Papierabruf, Zusammenfassung und Zitaterzeugung [1]

Nianlong Gu[2]

Abstract: Diese Arbeit präsentiert ein integriertes System für effizienten Abruf, Zusammenfassung und Erzeugung von Zitaten wissenschaftlicher Literatur. Wir schlagen ein Zwei-Stufen-Zitationsempfehlungssystem vor, das Geschwindigkeit und Genauigkeit ausbalanciert. Darüber hinaus stellen wir ein leichtgewichtiges Modell auf Basis von verstärkendem Lernen vor, um wissenschaftliche Artikel effizient zusammenzufassen. Wir präsentieren auch ein steuerbares Modell zur Zitaterzeugung, das durch bestimmte Zitatattribute gesteuert wird. Schließlich werden diese Teilsysteme in einer benutzerfreundlichen Benutzeroberfläche vereint, die zur KI-gesteuerten wissenschaftlichen Schlussfolgerung beiträgt und Autoren beim wissenschaftlichen Schreiben unterstützt.

1 Einleitung

Wissenschaftliche Schlussfolgerung kombiniert Beobachtung mit Vorwissen, um Schlussfolgerungen zu ziehen. Dieser Prozess führte zur Entdeckung von Neptun, als Unregelmäßigkeiten in der Umlaufbahn von Uranus auf ein unbekanntes Objekt hinwiesen. Diese Errungenschaft, die Newtons Gravitationsgesetz hervorhebt, hat seitdem zu anderen bedeutenden wissenschaftlichen Fortschritten wie der Quantenphysik und der Relativitätstheorie geführt. Wissenschaftliches Denken ist ein komplexer Prozess, der erhebliche Zeit erfordert, um relevante Literatur zu lesen und das erforderliche Vorwissen zu erlangen. Dennoch können menschliche kognitive Einschränkungen immer noch zu unzureichendem Wissen für unvoreingenommene, vernünftige Schlussfolgerungen führen. In dieser Studie untersuchen wir das Potenzial des Einsatzes künstlicher Intelligenz (KI), um wissenschaftliche Schlussfolgerungen zu automatisieren und damit diese Herausforderungen anzugehen.

Wissenschaftliche Schlussfolgerungen in wissenschaftlichen Arbeiten werden exemplarisch im Diskussionsteil des Manuskripts dargestellt, der oft auf den Ergebnisteil folgt, wie es in 63% der biomedizinischen Artikel im PubMed Central Open Access-Subset zu beobachten ist. Autoren vergleichen in der Regel experimentelle Ergebnisse mit verwandter Literatur, um Schlussfolgerungen zu ziehen. Dieser Prozess, der Ergebnisse als Beobachtungen, Literatur als Vorwissen und Diskussionsaussagen als Schlussfolgerungen umfasst, verkörpert wissenschaftliche Schlussfolgerungen. Da all diese Komponenten in natürlicher Sprache vorliegen, ist eine Automatisierung des Schreibens von Diskussionen mithilfe von Natural Language Processing (NLP)-Techniken möglich.

[1] Englischer Titel der Dissertation: "Paper Retrieval, Summarization and Citation Generation"

[2] ETH Zürich, Department of Information Technology and Electrical Engineering, Winterthurerstrasse 190, 8057 Zürich, Switzerland nianlong@ini.ethz.ch

Die Generierung von Diskussionsteilen durch NLP-Modelle stellt aufgrund der komplexen Natur der Abschnitte Ërgebnisseünd "Diskussionërhebliche Herausforderungen dar. Bestehende Sequence-to-Sequence-Modelle wie Transformers stoßen an Grenzen bei der Verarbeitung langer Eingabe- und Ausgabezeichenfolgen, die in der Regel auf 512 bis 1024 Tokens begrenzt sind. Anstatt darauf abzuzielen, den gesamten Diskussionsteil zu generieren, konzentriert sich dieser Artikel auf eine vorläufige Aufgabe: die Generierung von Zitatsätzen, die relevante Literatur referenzieren und diskutieren. Diese Entscheidung basiert darauf, dass das Zitieren und Diskutieren relevanter Arbeiten wesentliche Bestandteile des Diskussionsteils sind. Darüber hinaus kann das Verfassen von Zitatsätzen als mikroskopische wissenschaftliche Schlussfolgerung betrachtet werden, bei der Beobachtungen aus dem Kontext des Manuskripts und Vorwissen aus relevanten Arbeiten kombiniert werden, um Schlussfolgerungen oder Aussagen zu bilden. Diese Dissertation wurde in der ETH-Bibliothek veröffentlicht [Gu23].

