An LLM-based Approach for Translating Keywords in Scientific Publications

Luca De Santis[1,†], Paolo Pedinotti[1,†]

[1] *Net7 Srl, via Chiassatello 57, 56121 Pisa, Italy*

Abstract

We present herein a methodology and a working implementation for translating textual keywords of scientific publications. Using descriptive metadata to construct the context, this approach leverages Large Language Models (LLMs) to map keywords to entities of multilingual knowledge bases and controlled vocabularies, Wikidata in particular. By integrating these sources, it is not only possible to obtain keyword translations in multiple languages, but also to map them to Linked Data entities, disambiguating their meaning and improving the identification and classification of the associated publications. The methodology, developed during the ATRIUM research project, produced promising results when used with a commercial Large Language Model like ChatGPT. At the same time, our research highlights the challenges of reconciling free-form keywords, since the results can vary depending on the quality of the original metadata. While initially designed for the GoTriple discovery platform, this approach, along with its open-source example implementation, can be generalized to all situations where it is necessary to extract multilingual knowledge from text-based keywords.

Keywords

Metadata Enrichment, Text Processing, Multilingualism, Large Language Models, ChatGPT, Social Sciences and Humanities

1. Introduction

Multilingualism, defined as the practice of "writing and academic publishing in more than one language or having publications in more than one language" [1], is a common practice in many scientific disciplines, but it is in the context of the Social Sciences and Humanities (SSH) that this is prevalent. Several studies (e.g. [2], [3]) showed that, even if English is still the most used language for disseminating the outcomes of scientific research, SSH scholars often produce research papers in their local languages.

This phenomenon is not only a common practice, but it is highly encouraged amongst the SSH scientific community. For example, the OPERAS European research infrastructure put amongst its objectives "to support researchers that want to continue publishing in their own language and to develop transnational scientific cooperation at the same time" [4]. One of its services is the GoTriple multilingual discovery platform [5], which provides a central access point to find, access and reuse SSH-related materials such as articles, datasets, project descriptions and authors profiles.

At the time of writing, GoTriple indexes over 19 million documents metadata, 25 thousand project descriptions and cites over 22 million authors, all automatically acquired from harvesting more than 1,400 data sources. Finally, over 530 users have registered to the platform to have access to personalized features.

Being involved from the start in the development of GoTriple, our team has collaborated in implementing several automatic strategies to improve the multilingual support of the platform, including: the annotation of document descriptions with multilingual controlled vocabularies; the identification of languages of the textual metadata; language metadata normalization; the addition

[1] *4th International Conference on "Multilingual digital terminology today. Design, representation formats and management systems" (MDTT) 2025, June 19-20, 2025, Thessaloniki, Greece.*

[†] These authors contributed equally.

desantis@netseven.it (L. De Santis); pedinotti@netseven.it (P. Pedinotti)

0000-0003-0527-840X (L. De Santis); 0009-0009-2883-211X (P. Pedinotti)

of an English translation when absent, to facilitate the access of documents in local languages [6]. Moreover the GoTriple website is localized in 10 languages.

In GoTriple, the annotation of disciplines and controlled vocabulary terms enriches the original metadata of documents, which include after processing, labels in multiple European languages. This therefore facilitates the discovery of relevant content using a local language other than English. This added metadata are, as said, the result of an automatic process, based on Machine Learning and Natural Language Processing (NLP) techniques, which, while effective, cannot be defined as completely bulletproof (e.g. in [7] it is shown that the disciplines classifier only produces an average F1-score of around 50% over all its 11 supported languages).

Moreover, these classifications, while undoubtedly useful, cannot be considered as valuable as those originally applied by the document authors in the "keywords" attribute, that is, the free text descriptions added to provide a simple categorization of the content of the paper.

As indicated in [6], this specific document metadata proved to be problematic for automatic curation in GoTriple. In particular, the possibility to perform an automatic translation on them via a dedicated service (eTranslation [8] in the case of GoTriple) has been discarded: on the one hand, automatic systems may fail to perform well on short text, in particular when considered outside of a larger and more meaningful context (think of a term like "rock", that can be both applied to two distant subjects as Geology and Music).

On the other, for GoTriple metadata, it has been observed that articles often include keywords in multiple languages: in particular, when the text of an article is in a language different from English, the authors often add keywords in both the document language and in English, to ease the discovery of the article in scientific repositories.

To the keywords translation problem, a specific subtask (T3.4.1) of the on-going EU-funded research project ATRIUM (Advancing fronTier Research In the arts and hUManities) has been dedicated. The goal of ATRUM is to "bridge leading research infrastructures in arts and humanities (DARIAH), archaeology (ARIADNE), languages (CLARIN), and open scholarly communication in the social sciences and humanities (OPERAS)" [9].

