Détection automatique de citations erronées : jeu de données et méthodes

Qinyue Liu, Amira Barhoumi, Cyril Labbé

Détection automatique de citations erronées : jeu de données et méthodes

Qinyue LIU

Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
150 Pl. du Torrent, 38400 Saint-Martin-d'Hères, France
qinyue.liu@univ-grenoble-alpes.fr

RESUME. Les citations jouent un rôle important dans la recherche scientifique. Cependant, de nombreuses citations erronées sont identifiées au sein des publications scientifiques. Ces citations erronées, également appelées miscitations, peuvent conduire à une mauvaise interprétation des recherches citées, à une distorsion du message que l'auteur original souhaitait transmettre, et, potentiellement, à des conséquences plus graves. L'objectif de notre recherche est de détecter automatiquement les citations erronées selon le contexte de citation, et de construire un jeu de données contenant différents types de citations. Pour l'instant, nous avons constitué un jeu de données équilibré comprenant à la fois des citations fiables et erronées, issues de publications scientifiques en libre accès. En plus de ce jeu de données, notre étude propose deux méthodes basées sur le Traitement automatique des langues (TAL) pour distinguer automatiquement les citations erronées : une utilisant la similarité cosinus et l'autre, un classifieur de paraphrases, avec des plongements BERT en entrée. Selon nos résultats expérimentaux préliminaires, la similarité cosinus offre les meilleures performances sur notre base de données. Avec nos méthodes et le jeu de données équilibrés, nous nous concentrons pour l'instant sur la faisabilité de la détection automatique des citations fiables et erronées.

MOTS CLES. Citations erronées, Traitement automatique des langues, Article scientifique
ENCADREMENT. Cyril Labbé, Amira Barhoumi

1. États de l'art sur l'étude des citations

Certaines recherches sur les citations se concentrent quantifier la fréquence des citations erronée. Dans une étude sur les citations parue dans OHNS, 50 références aléatoires ont été analysées, révélant des erreurs dans 17% des cas, dont 34% considérées majeures (Armstrong et al., 2018). Il y a également des chercheurs qui ont analysé le contexte des citations pour identifier les tendances dans les sciences biomédicales (Jebari et al., 2021). Une autre recherche a développé une méthode pour étudier les thèmes cachés dans les publications, en analysant les résumés et les

citations d'un article source (Liu and Chen, 2013). D'autres études utilisant des techniques de TAL se sont consacrées à diverses tâches analytiques, telles que l'analyse du sentiment des citations (Liu, 2017) et la classification de la polarité des citations (Bordignon, 2022; Te et al., 2022).

2. Problématique : détecter automatiquement les citations erronées

De nombreuses études antérieures ont utilisé des techniques de TAL pour l'analyse des citations. Cependant, peu de recherches se sont concentrées sur l'évaluation automatique de la fiabilité des citations. À cet égard, notre étude vise à distinguer automatiquement les citations fiables et les citations erronées. Pour ceci, nous constituons un jeu de données en collectant des exemples de différentes types de citations, et testons différentes méthodes de TAL pour classifier automatiquement les citations.

3. Travaux réalisés : constitution du jeu de données et premiers tests de méthodes

Actuellement, nous concentrons sur l'évaluation de la similarité entre le contexte de citation et l'abstract des articles cités. Cette approche suppose qu'une citation fiable présente une grande similarité avec l'abstract. Nous avons donc collecté des citations pour créer notre jeu de données, sur laquelle nous avons ensuite testé nos méthodes.

3.1. Définition de différentes catégories de citations

Dans notre jeu de données, un contexte d'une citation "fiable" reflète correctement le contenu de l'article cité. À l'opposé, une citation "erronée" n'a aucun rapport avec le contenu de l'article cité (Voir Tableau 1).

3.2. Construction du jeu de données

Nous avons identifié deux méthodes pour collecter les données. La première méthode consiste à partir d'un article scientifique, parcourir la liste de ses références (articles cités par cet article). Puis, collecter dans cet article le contexte de citation pour chaque référence, ainsi que l'abstract de chaque référence. De cette manière, nous sommes capables d'évaluer la fiabilité des citations au sein d'un article scientifique.

La deuxième méthode consiste à partir d'un article et de son abstract , examiner tous les autres articles qui ont référencé cet article original. Pour chaque article qui a référencé l'article cité, nous identifions le contexte de citation, et nous le comparons avec l'abstract du article cité. De cette manière, nous pouvons découvrir combien de fois un article a été correctement cité.

Dans notre travail, nous avons appliqué principalement la deuxième méthode pour collecter les données. Les contextes de citation ont été manuellement rassemblés et

annotés à partir de divers articles en libre accès qui citaient 6 articles[1] dans des domaines scientifiques différentes. Au total, 199 citations ont été collectées pour le jeu de données. Pour garantir l'équilibre, 100 de ces citations sont fiables, tandis que 99 sont erronées.

