Beschleunigung von Ontologieentwicklung durch Sprachmodelle in den Digitalen Geisteswissenschaften - Nutzung der Formalisierung von Domänenwissen für eine effiziente Ontologieerstellung

Mitschunas, Johannes
johannes.mitschunas@uni-jena.de
Friedrich-Schiller-Universität, Deutschland
ORCID: 0009-0004-3579-1399

Beck, Clemens
clemens.beck@uni-jena.de
Friedrich-Schiller-Universität, Deutschland
ORCID: 0000-0001-5396-1612

Beckstein, Clemens
clemens.beckstein@uni-jena.de
Friedrich-Schiller-Universität, Deutschland
ORCID: 0000-0001-9099-7569

Stehfest, Robert Gramsch
robert.gramsch@uni-jena.de
Friedrich-Schiller-Universität, Deutschland
ORCID: 0000-0001-5939-5981

Einleitung

Strukturierte Wissensbasen, wie FactGrid, ermöglichen eine effiziente Datenabfrage und -analyse in der historischen Forschung. Ontologien sind hierbei zentral, da sie die Extraktion und Wiederverwendung von Informationen aus semistrukturierten Textkorpora unterstützen (MEPHISTO, 2020; HisQu-Projekt). Ein wesentliches Hindernis ist jedoch der zeitaufwändige Prozess der Formalisierung von Expertenwissen in Ontologiesprachen wie OWL.

Sprachmodelle (LLMs), insbesondere GPT-4, bieten eine Lösung, indem sie natürlichsprachliches Wissen in formalisierte Ontologiefragmente übersetzen können. Dies ermöglicht eine effizientere Entwicklung von Ontologien, während Experten die Validierung und Verfeinerung vornehmen (Mukanova et al. 2024).

Hintergrund

Im DFG-geförderten Projekt HisQu (HisQu 2024) ist die Entwicklung von Textparsern zur automatischen Datenextraktion aus semistrukturierten Texten, wie dem Repertorium Germanicum (RG), von zentraler Bedeutung. Diese Parser basieren auf einer expliziten ontologischen Grundstruktur. Als Grundlage dient hier das CIDOC-CRM-Modell, das kontinuierlich durch Ontologieschnipsel erweitert wird. Erste Studien, wie die von Mukanova et al. (2024), zeigen, dass LLMs wie GPT-4 in der Lage sind, Ontologien zu formalisieren und anzupassen, indem sie Expertenwissen für einen gegebenen Textkorpus nutzen und bestehende ontologische Strukturen verfeinern, statt sie zu erweitern.

Ziel

Dieses Poster präsentiert eine Pipeline, die darauf abzielt, Domänenexperten die Werkzeuge der Informatik zugänglich zu machen, um die Entwicklung von strukturierten Wissensbasen und Ontologien effizient voranzutreiben. Das Ziel ist es, ein Datenmodell zu entwickeln, das eine tiefere und automatisierte Extraktion von RG-Daten ermöglicht um diese in Form von Linked Open Data in Wissensgraphen wie FactGrid bereitzustellen.

Methodik

Die vorgeschlagene Pipeline umfasst die folgenden Schritte, welche iterativ wiederholt abgearbeitet werden, bis alle relevanten Bestandteile eines einzuordnenden Textkorpus abgedeckt werden:

1. Initiales Ontologie-Design: Ausgangspunkt ist eine bestehende Zielontologie (z.B. CIDOC-CRM). Diese bietet die Grundlage für die Formalisierung, indem sie zentrale Konzepte und Beziehungen definiert, welche dem LLM für die Formalisierung als Orientierung dienen.
2. Experteninput: Domänenexperten annotieren einen geschlossenen Abschnitt des Textkorpus, identifizieren relevante Begriffe und Beziehungen und liefern so die Grundlage für die Ontologieerweiterung.
3. LLM-Integration: Das annotierte Korpus und die initiale Ontologie werden einem LLM (z.B. GPT-4) übergeben. Das LLM schlägt formal strukturierte Ontologiefragmente vor, die das Expertenwissen aus dem Text repräsentieren. Die Rolle des LLMs ist dabei primär die Übersetzung von natürlichsprachlichem Wissen in formalisierte Ontologieschnipsel (Mukanova et al. 2024).
4. Konsistenzprüfung: Automatisierte Tools überprüfen die Konsistenz der generierten Ontologiefragmente. Zudem

validieren die Domänenexperten, ob die Abfragen des bereitgestellten Wissens mit der formalisierten Struktur übereinstimmen und sinnvoll beantwortet werden können.