2 Probleme und Motivation

Die Forschung in natürlicher Sprachverarbeitung (NLP) unterstützt Wissenschaftler, passende Artikel zu finden und zu zitieren. Es gibt Methoden, die allgemeine Informationen über Artikel nutzen, während andere auf spezifische Kontexte abzielen. Neuronale Modelle helfen, den Kerninhalt von Dokumenten zu verdichten und können verschiedene Textarten, auch wissenschaftliche Artikel, zusammenfassen. Die Zusammenfassungen können entweder ausgewählte Originalsätze [Zh20] oder neu erstellte Kurztexte sein [Hu21]. Ein besonderes Modell unterstützt Wissenschaftler beim Zitieren, indem es spezifische Textabschnitte und Zusammenfassungen der zitierten Artikel nutzt. Dieser Ansatz wurde um zusätzliche Informationen erweitert. Einige Studien erstellen auch einen Abschnitt über verwandte Arbeiten.

Trotz Fortschritten in automatischer Literaturrecherche, Zusammenfassung und Zitatgenerierung gibt es immer noch Herausforderungen bei der Nutzung dieser NLP-Techniken, um Forscher und Autoren bei der Erstellung wissenschaftlicher Artikel in der Praxis zu unterstützen.

2.1 Papierabruf

Die zunehmende Menge an wissenschaftlicher Literatur stellt eine Herausforderung für Empfehlungssysteme für Publikationen dar. Zum Beispiel enthalten Datenbanken wie S2ORC [Lo20] und PubMed Central Open Access (PMCOA) [Me03] Millionen von Artikeln, und diese Zahlen nehmen täglich zu. Während diese Fülle an Literatur den Forschern mehr Möglichkeiten bietet, erschwert sie gleichzeitig die Suche nach relevanten Papieren im Zeitalter der Informationsüberlastung.

Empfehlungsmodelle wie Graph Convolutional Networks (GCNs) [Je19] und große, vorab trainierte Sprachmodelle wie BERT [De19] können Zielzitate präzise abrufen. Allerdings sind diese Modelle rechentechnisch komplex, um mit einer großen Anzahl von Artikeln umzugehen. Andererseits bewerten Embedding-basierte Methoden [Gö20; PGJ18] Dokumente basierend auf ihrer Ähnlichkeit zu einer Abfrage mithilfe von Techniken wie Kosinus-Ähnlichkeit oder euklidischem Abstand. Diese Methoden sind effizient, da Dokumenten-Embeddings vorberechnet werden können und die Suche nach ähnlichsten Nachbarn beschleunigt werden kann. Allerdings sind sie, wie in Guo et al. [Gu19] diskutiert wird, weniger genau als BERT-basierte Abrufmethoden. Die Balance zwischen Genauigkeit und Geschwindigkeit ist entscheidend bei der Entwicklung eines praktischen Empfehlungssystems für wissenschaftliche Veröffentlichungen.

2.2 Wissenschaftliche Arbeiten zusammenfassen

Diese Studie konzentriert sich auf extraktive Zusammenfassungen, die in Syntax und Inhalt zuverlässiger sind als abstrakte Zusammenfassungen. Die Hauptherausforderung bei der extraktiven Zusammenfassung wissenschaftlicher Literatur ist die Länge der Dokumente. Unsere Forschung [GAH22] zeigt, dass wissenschaftliche Artikel aus dem arXiv-Datensatz durchschnittlich 5.206 Wörter und 206 Sätze im Hauptteil umfassen. Im Gegensatz dazu haben Dokumente im häufig verwendeten Dokumenten-Summary-Benchmark CNN/DM durchschnittlich nur 692 Wörter und 35 Sätze. Die längere Länge wissenschaftlicher Artikel erfordert ein Zusammenfassungsmodell, das effizient mit Dokumenten umgeht, die Hunderte von Sätzen und Tausende von Tokens enthalten.

