Projektstatus Report

Status gesamt: **grün**

Projekt: BibTexKonverter, Gruppe: Jürgen Bullinger, David Konieczny, Lars Lafleur, Constantin Schulz

29.05.2024

Erfolge, behobene Risiken oder Probleme (Farbcodierung = wie wichtig war Erledigung für Gesamterfolg)	• Erfolg 1 (wichtiger Milestone): Konvertierung von BibTeX-Code in Literaturstrings in Python (Bibliothek Pybtex) für die Stile Plain, APA und MLA: Die im Web gefundene Sammlung von BibTeX-Code „anthology.bib“ enthält Einträge des Typs Article, Proceedings und Inproceedings, die mithilfe von Pybtex konvertiert werden konnten. Daraus wurden Trainingsdaten mit den Attributen „Literaturstring“, „Literaturtyp“, „Style“ und „Bibtex“ erstellt. Pybtex kann später noch angepasst werden, um auch Referenzen in weiteren Stilen zu erzeugen. • Erfolg 2 (wichtiger Milestone): Eine Eruierung von ML-Modellen für NLP ist erfolgt. Die Plattform Hugging Face mit u.a. den Transformer- und Pytorch-Bibliotheken hat sich als vielversprechend herausgestellt. Erste Modelle/Ansätze wurden getestet, trainiert (Fine-Tuning) und auf Huggingface-Hub deployed. CUDA konnte erfolgreich installiert werden, sodass GPU zum Trainieren genutzt werden kann. • Erfolg 3 (wichtiger Milestone): Eine Pipeline für NER erkennt auch in Literatureinträgen die Named Entities. • Erfolg 4 (wichtiger Milestone): Es konnten bereits drei binäre Klassifizierer (auf Basis von DistilBERT) trainiert und getestet werden. Ein solcher Klassifizierer spezialisiert sich auf das Erkennen eines Formates. So schätzt zum Beispiel der APA-Klassifizierer, ob der Literatureintrag mit einer gewissen Wahrscheinlichkeit im APA-Stil ist oder nicht. • Erfolg 5 (mittelmäßig wichtiger Milestone): Eine erste Datenarchitektur konnte aufgebaut werden. • Erfolg 6 (mittelmäßig wichtiger Milestone): Aufbau GitHub-Repo: https://github.com/LaGit93/bibTexConverter_FaPra63065/tree/main • Erfolg 7 (mittelmäßig wichtiger Milestone): Installation einer standardisierten Entwicklungsumgebung (siehe conda.yaml) • Erfolg 8 (mittelmäßig wichtiger Milestone):): Meetingstruktur (2-3x wöchentliche Meetings), Agile Arbeitsweise (Kanban-Board) und schnelle Kommunikation über Discord-Server konnten etabliert werden
Risiken und Probleme (Farbcodierung = wie groß ist das Risiko für Gesamterfolg)	• Risiko 1 (mittleres Risiko): Das BibTeX-Konvertierungsproblem mit Seq2Seq-Modellen wie T5 zu lösen scheiterte, da das Trainieren (Fine-Tuning) zu ressourcenaufwendig war. Idee war es, die Konvertierung als Übersetzungsproblem („Rückübersetzung") zu behandeln. Es ist fraglich, wie ressourcenintensiv die binären Klassifizierer bei größeren Datenmengen werden. • Risiko 2 (mittleres Risiko): Bias in den Trainingsdaten: Durch die einheitliche Konvertierung von Bibtex-Code zu Literaturstrings mithilfe einer Bibliothek wird das “Rauschen” in Realbeispielen vernachlässigt. Auch wenn jemand den Stil APA benutzt, benutzt er vielleicht hier ein Komma statt einem Punkt, der andere klammert das Jahr und der Dritte nicht. Manche schreiben nur die Seitenzahlen, andere setzen davor pages, p., oder pp. Die von uns erzeugten Trainingsdaten werden dieses "Rauschen" nicht abbilden. Andererseits sind diese Trainingsdaten mit dem jeweiligen Stil gelabelt, d.h. so kann das Modell eine grobes Muster für den jeweiligen Stil lernen. Das ist der Vorteil zu gescrapten Referenzen, deren Stil ohne weiteres unklar ist. Es wird sich herausstellen, inwieweit sich hieraus Probleme ergeben. • Risiko 3 (mittleres Risiko): Seit 11 Tagen keine Kommunikation/Reaktion von Jürgen Bullinger. Sollten in der kommenden Woche keine Rückmeldung kommen, muss der Umfang des Projektes ggf. angepasst werden.
Geplante nächste Aktivitäten (Farbcodierung = wie wichtig ist Erledigung für Gesamterfolg)	• Aktivität 1 (wichtige Aktivität): Überprüfung der Qualität der Trainingsdaten (Sind die mit Pybtex erzeugten Literaturstring nach den jeweiligen Richtlinien des Styles formatiert?) • Aktivität 2 (wichtige Aktivität): Recherche von BibTeX-Code für books, incollections, phdthesis oder Erzeugung gemockter Daten. • Aktivität 3 (wichtige Aktivität): Es soll für die Formate APA, MLA, Havard, ACM und IEEE jeweils ein binärer Klassifizierer trainiert werden. Jeder Klassifizierer spezialisiert sich auf das Erkennen von einem Style. • Aktivität 4 (wichtige Aktivität): Es soll für die Typen book, article, proceedings, inproceedings und incollection jeweils ein binärer Klassifizierer trainiert werde. Jeder Klassifizierer spezialisiert sich auf das Erkennen von einem Typen.

	• Aktivität 5 (wichtiger Aktivität): Verwendung der generierten Trainingsdaten für das ML-Modell: Liefert das Modell anhand der selbst erzeugten Trainingsdaten zufriedenstellende Ergebnisse? • Aktivität 6 (wichtiger Aktivität): Es soll zusätzlich ein NER evaluiert werden, der auch Jahreszahlen und Datum erkennen kann. • Aktivität 7 (wichtiger Aktivität): Es sollen in Abhängigkeit von Style und Typ bestimmte Muster erkannt und durch reguläre Ausdrücke formuliert werden. • Aktivität 8 (mittelmäßig wichtige Aktivität): Testen der Bibliotheken zum Dursuchen von Litertaturdatenbanken zur Auflösung von Abkürzungen: Pybliometrics, Scholarly, Arxiv, Pybliographer
Offene Fragen, generelle Anmerkungen	Aktuell wird ein Ensemble-Learning-Ansatz verfolgt: Ein Literatureintrag soll zunächst den binären Klassifizierern zugeführt werden, um Style (IEEE,…) und Typ (article,…) zu erkennen. Der binäre Klassifizierer, der sich am sichersten ist, kommt zum Zuge. Anschließend soll es mithilfe der NER und den regulären Ausdrücken möglich sein, die relevanten BibTeX-Felder im Literatureintrag zu erkennen und zu extrahieren. Zusätzlich kann der Algorithmus die erkannten Felder nutzen, um sie in den externen Litertaturdatenbanken nachzuschlagen und um somit seine Schätzung zu verifizieren.