5. Iterative Verfeinerung: Inkonsistente oder unvollständige Fragmente werden von den Experten überarbeitet, anschließend wird das LLM mit den neuen Informationen versorgt. Dieser Prozess wiederholt sich, bis die Ontologie konsistent und abfragbar ist.

6. Ontologieerweiterung: Validierte Ontologiefragmente werden in die Zielontologie integriert, um diese schrittweise zu erweitern und zu verfeinern.

Test und Validierung der Pipeline

Das HisQu-Projekt bietet mit der Fallstudie des Repertorium Germanicum (RG) ideale Bedingungen für die Erprobung der Pipeline. Das RG zeichnet sich durch eine wiederkehrende Struktur in den dort dokumentierten Prozessen aus, die eine effektive Modellierung in einer Ontologie ermöglicht. Durch die Vorarbeiten zu ANTLR-Parsern und die semantischen Strukturen in den Texten der Archivare können präzise, anwendungsbezogene Tests durchgeführt werden. Diese strukturierte Datenbasis, zusammen mit den Expertenannotationen, bildet den idealen Nährboden für die effiziente und präzise Entwicklung der Zielontologie und ermöglicht eine systematische Validierung der Pipeline.

Einschränkungen

Die größte Hürde bei der Ontologieentwicklung liegt traditionell darin, dass die Erstellung und Pflege von Ontologien tiefgehendes Fachwissen sowohl in der jeweiligen Domäne als auch in den formalen Techniken der Ontologieentwicklung erfordert. Diese Doppelkompetenz – das Verständnis der gesamten Ontologiestruktur und die Beherrschung der Ontologiesprachen wie OWL – ist selten in einer Person vereint. Die hier vorgestellte Pipeline mindert dieses Problem. Eine zusätzliche Herausforderung bleibt jedoch die Validierung der Ontologie, da diese nur von Domänenexperten zuverlässig durchgeführt werden kann. Darüber hinaus müssen LLMs aufgrund begrenzter Kontextlängen häufig mit segmentierten Korpora arbeiten, was mehrere Iterationen erforderlich macht.

Ergebnisse und Ausblick

Durch die Implementierung dieser Pipeline kann die Ontologieentwicklung erheblich beschleunigt und vereinfacht werden. Dies fördert die Digitalisierung und Zugänglichkeit historischen Wissens und ermöglicht eine effizientere Zusammenarbeit zwischen Domänenexperten und Informatikern. Besonders im Kontext der historischen Forschung bietet die Pipeline eine vielversprechende Lösung, um semistrukturierte und komplexe Textkorpora wie das Repertorium Germanicum systematisch zu modellieren. Die Fähigkeit der LLMs, komplexe Zusammenhänge zu verstehen und in formalisierte Ontologiefragmente zu übersetzen, unterstützt nicht nur die Strukturierung und Extraktion von Daten, sondern erleichtert auch die kontinuierliche Erweiterung und Anpassung der Ontologie an neue Erkenntnisse.

Bibliographie

Beckstein, Clemens, Robert Gramsch-Stehfest, Clemens Beck, Jan Engelhardt, Christian Knüpfer, und Oskar Jauch. 2022. „Digitale Prosopographie. Automatisierte Auswertung und Netzwerkanalyse eines Quellenkorpus zur Geschichte gelehrter deutscher Eliten des 15. Jahrhunderts.“ In: Digital History. Konzepte. Methoden und Kritiken digitaler Geschichtswissenschaften. Oldenbourg. DOI: https://doi.org/10.1515/9783110757101.

Mukanova, Assel, Marek Milosz, Assem Dauletkaliyeva, Aizhan Nazyrova, Gaziza Yelibayeva, Dmitrii Kuzin, und Lazzat Kussepova. 2024. „LLM-Powered Natural Language Text Processing for Ontology Enrichment.“ Applied Sciences 14, Nr. 13: 5860. https://doi.org/10.3390/app14135860.

Schulhoff, Sander, et al. 2024. „The Prompt Report: A Systematic Survey of Prompting Techniques.“ In: arXiv:2406.06608v3. DOI: https://doi.org/10.48550/arXiv.2406.06608.

Dell'Acqua, Fabrizio, et al. „Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality.“ Harvard Business School Technology & Operations Mgt. Unit Working Paper No. 24-013. DOI: http://dx.doi.org/10.2139/ssrn.4573321.