Die Anwendung von Transformer-basierten Modellen wie BERT zur Textzusammenfassung stößt oft auf Herausforderungen aufgrund ihrer Beschränkungen hinsichtlich der Sequenzlängen. Die meisten dieser Modelle können Sequenzen von maximal 512 oder 1024 Tokens verarbeiten. Diese Beschränkung ergibt sich aus der hohen Speicherkomplexität, die in den Aufmerksamkeitsmechanismen dieser Modelle zum Einsatz kommt.

Die Speicherkomplexität stellt einen Hauptfaktor dar, der die Kapazität dieser Modelle bei der Verarbeitung langer Textsequenzen einschränkt. Die Aufmerksamkeitsmechanismen generieren Abfrage-, Schlüssel- und Wertmatrizen:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

$Q, K, V \in R^{n \times d_k}$ repräsentieren die Abfrage-, Schlüssel- und Wertmatrizen in jeder Aufmerksamkeitsschicht für jedes Token in der Sequenz, wobei d_k die verborgene Dimension darstellt. Da die Speicherkomplexität proportional zum Quadrat der Anzahl der Tokens ist ($O(n^2)$), steigt der Speicherbedarf mit zunehmender Sequenzlänge rapide an.

Um diese Herausforderung zu überwinden, haben Huang et al. [2021] einen effizienten Transformer vorgeschlagen, der die Speicherkomplexität durch eine modifizierte Aufmerksamkeitsberechnung auf einen linearen Faktor reduziert. Obwohl dieser Ansatz dazu beiträgt, die Einschränkungen herkömmlicher Transformer-Modelle zu überwinden, hat er immer noch viele trainierbare Parameter, was das Training von Grund auf erschwert und den Inferenzprozess verlangsamt.

Um diese Herausforderungen zu adressieren, schlagen wir ein Modell für die extraktive Zusammenfassung vor, das speziell darauf abzielt, lange wissenschaftliche Dokumente effizient zusammenzufassen. Unser Modell ist auf die Optimierung von Speicher- und Rechenzeit ausgelegt und erfordert keine Trunkierung langer Dokumente. Damit bietet es eine schnelle und effiziente Lösung für die Zusammenfassung von wissenschaftlichen Texten.

2.3 Erzeugung von Zitaten

Die Zitaterzeugung zielt darauf ab, eine Reihe von Zitaten zu erstellen, die innerhalb des Kontexts eines Manuskripts auf ein Papier verweisen und es diskutieren. Traditionell wurde diese Aufgabe als Problem der abstrakten Zusammenfassung angegangen. Zum Beispiel schlug [Ge21; XFW20] ein Sequence-to-Sequence-Modell vor, das kontextuelle Sätze aus dem Manuskript und die Zusammenfassung des zitierten Artikels als Eingabe nimmt und einen Zitatssatz als Ausgabe generiert. Diese Herangehensweise vernachlässigt jedoch, dass die Art und Weise, wie ein Papier zitiert wird, von verschiedenen bedingten Faktoren abhängen kann. Diese Faktoren umfassen die Zitationsabsicht (z. B. Einführung in den Hintergrund, Beschreibung der Methode oder Vergleich der Ergebnisse), die Schlüsselwörter, die zur Zitation eines bestimmten Papiers führen, und relevante Phrasen im Text des zitierten Papiers, die Details enthalten, auf die der Autor verweisen möchte. Wir argumentieren, dass Zitationsmodelle, die in diesem vereinfachten Kontext trainiert wurden, Schwierigkeiten haben können, die generierten Zitate an benutzerspezifische Bedingungen anzupassen, was ihre Nützlichkeit in realen Szenarien einschränkt, in denen Autoren Hilfe bei der Zitierung von Artikeln benötigen.

Anstatt sich ausschließlich auf die Automatisierung zu konzentrieren, wie in früheren Studien von Xing et al. [XFW20], ist unser Ziel die Entwicklung eines Zitationssystems mit größerer Kontrolle. Dieses System kann unterschiedliche Zitate generieren, wenn Benutzer verschiedene Attribute für den gewünschten Satz von Zitaten angeben. Die verbesserte Kontrollierbarkeit unseres Zitationssystems ermöglicht es Benutzern, die generierten Zitate flexibel anzupassen und so die Wahrscheinlichkeit zu erhöhen, dass geeignete Zitatsätze erzeugt werden, die den Zitationsanforderungen des Autors entsprechen.