Tableau 1. Exemples d'une citation erronée et d'une citation fiable

Catégorie	Contexte de citation	Abstract d'article cité
Fiable	For instance, other approaches for topic modelling can be tested.	Semantic similarity detection is a fundamental task in natural language understanding. Adding topic information has been useful for previous feature-engineered semantic similarity models, as well as neural models for other tasks. (Peinelt et al., 2020)
Erronée	Eddy covariance devices or lysimeters can be used to determine ET0	Male moths compete to arrive first at a female releasing pheromone. A new study reveals that additional pheromone cues released only by younger females may prompt males to avoid them in favor of older but more fecund females. (Vickers, 2017)

3.3. Méthodes pour classifier les citations

3.3.1. Similarité Cosinus

La similarité cosinus est largement utilisée pour mesurer la similarité entre deux textes sous formes de vecteurs (plongements qui capturent le contenu sémantique). Nous avons utilisé le modèle BERT (Devlin et al., 2018) pour générer les plongements du contexte de citation et de l'abstract d'article cité et comparer leurs similarités.

3.3.2. Classifieur de paraphrase

Nous avons fine-tune un classifieur également basé sur BERT (Devlin et al., 2018) en utilisant le corpus MSRP[2] pour différencier les citations fiables et erronées. La sortie du classifieur est catégorisée soit comme 'paraphrase', soit comme 'non paraphrase'. Dans notre cas, une sortie 'paraphrase' signifie une citation fiable ; Une sortie 'non paraphrase' signifie une citation erronée.

4. Travaux Futurs

Dans cette étude, notre objectif principal est d'évaluer la fiabilité des citations dans les articles scientifiques. Nous avons construit un jeu de données comprenant 199 contextes de citation, proposé et évalué deux méthodes sur nos données. La méthode

[1] Les DOI de 6 articles pour construire notre base de données : 10.1177/1609406919841251, 10.1371/journal.pone.0090972, 10.1007/s10900-017-0360-5, 10.18653/v1/2020.acl-main.630, 10.48550/ARXIV.1706.03762, 10.1016/j.cub.2017.05.0641

[2] https://www.microsoft.com/en-us/download/details.aspx?id=52398

de Similarité Cosinus a donné de meilleurs résultats, atteignant une précision de 93% sur notre jeu de données. La méthode de classifieur a atteint une précision de 87.4%. Pour les travaux futurs, nous envisageons d'ajouter d'autres types de citations erronées et d'agrandir le jeu de données. Nous souhaiterions également tester nos méthodes sur des articles scientifiques où le nombre de citations fiables dépasse celui des citations erronées, plutôt que tester sur notre jeu de données équilibré. De plus, nous envisageons de mener une recherche statistique pour évaluer si la section abstract d'un article cité suffit à justifier le contexte de citation dans l'article citant. Cette analyse aiderait à déterminer s'il est nécessaire d'analyser l'ensemble du article cité ou si se concentrer uniquement sur la section abstract est suffisant.

Remerciements

Nous remercions le projet NanoBubbles, financé par Conseil Européen de la Recherche (ERC) sous forme de subvention Synergie, dans le cadre du programme Horizon 2020 de l'Union Européenne, accord de subvention n° 951393.

Bibliographies

Armstrong, M.F., Conduff, J.H., Fenton, J.E., Coelho, D.H., 2018. Reference Errors in Otolaryngology–Head and Neck Surgery Literature. Otolaryngol.--head neck surg. 159, 249–253. https://doi.org/10.1177/0194599818772521

Bordignon, F., 2022. Critical citations in knowledge construction and citation analysis: from paradox to definition. Scientometrics 127, 959–972. https://doi.org/10.1007/s11192-021-04226-0

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/10.48550/ARXIV.1810.04805

Jebari, C., Herrera-Viedma, E., Cobo, M.J., 2021. The use of citation context to detect the evolution of research topics: a large-scale analysis. Scientometrics 126, 2971–2989. https://doi.org/10.1007/s11192-020-03858-y

Liu, H., 2017. Sentiment Analysis of Citations Using Word2vec. https://doi.org/10.48550/ARXIV.1704.00177

Liu, S., Chen, C., 2013. The differences between latent topics in abstracts and citation contexts of citing papers. J Am Soc Inf Sci Tec 64, 627–639. https://doi.org/10.1002/asi.22771

Payton, E., Khubchandani, J., Thompson, A., Price, J.H., 2017. Parents' Expectations of High Schools in Firearm Violence Prevention. J Community Health 42, 1118–1126. https://doi.org/10.1007/s10900-017-0360-5

Te, S., Barhoumi, A., Lentschat, M., Bordignon, F., Labbé, C., Portet, F., n.d. Citation Context Classification: Critical vs Non-critical.

Vickers, N.J., 2017. Animal Communication: When I'm Calling You, Will You Answer Too? Current Biology 27, R713–R715. https://doi.org/10.1016/j.cub.2017.05.064