WOKIE - FAIR in allen Sprachen

Ein automatisiertes, LLM-gestütztes Übersetzungssystem für SKOS-Thesauri

Kraus, Felix
felix.kraus[at]kit.edu
Karlsruher Institut für Technologie (KIT), Deutschland
ORCID: 0000-0002-2102-4170

Tonne, Danah
danah.tonne[at]kit.edu
Karlsruher Institut für Technologie (KIT), Deutschland
ORCID: 0000-0001-6296-7282

Zusammenfassung. Thesauri sind zentrale Werkzeuge in den Digital Humanities, um heterogene Forschungsdaten zu strukturieren, auffindbar zu machen und auszuwerten. Eine Veröffentlichung nach den FAIR-Prinzipien wird insbesondere durch Mehrsprachigkeit erheblich verbessert, da sie Nachnutzung und Auffindbarkeit über Sprach- und Ländergrenzen hinweg ermöglicht. Bisher wird dies jedoch nicht durchgängig umgesetzt: Manuelle Übersetzungen sind aufwändig und externe Dienste liefern teils nur ungenaue Fachübersetzungen. Vor diesem Hintergrund stellen wir WOKIE vor, ein Open-Source-Werkzeug zur automatisierten, kontextsensitiven Übersetzung von SKOS (Simple Knowledge Organization System)-Thesauri. WOKIE kombiniert frei wählbare Übersetzungsdienste wie Google Translate, Argos oder PONS mit Large Language Models (LLMs). Durch den Einbezug von Definitionen und Kontextinformationen wird eine hohe Übersetzungsqualität erreicht, was in einer ersten Evaluation gezeigt wurde. Das Werkzeug ist dabei auf handelsüblichen PCs lauffähig. Im Vortrag demonstrieren wir typische Herausforderungen bei der Übersetzung, diskutieren die Bedeutung von Mehrsprachigkeit für die FAIRness von Thesauri und reflektieren Fragen der Urheberschaft bei LLM-generierten Übersetzungen. Dabei liegt der Schwerpunkt auf dem praktischen Nutzen und den Herausforderungen im geisteswissenschaftlichen Forschungsalltag.

1 Motivation

Kontrollierte Vokabulare, Taxonomien und Thesauri[1] spielen eine immer wichtigere Rolle bei der Strukturierung, Erschließung und Vernetzung von Forschungsdaten[2]. Gerade in den digitalen Geisteswissenschaften sind sie unverzichtbar, um komplexe, heterogene und teils räumlich stark verteilte Forschungsgegenstände zu beschreiben. Auch in Services und Infrastrukturen finden sie Verwendung, um Annotation, Suche oder computergestützte Auswertung zu ermöglichen.

Die aufgebauten Thesauri sind allerdings nicht nur Mittel zum Zweck, sondern wichtige, eigenständige Forschungsleistungen, die gemäß den FAIR[3]-Kriterien[4] veröffentlicht werden sollten. Ein zentraler Aspekt zur Verbesserung der *Findability* und *Reusability* von Thesauri ist ihre Mehrsprachigkeit, die bisher nicht durchgängig umgesetzt ist.

Beispielsweise wurde bei einem Thesaurus der Archäologie ein Drittel aller Begriffe von Deutsch auf Englisch übersetzt, andere Thesauri sind ausschließlich auf Deutsch verfügbar. Die Ursache liegt oftmals in den begrenzten Ressourcen: Thesauri werden häufig begleitend zu anderen Projektaktivitäten für einen spezifischen Anwendungsfall entwickelt, sodass eine manuelle Übersetzung aller Begriffe kaum leistbar ist. Auch der Einsatz externer Übersetzungstools bringt kaum Zeitersparnis, da sehr spezifisches Fachvokabular nicht in jedem Übersetzungsdienst vorkommt und die manuelle Integration in den Thesaurus zeitaufwendig ist.

[1] Im Folgenden wird nur noch der Begriff Thesaurus verwendet, welcher aber auch kontrollierte Vokabulare und Taxonomien einschließt.

[2] Hyvönen (2020)

[3] Findable, Accessible, Interoperable, Reusable

[4] Wilkinson u. a. (2016)

2 WOKIE: Funktionsweise und Mehrwert

Vor diesem Hintergrund haben wir das Übersetzungswerkzeug WOKIE (veröffentlicht unter Open-Source-Lizenz[5]) entwickelt[6], das einen Thesaurus automatisiert übersetzen und direkt in den Ausgangsthesaurus integrieren kann. WOKIE wurde mit dem Ziel konzipiert, den Aufwand für die Übersetzung signifikant zu reduzieren und gleichzeitig eine möglichst hohe Übersetzungsqualität zu erzielen.
Das System übersetzt zweistufig: In einem ersten Schritt nutzt es mehrere, frei wählbare Übersetzungsdienste wie Google Translate[7] (194 Sprachen), Microsoft Translator[8] (136 Sprachen), PONS[9] (unterstützt auch Latein) oder lokal ausführbare Systeme wie Argos Translate[10], was insbesondere bei datenschutzrechtlichen Anforderungen wichtig ist.
Im nächsten Schritt werden die Übersetzungen dieser Dienste verglichen. Wenn dabei unterschiedliche Übersetzungen entstehen, kommt ein Large Language Model (LLM) zum Einsatz. Dieses generiert unter Berücksichtigung von Kontextinformationen des Vokabulars (z. B. Definitionen) eine präzisere Übersetzung. Die verwendeten LLMs sind austauschbar, es kann ein lokal laufendes Modell verwendet werden. Außerdem wurde Wert darauf gelegt, dass WOKIE auf einem handelsüblichen PC oder Mac lauffähig ist und keine zusätzliche hochperformante Server-Infrastruktur notwendig ist.