2.4 Lösung für gemeinsames Retrieval, Zusammenfassung und Erstellung von Zitaten

Die Unterstützung von wissenschaftlichen Schreibprozessen durch getrennte NLP-gestützte Funktionen kann komplex und ineffizient sein. Forscher müssen häufig unterschiedliche Tools und Plattformen nutzen, um relevante Literatur zu finden, Zusammenfassungen zu erstellen und Zitiersätze zu generieren. Ein solcher fragmentierter Ansatz kann eine Barriere für die Akzeptanz und Nutzung dieser fortschrittlichen Tools darstellen.

Deshalb zielt diese Arbeit darauf ab, eine integrierte Plattform zu entwickeln, die den gesamten Prozess der wissenschaftlichen Literaturrecherche und -schreibung in einer einzigen, umfassenden Umgebung unterstützt. Dazu gehören das Durchsuchen von Literatur, das Lesen und Zusammenfassen relevanter Artikel und das Erzeugen von Zitattexten.

Wir planen, unsere entwickelten Module für Literaturrecherche, extraktive Zusammenfassung und Zitiergenerierung in die Plattform zu integrieren. Diese kombinierte Plattform wird den Nutzern ermöglichen, relevante Artikel zu suchen, wichtige Punkte herauszufiltern und Referenztexte zu generieren, alles in einem einzigen System.

Darüber hinaus bietet diese Plattform die Flexibilität, weitere NLP-gestützte Funktionen einzuführen, um die Benutzererfahrung weiter zu verbessern und den gesamten wissenschaftlichen Schreibprozess zu vereinfachen. Diese integrative Herangehensweise wird es Autoren ermöglichen, ihre Forschungs- und Schreibprozesse effizienter und produktiver zu gestalten.

3 Beiträge der Dissertation

Diese Studie untersucht umfassend vier Unterkategorien: 1) Entwurf und Implementierung eines Papierabrufsystems, das einen Ausgleich zwischen Geschwindigkeit und Genauigkeit erzielt; 2) Erforschung effizienter extraktiver Zusammenfassungsmodelle für lange wissenschaftliche Dokumente, die Redundanz effektiv reduzieren, ohne den Text übermäßig zu verkürzen; 3) Entwicklung eines steuerbaren Modells zur Generierung von Zitattexten, das Benutzern die Kontrolle über Eigenschaften wie Absicht, Schlüsselwörter und relevante Phrasen ermöglicht; 4) Entwurf und Implementierung einer Online-Plattform, die die Funktionen für Abruf, Zusammenfassung und Zitaterzeugung in einen nahtlosen Arbeitsablauf integriert.

3.1 Effizientes System zur Empfehlung lokaler Zitate für die Dokumentensuche [GGH22]

Um das optimale Verhältnis zwischen Effizienz und Genauigkeit bei der Papierempfehlung zu erreichen, haben wir einen zweistufigen Prozess vorgeschlagen. Er besteht aus einer

Prefetching-Phase, bei der wir eine kleine Kandidatenauswahl mit einem schnellen Abruf-Algorithmus erstellen. Die zweite Phase ist das Reranking, bei dem wir die Kandidaten neu bewerten und ordnen, basierend auf einem genaueren Ähnlichkeitsmodell. Hier nutzen wir ein auf Einbettungen basierendes System, das mittels eines siamesischen Textkodierers Vektoreinbettungen für jedes Paper berechnet. Unser Reranking verwendet ein fein abgestimmtes SciBERT-Modell, welches die Verbindung zwischen der Suchanfrage und dem Papier ermittelt und bewertet.

Um das Fehlen von umfangreichen Trainingsdatensätzen auszugleichen, haben wir einen neuen Datensatz erstellt. Dieser enthält 1,7 Millionen arXiv-Papiere und 3,2 Millionen lokale Zitatsätze, sowie Titel und Zusammenfassungen der zitierenden und zitierten Artikel. Wir haben auch die Auswirkungen unterschiedlicher Trainingsverluste auf die Leistung der Empfehlung analysiert und Strategien zur Verbesserung der Trainingskonvergenz entwickelt.

Unser Prefetching-System mit HAtten-Technologie hat in Vergleichstests besser abgeschnitten als alle anderen getesteten Systeme. Es profitiert zudem von der Kompatibilität mit GPU-Beschleunigung. Unsere Studie hat gezeigt, dass ein leistungsfähiges Prefetching-Modell für groß angelegte Zitationsempfehlungs-Pipelines von entscheidender Bedeutung ist. Es ermöglicht den Reranking-Modellen, weniger Kandidaten zu bewerten, ohne die Empfehlungsleistung zu beeinträchtigen. Dies führt zu einem besseren Gleichgewicht zwischen Geschwindigkeit und Genauigkeit.