The keywords translation task, albeit simple and straightforward in theory, presents numerous challenges, such as short keywords, lack of context, unidentified languages, and the use of multiple languages within a single publication's keywords.

In this article we present the work done by our team in this context. We start by presenting a review of interesting LLM-based approaches for metadata enrichment (Section 2). In Section 3, the methodology proposed in the ATRIUM project is presented, followed by a description of its implementation (Section 4), in the form of a publicly available Python code repository and an SSH Open Marketplace workflow [10], which describes a step-by-step documentation on how to use the aforementioned code. In Section 5, we present the experimental results obtained by applying the methodology on a selection of multilingual documents extracted from the GoTriple platform, while the conclusions provide a summary and suggest possible directions for future work.

2. Review of the use of LLMs for metadata enrichment

Numerous studies are available to demonstrate the great potential of using an LLM, and ChatGPT in particular, for enriching the descriptions of textual documents.

For example, in [11] ChatGPT was used for automatic genre classification on texts in English and Slovenian, providing results, even with a zero-shot approach, that outperformed a machine learning model fine-tuned on manually annotated datasets. In [12], the use of ChatGPT to classify hate speech has been empirically tested, showing a positive result even in presence of implicit hateful content, in 80% of the cases. In [13], it is stated that "ChatGPT outperforms crowd workers for several annotation tasks, including relevance, stance, topics, and frame detection" with "the zero-shot accuracy of ChatGPT exceeding that of crowd workers by about 25 percentage points on average".

More similar to our goals are the research described in [14], which showed how LLMs can be used to annotate subject metadata by providing classification examples through an in-context learning approach. The results obtained have been considered "promising" although the experiments have been conducted by using ChatGPT-3.5 which, as mentioned by the authors, performed poorly for the categorization of documents of specific disciplines (e.g. History and archaeology).

While the potential of using LLMs is widely recognized and proven, it is important at the same time not to ignore the potential problems that can arise in their use, in particular, the issue of generating the so-called hallucinations.

The research in [15], which focuses on the use of LLMs to create systematic reviews, highlights problems in obtaining accurate references, with a risk of having hallucinations "at a rate between 28% to 91%", according to the model used, stating also that "any references generated by such models warrant thorough validation by researchers".

In order to limit the risk of hallucination, [16] proposes an interesting approach based on refining the original LLM response by searching for supporting documents to verify and enforce any citation contained in it.

The idea of using an external corpus to support the LLM responses is defined in the article as "citation augmented strategies", which can be either "parametric", that is based on "information internalized from the training data" or "non-parametric", that is methods that "involve querying relevant information and seamlessly integrating the retrieved content from outside corpus" to enrich the LLM original response.

The validity of this approach finds confirmation in other studies, like [17] and [18].

3. The proposed methodology

Keyword translation has been approached by using an LLM to map the keywords to entities of multilingual controlled vocabularies, Wikidata in particular. Bibliographic keywords are included in publications in a "bag of words" manner, using concepts composed of one or more words, typically separated by commas, that describe the article's content but are not necessarily in a strict semantic relationship with each other: their meaning emerges when considered in relation to the content of the article.

Our methodology uses the idea of recreating a significant context for the interpretation of keywords by using the other publication's metadata, in particular its title, abstract and the language in which it is written. With this context, we produce a prompt to ask the LLM to recognize for each keyword a concept from Wikidata, returning also the URL of the corresponding Wikidata page.

We started our initial experiments by using ChatGPT with a prompt similar to the one indicated below.

> *We process a scientific article of which we have the TITLE, the ABSTRACT and the KEYWORDS separated by commas.*
>
> *The language of the article is <document_language> but the KEYWORDS can be in different languages.*
>
> *The goal is to map each keyword to a corresponding entity of Wikidata.*
>
> *Use the TITLE and ABSTRACT as context. Use this context to suggest a mapping of each keyword to a Wikidata entity, returning also its URL.*
>
> *TITLE: <document_title>*
>
> *ABSTRACT: <document_abstract>*
>
> *KEYWORDS: <document_keywords>.*

From the very start we noticed two important aspects. On the one hand, the results obtained were very often accurate, with the LLM able to identify and reconcile keywords to their exact Wikidata counterpart or to very logically close entities. On the other, the URLs provided were always wrong. ChatGPT was able to return correct Wikidata URLs which correspond to different entities.

