Abstraction ou Hallucination ?

Etat des lieux et évaluation du risque pour les modèles de génération de résumés automatiques de type séquence-à-séquence

Eunice Akani[1,2] & Frederic Bechet[1] & Benoit Favre[1]

[1]Aix-Marseille Univ, CNRS, LIS

[2]Enedis

30 juin 2022

Introduction

- Résumé automatique de texte ? : Générer automatiquement un résumé contenant les informations saillantes du texte en entrée.

- Problème ? :
  - L'insuffisance des mesures actuelles d'évaluation de résumé automatique telle que ROUGE [Lin, 2004].
  - Non fidélité des résumés prédits par rapport au document source.

Document : Il tourna sept films de la saga, dont "L'espion qui m'aimait", "Rien que pour vos yeux" et "Octopussy". Outre sa carrière cinématographique, Roger Moore s'était illustré au début de sa carrière dans plusieurs séries télévisées comme "Ivanhoé", "Le Saint" ou "Amicalement vôtre", où il partageait l'affiche avec Tony Curtis. Avec plus de cinquante films à son actif, Roger Moore avait quelque peu délaissé le grand écran ces dernières années. Ses dernières apparitions se sont faites essentiellement dans des téléfilms ou des séries. En 2003, il est fait chevalier commandeur de l'Ordre de l'Empire britannique et obtient également, en 2008, le titre de Commandeur des Arts et des Lettres décerné par la France. Très sensible à la cause animale, il soutenait activement l'association PETA. Après trois divorces, Roger Moore était marié depuis 2002 à une richissime danoise, Kristina Tholstrup.

Barthez : L'acteur et réalisateur américain Roger Moore, décédé à l'âge de 95 ans, est connu pour son rôle dans "L'espion qui m'aimait", "Rien que pour vos yeux" et "Octopussy".

C2C : L'acteur britannique Roger Moore, connu notamment pour son rôle dans "L'espion qui m'aimait", est décédé mardi à l'âge de 87 ans, a annoncé sa famille à la télévision.

PtGen : "roger moore, association peta, est décédé dimanche à l'âge de 85 ans, a annoncé sa famille à l'afp .l'association peta a également fait part de son passage à roger moore, "" ivanhoé"" ."

Travaux connexes & Contributions

Travaux connexes

- Mesure d'entailment pour évaluer la fidélité d'un résumé par rapport au document source [Maynez et al., 2020].
- [Maynez et al., 2020] a proposé une annotation des erreurs (hallucination extrinsèque et intrinsèque)
- Mesure d'évaluation du résumé prédit à l'aide d'un système de question-réponse [Durmus et al., 2020].

Contributions

- Proposition d'une typologie d'erreurs (pour les résumés candidats) et d'abstraction (pour les résumés de références)
- Mesure du risque d'hallucination des différents systèmes sur les entités.

Contributions

Typologie des erreurs

Pour l'annotation des résumés prédits par les systèmes de résumé automatique.

- Hors du document
- Agrammaticalité
- Erreur de référence
- Contresens
- Non inférable
- Autre

Typologie des abstractions

Pour l'annotation des résumés de référence.

- Abstraction inférable à partir du document
- Abstraction inférable à partir des connaissances de l'annotateur
- Abstraction non inférable

Typologie : Exemple

Document : Il tourna sept films de la saga, dont "L'espion qui m'aimait", "Rien que pour vos yeux" et "Octopussy". Outre sa carrière cinématographique, Roger Moore s'était illustré au début de sa carrière dans plusieurs séries télévisées comme "Ivanhoé", "Le Saint" ou "Amicalement vôtre", où il partageait l'affiche avec Tony Curtis. Avec plus de cinquante films à son actif, Roger Moore avait quelque peu délaissé le grand écran ces dernières années. Ses dernières apparitions se sont faites essentiellement dans des téléfilms ou des séries. En 2003, il est fait chevalier commandeur de l'Ordre de l'Empire britannique et obtient également, en 2008, le titre de Commandeur des Arts et des Lettres décerné par la France. Très sensible à la cause animale, il soutenait activement l'association PETA. Après trois divorces, Roger Moore était marié depuis 2002 à une richissime danoise, Kristina Tholstrup.

C2C : [L'acteur britannique] Roger Moore, connu notamment pour son rôle dans "L'espion qui m'aimait", [est décédé] [mardi] [à l'âge de 87 ans], [a annoncé sa famille à la télévision].

PtGen : "roger moore, association peta, [est décédé] [dimanche] [à l'âge de 85 ans], [a annoncé sa famille à l' afp] .l'association peta a également [fait part de son passage] à roger moore, "" ivanhoé"" ."