Ein Beispiel aus dem DEFC (Digitizing Early Farming Cultures) Thesaurus[11] soll den Vorteil von kontextsensitiven LLMs zur Übersetzung illustrieren: Der englische Begriff *pulse* wird von gängigen Übersetzungsdiensten auf Deutsch als *Puls* oder *Impuls* übersetzt. Zwar sind auch diese Systeme in der Lage, Kontextinformationen zu berücksichtigen, erfordern in der Regel aber einen ganzen Satz und stoßen bei der Übersetzung eines alleinstehenden Terms an ihre Grenzen. Beim LLM hingegen wird durch zusätzliche Informationen wie

---

[5] https://doi.org/10.5281/zenodo.15313563

[6] Diese Forschungsarbeit wurde finanziert durch das Forschungsprogramm *Engineering Digital Futures* der Helmholtz-Gemeinschaft Deutscher Forschungszentren und der *Helmholtz Metadata Collaboration Platform* (HMC)

[7] https://cloud.google.com/translate

[8] https://www.microsoft.com/de-de/translator

[9] https://de.pons.com/p/ubersetzungsapi

[10] https://github.com/argosopentech/argos-translate

[11] https://vocabs.acdh.oeaw.ac.at/defcthesaurus/pulse/4.32

der Vokabularbeschreibung *pulse* korrekt als *Hülsenfrüchte* übersetzt, was mit der im DEFC angegebenen Übersetzung übereinstimmt.

3 Evaluation

Um das Werkzeug zu evaluieren, wurde bei bestehenden geisteswissenschaftlichen Thesauri eine Sprachversion entfernt und wieder in diese Sprache zurückübersetzt. Anschließend wurden Original und Rückübersetzungen (beide dann in derselben Sprache) verglichen[12]. Um eine untere Schranke für die Qualität der Rückübersetzung zu erhalten, wird diese nur als korrekt bewertet, wenn sie (abgesehen von Numerus) identisch mit dem Original ist. Beim DEFC-Thesaurus ist dies für 45% der Übersetzungen (Zielsprache Deutsch) der Fall, beim UNESCO-Thesaurus[13] für 90% der Übersetzungen (Zielsprache Englisch).

4 Zusammenfassung und Ausblick

WOKIE unterstützt Forschende innerhalb der Digital Humanities, ohne erheblichen Mehraufwand einen breiteren, mehrsprachigen Zugang zu ihren Thesauri zu schaffen. Die so erzeugten Thesauri stellen einen idealen Ausgangspunkt zur Qualitätsprüfung der Übersetzungen dar, besonders in Verbindung mit kollaborativen Vokabulareditoren wie dem von uns entwickelten EVOKS[14]. Aber auch ohne Nachbearbeitung kann ein Mehrwert erreicht werden, indem ein Hinweis zur maschinellen Übersetzung bei den entsprechenden Sprachen im Thesaurus hinzugefügt wird.
In unserem Vortrag stellen wir WOKIE vor und zeigen typische Übersetzungsprobleme im Umgang mit geisteswissenschaftlichen Begriffen, die häufig stark kontextabhängig sind. Außerdem diskutieren wir den potenziellen Mehrwert automatisierter Übersetzungen für die FAIRness von Thesauri. Wir stellen Grenzen und Fallstricke bei der Nutzung eines LLM-unterstützten Werkzeugs vor, um nicht zuletzt auch die Frage der Urheberschaft zu thematisieren: Wer 'verfasst' eine Übersetzung, wenn sie von einem LLM erzeugt wird?

[12] Rohdaten veröffentlicht unter https://doi.org/10.5281/zenodo.15313374

[13] http://vocabularies.unesco.org/thesaurus

[14] Ernst, Frank, und Götzelmann (2020)

Der Schwerpunkt des Vortrags liegt nicht auf technischen Details, sondern auf dem konkreten Nutzen und den Herausforderungen, die Anwenderinnen und Anwender im geisteswissenschaftlichen Forschungsalltag meistern müssen, um Thesauri durch Mehrsprachigkeit FAIRer zu gestalten.

Bibliografie

- Ernst, Felix, Laura Frank, und Germaine Götzelmann. „EVOKS - Benutzerfreundliche Erstellung Kontrollierter Vokabulare Für Die Geisteswissenschaften“. In *FORGE 2023 - Forschungsdaten in Den Geisteswissenschaften: Anything Goes?! Forschungsdaten in Den Geisteswissenschaften - Kritisch Betrachtet. Konferenzabstracts*. Tübingen, Germany, 2023. https://doi.org/10.5281/zenodo.8386468.

- Hyvönen, Eero. „Using the Semantic Web in Digital Humanities: Shift from Data Publishing to Data-Analysis and Serendipitous Knowledge Discovery“. *Semantic Web* 11, Nr. 1 (31. Januar 2020): 187–93. https://doi.org/10.3233/SW-190386.

- Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, u. a. „The FAIR Guiding Principles for Scientific Data Management and Stewardship“. *Scientific Data* 3, Nr. 1 (Dezember 2016): 160018. https://doi.org/10.1038/sdata.2016.18.