As the retrieved concepts were accurate, we decided to apply to the LLM response a non-parametric citation augmented strategy, as defined in [16]: in our case, we query Wikidata via its API to obtain the correct URLs of the entities recognized by the LLM.

If the keyword text directly matches a label of a Wikidata entity, we can export its translations for all the languages that we need. More generically, we can establish a strong semantic association between the keyword and the Wikidata entity by using a predicate of the SKOS ontology [19], such as “exactMatch”.

If the keyword doesn’t literally correspond to the associated Wikidata entity’s labels, we will not use them as translations, but we can, in any case, create a weaker semantic link using the SKOS predicate “relatedMatch”.

4. The experimental implementation

The methodology described herein has been implemented using the Python programming language: all of the source code is freely available on GitHub [20].

The most significant code file is *main_functions.py*, which encapsulates the logic required to perform keyword translation tasks.

The first step involves text preprocessing, which, given an article, identifies and extracts the relevant metadata, in particular the title, the abstract and the keywords.

Then the LLM prompt is created, which includes the context built with the extracted metadata. To interact with LLMs, the Groq APIs [21] have been used, as they provide a convenient way to interact with multiple language models, both commercial and open source.

The prompt instructs the model to return recognized entities enclosed in square brackets, so that they can be easily retrieved using a regular expression.

Each entity is then used to query Wikidata’s APIs, in order to retrieve its URL along with the available translations provided on the platform.

The implementation is designed to be flexible, supporting both commercial and open-source large language models to accommodate diverse deployment requirements.

5. Measurements and results

The effectiveness of the proposed algorithm was measured on an annotated dataset of 200 articles extracted from GoTriple. This dataset was constructed by selecting documents in 21 languages, including English, belonging to 23 SSH disciplines, each containing at least four keywords (9.53 on average).

The annotation, made by the authors of this paper, consisted of manually mapping each entity to a corresponding Wikidata entity. Both "exact matches" and "related matches" were considered, with the latter category including similar, but not entirely precise, correspondences found in Wikidata. It was possible to include more than one Wikidata URL in those situations in which more entities could be significantly associated with a keyword. When no match was possible, the original keyword was left without any Wikidata association.

Examples of these annotations follow:

- Exact match: nation navajo -> https://www.wikidata.org/wiki/Q1783171 (Navajo Nation)

- Related match: négociations interethniques (interethnic negotiations) -> https://www.wikidata.org/wiki/Q118985945 (interethnic relations)
- No match: entreprise missionnaire contemporaine (contemporary missionary enterprise).

The experiment was limited to verifying the effectiveness of the algorithm in performing the reconciliation of keywords with Wikidata entities. Once the correspondence is created, the translations can be easily obtained, in multiple languages, using Wikidata APIs: therefore, this last step was not included in the test.

While annotating the keywords for the experiment, we noted that only a percentage of keywords could be safely associated with a Wikidata entity. In around 80% of the cases, it was possible to find a real association, both exact or related, the former in 64.72% of the cases.

The manual annotation represented the ground truth against which the results of the algorithm were evaluated. The metrics of precision and recall have been used for the evaluation.

The algorithm was tested on this dataset by using gpt-4o-mini. Results are shown in Table 1.

Table 1
Algorithm performance

	Precision	Recall	F1
Overall	0.58	0.56	0.56
Exact match	0.66	0.64	0.65
Related match	0.23	0.19	0.21

The algorithm's performance proves to be more effective in identifying an exact match, with a precision of 0.66 and a recall of 0.64 against the ground truth. On the other hand, the results for the looser matches were quite disappointing, as inevitably their choice brings greater uncertainty and, possibly, also a personal bias of the human who performs the annotation.

The script and dataset used for this experiment have been made freely available in the "evaluation_files" directory of the software's GitHub repository [20].

6. Conclusions and future work

We presented a methodology for translating the keywords of scientific publications by leveraging a Large Language Model to reconcile them with Wikidata entities. This reconciliation enables the retrieval of translations by utilizing the multilingual capabilities of this collaborative knowledge base.

The methodology and its associated implementation were evaluated against a testbed of manual annotations, demonstrating strong performance (around 65%) on a limited set of keywords that correspond readily to Wikidata concepts.

Of course, the benefits of this approach are not limited to translations. Reconciling keywords with Wikidata entities facilitates article classification and enhances the understanding of its main

subjects. This is particularly useful in a multilingual discovery platform like GoTriple, which features articles in many different languages.

