 10.1145/3529754.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2020). CamemBERT : a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7203–7219, Online : 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., GULCEHRE C., XIANG B. *et al.* (2016). Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv :1602.06023*.

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.

PAPINENI K., ROUKOS S., WARD T. & ZHU W. (2002). Bleu : a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, p. 311–318 : ACL.

QI W., YAN Y., GONG Y., LIU D., DUAN N., CHEN J., ZHANG R. & ZHOU M. (2020). Prophetnet : Predicting future n-gram for sequence-to-sequence pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : Findings, EMNLP 2020*, p. 2401–2410 : Association for Computational Linguistics.

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. *The Journal of Machine Learning Research*, **21**(1), 5485–5551.

RUMELHART D. E., HINTON G. E. & WILLIAMS R. J. (1986). Learning representations by back-propagating errors. *nature*, **323**(6088), 533–536.

SAINI N., SAHA S. & BHATTACHARYYA P. (2019). Multiobjective-based approach for microblog summarization. *IEEE Trans. Comput. Soc. Syst.*, **6**(6), 1219–1231.

SANDHAUS E. (2008). The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, **6**(12), e26752.

SCIALOM T., DRAY P.-A., GALLINARI P., LAMPRIER S., PIWOWARSKI B., STAIANO J. & WANG A. (2021). Questeval : Summarization asks for fact-based evaluation. *arXiv preprint arXiv :2103.12693*.

SEE A., LIU P. J. & MANNING C. D. (2017). Get to the point : Summarization with pointer-generator networks. *arXiv preprint arXiv :1704.04368*.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł. & POLOSUKHIN I. (2017). Attention is all you need. *Advances in neural information processing systems*, **30**.

VINYALS O., FORTUNATO M. & JAITLY N. (2015). Pointer networks. In *NIPS*.

WANG A., CHO K. & LEWIS M. (2020). Asking and answering questions to evaluate the factual consistency of summaries. *arXiv preprint arXiv :2004.04228*.

ZHANG J., ZHAO Y., SALEH M. & LIU P. (2020a). PEGASUS : Pre-training with extracted gap-sentences for abstractive summarization. In H. D. III & A. SINGH, Éds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 de *Proceedings of Machine Learning Research*, p. 11328–11339 : PMLR.

ZHANG T., KISHORE V., WU F., WEINBERGER K. Q. & ARTZI Y. (2020b). Bertscore : Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020* : OpenReview.net.

ZHONG M., LIU P., CHEN Y., WANG D., QIU X. & HUANG X. (2020). Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, p. 6197–6208 : Association for Computational Linguistics.

ZHOU Y., PORTET F. & RINGEVAL F. (2022). Effectiveness of french language models on abstractive dialogue summarization task.