3.2 Leichtgewichtige Dokumentenzusammenfassung mit Kenntnis der Extraktionsgeschichte [GAH22]

Um ein effizientes und redundanzresistentes extraktives Zusammenfassungssystem zu entwickeln, haben wir uns von der menschlichen Leseweise inspirieren lassen. Wir sind davon ausgegangen, dass Menschen nach dem Lesen eines Dokuments Satz für Satz eine Zusammenfassung erstellen und dabei bereits gewählte Sätze berücksichtigen. Redundanzen vermeiden sie intuitiv. Dieses Verhalten kann man als Markov-Entscheidungsprozess modellieren, bei dem die Auswahlhistorie zur Vermeidung redundanter Sätze hilft. Zudem betonten wir die Bedeutung eines effizienten Modells für die Zusammenfassung langer Dokumente, ohne den Text stark zu reduzieren.

Entsprechend dieser Leitlinien haben wir MemSum vorgestellt, ein Modell, das extraktive Zusammenfassungen als mehrstufigen episodischen Markov-Entscheidungsprozess behandelt. Es betrachtet ein Dokument als Liste von Sätzen und bewertet diese iterativ. Der Satz mit der höchsten Bewertung wird in die Zusammenfassung aufgenommen. Die Bewertungen der restlichen Sätze werden bei jeder Auswahl aktualisiert, basierend auf dem Satztext, dem globalen Kontext im Dokument und der Auswahlhistorie.

Unsere Ergebnisse zeigten, dass MemSum mit Hilfe der Auswahlhistorie kompaktere und redundanzfreiere Zusammenfassungen erstellen kann als Modelle ohne diese Histo-

rie. Dank eines effizienten Modells auf Basis bidirektionaler LSTMs und eines flachen Aufmerksamkeitsnetzes konnten wir auch lange wissenschaftliche Dokumente speichereffizient zusammenfassen. Unsere Fallstudien und menschlichen Bewertungen zeigen, dass die von MemSum erstellten Zusammenfassungen qualitativ hochwertiger sind als die konkurrierender Ansätze und weniger Redundanzen aufweisen.

3.3 Steuerbare Generierung von Zitationstexten [GH22]

Unser Konzept für ein kontrollierbares Zitiersystem basiert auf der Untersuchung konditionaler generativer Modelle. Die Idee ist, bedingte Eingaben zu nutzen, um die Generierung bestimmter Ausgaben zu fördern, wie dies in früheren Arbeiten gezeigt wurde.

Wir übertragen diesen Ansatz auf die Generierung von Zitaten. Dafür führen wir eine Reihe von zitatbezogenen Attributen als bedingten Eingabetext ein, wenn wir das Encoder-Decoder-Modell trainieren. Speziell entwickeln wir ein Conditional Citation Generation (CCG)-Modell, das kontextuellen Text aus dem Manuskript sowie den Titel und die Zusammenfassung der zitierten Arbeit als Grundeingaben und Zitierattribute der Zielzitate als bedingte Eingaben verwendet. Das CCG-Modell wird trainiert, um den Zielzitatensatz zu generieren, der die zitierte Arbeit im Kontext des Manuskripts zitiert.

Außerdem schlagen wir ein auf SciBERT basierendes Modul zur Vorschlag von Attributen vor. Es kann mögliche Schlüsselwörter, Sätze und die wahrscheinlichste Zitierabsicht vorschlagen. Das Modul ermöglicht es dem Benutzer, die vorgeschlagenen Attribute auszuwählen, um die Generierung durch das CCG-Modul zu steuern.

Unsere Experimente zeigen, dass unser CCG-Modul eine gute Kontrolle über die verschiedenen angebotenen Zitierattribute ermöglicht und unser Attributvorschlagsmodul effektiv relevante Zitierattribute vorschlägt.

3.4 SciLit: Integrierte Plattform für kollaboratives Retrieval, Zusammenfassung und Zitatgenerierung [akzeptiert von ACL 2023]

In dieser Systemvorstellung präsentieren wir SCILIT, eine Plattform für die Suche, Zusammenfassung und Zitierung wissenschaftlicher Arbeiten. SCILIT besteht aus einem NLP-Backend und einem benutzerfreundlichen Frontend, das mit React JS erstellt wurde.