Exemple : Roger Moore [n'a tournée que deux films] de la saga.

Exemple : [Tony Curtis] tourna sept films de la saga, dont "L'espion qui m'aimait".

REF : Roger Moore [s'est éteint]. [L'acteur britannique] connu pour [son élégance en toutes circonstances et son humour] avait endossé [le costume de James Bond] [de 1973 à 1985].

Expériences

Idée : Étudier les hallucinations sur les entités.

Hypothèse : Un résumé avec le moins d'entités hors du document aura moins d'hallucinations.

- Annotation de résumés générés par différents systèmes à base de RNN et Transformers (100 résumés par système).
- Détection manuelle et automatique des entités nommées des résumés qui ne sont pas dans le document source
- Calcul du risque d'hallucination en supposant la référence comme information complémentaire au document source.

Cadre expérimental

Données :

- **OrangeSum** [Kamal Eddine et al., 2021] — Corpus d'actualités provenant du site de Orange Actu.

Systèmes utilisés:

- **PtGen** [See et al., 2017] — Modèle RNN basé sur les mécanismes de pointeur/générateur et couverture.
- **CamemBERT2CamemBERT (CTC)** [Martin et al., 2020] — Modèle séquence à séquence à base de Transformer (CamemBERT).
- **Barthez**[1] [Kamal Eddine et al., 2021] — Modèle français de BART.
- **mT5** [Xue et al., 2021] — Modèle multilingue.

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

Annotation des résumés et Résultats

- Les informations non fidèles au document source des résumés de références et candidats ont été annotées suivant la typologie des abstractions et des erreurs respectivement.

Models	% de résumés avec au moins une erreur	#nb moyen d'erreurs par résumé
PtGen	88	1.29
C2C	44	1.26
Barthez	43	1.27
mT5	**30**	**1.12**

Table: Statistiques des erreurs de chaque système.

	%abstraction
AbsDoc	26.5
AbsInf	13.68
AbsNInf	**59.83**

Table: Pourcentage d'apparition des différents types d'abstraction dans les 100 résumés de référence annotés.

** Pour 100 résumés annotés

Résultats I

	Manuel (100 résumés)		Auto (1500 résumés)	
Models	$\neg Doc$	$\neg Doc \cap \neg Ref$	$\neg Doc$	$\neg Doc \cap \neg Ref$
PtGen	**5.4**	100	**6.6**	97.2
C2C	16.53	90.47	9.7	87.5
Barthez	21.12	**74.07**	15.4	**80.8**
mT5	13.39	93.33	8.6	88.7

- PtGen est le système qui prend le moins de risque mais contient le plus d'erreurs
- Barthez prend le plus de risque ; il a le plus d'entités hors du document

Résultats II

Models	R-1	R-2	R-L	BERTScore	Risque (Manuel)	Risque (Auto)
PtGen (RNN)	28.16	9.55	19.12	19.80 / 69.95	5.4	6.6
C2C	**31.83**	**13.22**	22.87	27.04 / 72.66	16.53	9.7
Barthez	31.81	13.20	**23.07**	**28.63** / **73.25**	21.12	15.4
mT5	30.77	12.64	22.49	27.59 / 72.86	**13.39**	**8.6**

- Barthez est le meilleur système en terme de BERTScore ; mais sa valeur du risque est la plus élevée
- mT5 est le meilleur système (bon compromis entre ROUGE Score et risque)

Conclusion et perspectives

- Les modèles séquences à séquences prennent énormément de risque en générant des entités hors du document.
- La mesure du risque dépend du système d'extraction de résumé.
- Les corpus d'actualité utilisés pour le résumé de texte encouragent les modèles à halluciner.

- Évaluer le comportement des différents modèles lorsque l'on enlève les résumés dont la référence à des informations (entités) hors du document source.
- Réduire les hallucinations sur entités dans les résumés générés en utilisant un modèle de sélection de résumés.

Références I


Durmus, E., He, H., and Diab, M. (2020).
FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization.
In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.


Kamal Eddine, M., Tixier, A., and Vazirgiannis, M. (2021).
BARThez: a skilled pretrained French sequence-to-sequence model.
In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9369–9390, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.


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


Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de la Clergerie, É. V., Seddah, D., and Sagot, B. (2020).
Camembert: a tasty french language model.
In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Références II

Maynez, J., Narayan, S., Bohnet, B., and McDonald, R. (2020).
On faithfulness and factuality in abstractive summarization.
In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

See, A., Liu, P., and Manning, C. (2017).
Get to the point: Summarization with pointer-generator networks.
In *Association for Computational Linguistics*.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021).
mT5: A massively multilingual pre-trained text-to-text transformer.
In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Merci pour votre attention !