On the other hand, the lack of standard workflows for creating keywords, along with noise introduced by data aggregators that may include classification codes (such as the Dewey Decimal Classification - DDC) as keywords, makes processing this metadata particularly challenging. In fact, manual annotation of our test data was generally feasible for 80% of the keywords, but an exact correspondence with Wikidata entities was achieved in only 64.72% of the cases.

Future directions for this work include exploring the possibility of reconciling keywords with other standard classification taxonomies and controlled vocabularies, such as the Library of Congress Subject Headings or DDC. Additionally, experimenting with open-source LLMs to compare their performance with that of ChatGPT will also be pursued.

Acknowledgements

This research was funded by the European Union, grant agreement number 101132163 (ATRIUM [9]).

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT 4o in order to: Grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] E. Kulczycki, T. C. E. Engels, J. Pölönen, Multilingualism of social sciences, in: E. Kulczycki, T. C. E. Engels (Ed.), Handbook on Research Assessment in the Social Sciences, ed., Edward Elgar Publishing, Cheltenham, UK, 2022, pp. 350-366. doi:10.4337/9781800372559.00031.

[2] E. Kulczycki, R. Guns, J. Pölönen, et al., Multilingual publishing in the social sciences and humanities: A seven-country European study, J. Assoc. Inf. Sci. Technol. 71 (2020) 1371–1385. doi:10.1002/asi.24336.

[3] N. Kancewicz-Hoffman, J. Pölönen, Does excellence have to be in English? Language diversity and internationalisation in SSH research evaluation, 2020. URL: https://enressh.eu/wp-content/uploads/2017/09/OverviewPeerReviewENRESSH-1.pdf.

[4] L. Delfim, M. Angelaki, A. Bertino, S. Dumouchel, F. Vidal, OPERAS Multilingualism White Paper, (2018). doi:10.5281/zenodo.1324026.

[5] GoTriple Discovery Platform. URL: https://gotriple.eu.

[6] L. De Santis, TRIPLE Deliverable: D2.5 - Report on Data Enrichment, (2022). doi:10.5281/zenodo.7359654.

[7] L. De Santis, F. Cau, D. Giacomi, T. Agosta, J. Homo, V. Ardizzone, I. Lamata Martínez, TRIPLE Deliverable D4.4 Technical and User Documentation for the TRIPLE system, (2023). doi:10.5281/zenodo.7708784.

[8] eTranslation, The European Commission's Machine Translation system. URL: https://commission.europa.eu/resources-partners/etranslation_en.

[9] ATRIUM Project website. URL: https://atrium-research.eu/.

[10] P. Pedinotti, LLM-Powered Mapping of Keywords of a Research Article to Linked Data, SSH Open Marketplace workflow, 2024. URL: https://marketplace.sshopencloud.eu/workflow/rEet9L.

[11] T. Kuzman, I. Mozetič, N. Ljubesic, ChatGPT: Beginning of an End of Manual Linguistic Data Annotation? Use Case of Automatic Genre Identification, (2023). doi:10.48550/arXiv.2303.03953.
[12] F. Huang, H. Kwak, J. An, Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In Companion proceedings of the ACM web conference (2023) 294-297. doi:10.48550/arXiv.2302.07736.
[13] F. Gilardi, A. Meysam, K. Maël, ChatGPT outperforms crowd workers for text-annotation tasks, Proceedings of the National Academy of Sciences 120.30 (2023).
[14] S. Zhang, W. Mingfang, Z. Xiuzhen, Utilising a Large Language Model to Annotate Subject Metadata: A Case Study in an Australian National Research Data Catalogue, (2023). doi:10.48550/arXiv.2310.11318.
[15] M. Chelli, J. Descamps, V. Lavoué, C. Trojani, M. Azar, M. Deckert, J. Raynier, G. Clowez, P. Boileau, C. Ruetsch-Chelli, Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis, J Med Internet Res, (2024). doi: 10.2196/53164.
[16] W. Li et al, Citation-Enhanced Generation for LLM-based Chatbot, (2024). doi:10.48550/arXiv.2402.16063.
[17] T. Gao, H. Yen, J. Yu, D. Chen, Enabling Large Language Models to Generate Text with Citations, Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 6465–6488.
[18] J. Menick, M. Trebacz, V. Mikulik, J. Aslanides, F. Song, M. Chadwick, N. McAleese, Teaching language models to support answers with verified quotes, (2022). doi:10.48550/arXiv.2203.11147.
[19] SKOS ontology. URL: https://www.w3.org/2009/08/skos-reference/skos.html
[20] Keyword translation code. URL: https://github.com/atrium-research/T3.4.1_KeywordsTranslation.
[21] Groq. URL: https://groq.com/.