Für das Backend haben wir Algorithmen entwickelt, die auf unserer aktuellen Forschung basieren, für das Durchsuchen von Artikeln, das Zusammenfassen und das Generieren von Zitaten. Mehr als 136 Millionen wissenschaftliche Arbeiten aus S2ORC haben wir verarbeitet und indexiert, um Nutzern eine umfangreiche Datenbank zum Durchsuchen, Zusammenfassen und Zitieren von Arbeiten zu bieten.

Wir haben die Effizienz von SCILIT in Bezug auf die Literatursuche, Artikelzusammenfassung und Zitatgenerierung bewertet und einen Absatz zur Generierung verwandter Arbeiten demonstriert, um unseren vorgeschlagenen Arbeitsablauf zu illustrieren. Zudem haben wir unsere Datenbank und die Algorithmen, die wir zur Erstellung unseres Systems verwendet haben, der Öffentlichkeit durch Open-Sourcing zur Verfügung gestellt.

4 Schlussfolgerung und Ausblick

In dieser Arbeit haben wir den Einsatz von NLP-Techniken zur Unterstützung von Autoren beim Durchsuchen, Zusammenfassen und Zitieren wissenschaftlicher Artikel erforscht. Durch die Untersuchung von drei separaten Forschungsthemen - Zitierempfehlung, extraktive Zusammenfassung langer Dokumente und steuerbare Zitiergenerierung - haben wir das Potenzial von NLP-Techniken für die Automatisierung wissenschaftlicher Inferenzen evaluiert.

Unsere Studien haben eine Grundlage für weitere Forschung in diesem Bereich gelegt und den Weg für die Entwicklung eines ganzheitlichen Systems sowie einer Webplattform geebnet. Diese ermöglicht es Benutzern, unseren Arbeitsablauf effizient zu nutzen.

Wir betrachten die Generierung von Zitaten als einen speziellen Fall wissenschaftlicher Inferenz auf mikroskopischer Ebene. Das Lösen dieses Problems könnte den Prozess des Schreibens des Diskussionsteils wissenschaftlicher Artikel erleichtern, was die Möglichkeit einer vollautomatischen Diskussionsgenerierung und KI-gesteuerten wissenschaftlichen Inferenz auf der Ebene von Abschnitten oder Kapiteln eröffnet.

Für zukünftige Arbeiten schlagen wir vor, effektivere Triplet-Mining-Strategien für das Auffinden von Dokumenten in großen Datenbanken zu erforschen, speichereffiziente Reranking-Systeme zu untersuchen und eine Pipeline für die extraktive und abstrakte Zusammenfassung wissenschaftlicher Dokumente zu entwickeln. Des Weiteren planen wir, das Training des Attributvorschlagsmoduls und des bedingten Zitiermoduls zu integrieren, um die Systemleistung zu verbessern.

Darüber hinaus ist die Verwendung von großen vortrainierten Sprachmodellen wie Galactica oder ChatGPT für die Aufgabe der automatischen wissenschaftlichen Inferenz ein wichtiger Bereich von Interesse für zukünftige Forschungen.

Literatur

[De19] Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs]/, arXiv: 1810.04805, Mai 2019, URL: http://arxiv.org/abs/1810.04805, Stand: 11. 03. 2020.

[GAH22] Gu, N.; Ash, E.; Hahnloser, R.: MemSum: Extractive Summarization of Long Documents Using Multi-Step Episodic Markov Decision Processes. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Dublin, Ireland, S. 6507–6522, Mai 2022, URL: https://aclanthology.org/2022.acl-long.450.

[Ge21] Ge, Y.; Dinh, L.; Liu, X.; Su, J.; Lu, Z.; Wang, A.; Diesner, J.: BACO: A Background Knowledge- and Content-Based Framework for Citing Sentence Generation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, Online, S. 1466–1478, Aug. 2021, URL: https://aclanthology.org/2021.acl-long.116, Stand: 22. 04. 2022.

[GGH22] Gu, N.; Gao, Y.; Hahnloser, R. H. R.: Local Citation Recommendation with Hierarchical-Attention Text Encoder and SciBERT-Based Reranking. In (Hagen, M.; Verberne, S.; Macdonald, C.; Seifert, C.; Balog, K.; Nørvåg, K.; Setty, V., Hrsg.): Advances in Information Retrieval. Springer International Publishing, Cham, S. 274–288, 2022, ISBN: 978-3-030-99736-6.

[GH22] Gu, N.; Hahnloser, R. H.: Controllable Citation Text Generation. arXiv preprint arXiv:2211.07066/, 2022.

[Gö20] Gökçe, O.; Prada, J.; Nikolov, N. I.; Gu, N.; Hahnloser, R. H.: Embedding-based Scientific Literature Discovery in a Text Editor Application. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Association for Computational Linguistics, Online, S. 320–326, Juli 2020, URL: https://aclanthology.org/2020.acl-demos.36.

[Gu19] Guo, J.; Fan, Y.; Pang, L.; Yang, L.; Ai, Q.; Zamani, H.; Wu, C.; Croft, W. B.; Cheng, X.: A Deep Look into Neural Ranking Models for Information Retrieval. arXiv:1903.06902 [cs]/, arXiv: 1903.06902, Juni 2019, URL: http://arxiv.org/abs/1903.06902, Stand: 01. 04. 2020.

[Gu23] Gu, N.: Paper Retrieval, Summarization and Citation Generation, en, Doctoral Thesis, Zurich: SNF, 2023.

[Hu21] Huang, L.; Cao, S.; Parulian, N.; Ji, H.; Wang, L.: Efficient Attentions for Long Document Summarization. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Online, S. 1419–1436, Juni 2021, URL: https://aclanthology.org/2021.naacl-main.112.

[Je19] Jeong, C.; Jang, S.; Shin, H.; Park, E.; Choi, S.: A Context-Aware Citation Recommendation Model with BERT and Graph Convolutional Networks. arXiv:1903.06464 [cs]/, arXiv: 1903.06464, März 2019, URL: http://arxiv.org/abs/1903.06464, Stand: 17. 03. 2020.

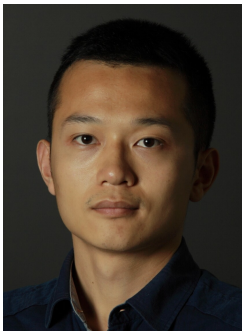
[Lo20] Lo, K.; Wang, L. L.; Neumann, M.; Kinney, R.; Weld, D.: S2ORC: The Semantic Scholar Open Research Corpus. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, S. 4969–4983, Juli 2020, URL: https://www.aclweb.org/anthology/2020.acl-main.447.

[Me03] of Medicine, B. (N. L.: PMC Open Access Subset [Internet], https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/, Accesses: 2022-07-30, 2003.

[PGJ18] Pagliardini, M.; Gupta, P.; Jaggi, M.: Unsupervised Learning of Sentence Embeddings Using Compositional n-Gram Features. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Association for Computational Linguistics, New Orleans, Louisiana, S. 528–540, Juni 2018, URL: https://www.aclweb.org/anthology/N18-1049.

[XFW20] Xing, X.; Fan, X.; Wan, X.: Automatic Generation of Citation Texts in Scholarly Papers: A Pilot Study. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, S. 6181–6190, Juli 2020, URL: https://aclanthology.org/2020.acl-main.550, Stand: 04. 10. 2021.

[Zh20] Zhong, M.; Liu, P.; Chen, Y.; Wang, D.; Qiu, X.; Huang, X.: Extractive Summarization as Text Matching. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, S. 6197–6208, Juli 2020, URL: https://www.aclweb.org/anthology/2020.acl-main.552.



Nianlong Gu hat seinen Doktortitel in Informationstechnologie und Elektrotechnik von der ETH Zürich im Jahr 2023 erlangt. Während seines Promotionsstudiums hat er Forschung im Bereich Natural Language Processing durchgeführt, insbesondere in den Bereichen Information Retrieval, Textzusammenfassung und Zitaterzeugung. Derzeit ist Nianlong Mitglied des NCCR@LiRI-Teams, wo er als Spezialist für maschinelles Lernen tätig ist. In dieser Funktion bietet er Beratung und Unterstützung für die Wissenschaftler des NCCR Evolving Language, insbesondere auf dem Gebiet des Sprachwandels, einschließlich Themen wie vokale Segmentierung unter Verwendung von maschinellem Lernen.