Automatic Translation and Multilingual Cultural Heritage Retrieval: A Case Study with Transcriptions in Europeana

Mónica Marrero[1](✉), Antoine Isaac[1,2], and Nuno Freire[3]

[1] Europeana Foundation, The Hague, The Netherlands
monica.marrero@europeana.eu
[2] Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
[3] INESC-ID, Lisbon, Portugal

Abstract. Multilinguality is of particular interest for digital libraries in Cultural Heritage (CH), where the language of the data may not match users' languages. However, multilingual access is rarely implemented beyond the use of multilingual interfaces. We have run an experiment using the Europeana CH digital library as a use case. We evaluate the effectiveness of a multilingual information retrieval strategy using machine translations to English as pivot language. We conducted an indirect evaluation that should be considered preliminary. Yet, together with a manual analysis of the query translations, it already shows (or confirms) some of the benefits and challenges of deploying such systems in CH.

Keywords: Multilinguality · Cultural Heritage · Machine translation

1 Introduction and Related Work

Multilingual access to metadata and contents is of particular interest for international digital libraries (DL) in the area of Cultural Heritage (CH), which have collections in multiple languages, and users from different countries and with different cultural backgrounds. However, Multilingual Information Retrieval (MLIR) is rarely implemented in this domain beyond the interface language [9,15]. Only a few practical cases have been reported in the literature (see extensive reviews in Vassilakaki and Garoufallou [19], Diekema [3], and Chen [2]), and most of them use human translations and specialized vocabularies. This is the case for example of the World Digital Library [11], or the International Children's Digital Library[1], where contents are manually translated. In query translation, Bonet et al. [5] obtained good results using specialized dictionaries, while Kools et al. [7] obtained satisfactory results using machine translation. Matusiak et al. [9] reports an experiment using Google Translate to translate to English a collection of Chinese artworks, but they finally opted for human translation given

[1] http://en.childrenslibrary.org/.

G. Berget et al. (Eds.): TPDL 2021, LNCS 12866, pp. 133–138, 2021.
https://doi.org/10.1007/978-3-030-86324-1_17

the limitations found. In other domains machine translation seems to work well for the most widely spoken languages [4], with only a decrease of performance of 5–12% compared to the monolingual setting [13]. This lack of use of machine translation in DLs could be explained by the translation ambiguity and the insufficient lexical tools' coverage, considered to be among the most prominent problems in MLIR [12].

Europeana, a European digital library that aggregates content from libraries, archives and museums from all around Europe[2], is also a good example of this situation. It provides access to more than 60 million objects, from textual documents, like books or newspapers, to multimedia objects like audio, videos and paintings, which are primarily associated with 38 different languages. The data of these objects (i.e. metadata and content) is indexed in a search engine that provides a search functionality over all collections, however, in most cases, this data is only available in one language. Europeana performs data enrichment, adding persons, locations and concepts described in multiple languages to its metadata records. Yet the coverage of this approach is incomplete: there is no wide-spread translation of metadata, content and/or queries.

We have run an experiment using part of Europeana's collections to see the effectiveness of a MLIR system in this domain. We have focused on the content, not the metadata, and we have adopted a mixed approach where queries and object content are automatically translated to English as a pivot language, following the Europeana Multilingual Strategy [10]. Although document translation is considered more effective [12,13,17], this hybrid approach has outperformed other strategies in an experiment conducted [13], and it is more scalable when the number of different languages is considerable. Also, English is the most present language in these collections, and its effectiveness in machine translation is higher [4,13]. We have used the CEF translation service [1] as it is intended as a free, secure service for public bodies, which can be appealing for CH institutions, especially in Europe. The repository with the data of the experiment [8] and the client [6] used to get the translations are publicly available.

2 Data and Evaluation

We have selected a sample of 18,257 handwriting transcriptions of documents from the Europeana 1914–1918 thematic collection[3], obtained from the Transcribathon crowdsourcing platform [18]. This collection includes many World War I related objects contributed by members of the public all over Europe, like soldiers' diaries or letters. After removing 18 transcriptions that lacked indication of the original language, and those originally in English, we submitted 13,996 transcriptions to the service for translation to English. We received errors for 404 of them (2.9%), either because the language is not supported or because the text is too long and a different interface should then be used (this is part

[2] http://europeana.eu.

[3] https://www.europeana.eu/en/collections/topic/83-1914-1918.

Table 1. Original language of the transcriptions and queries (assuming for the queries it is the same as the language of the portal), and number of successful English translations.

Language tag	de	en	fr	it	ro	nl	el	lv	bs	cs	da	sl	hu	es	pl	sk	hr	Total
Transcriptions	9300	4243	1669	992	578	455	364	226	215	90	90	7	3	2	2	2	1	**18239**
Translated	9151	0	1659	973	577	454	356	226	0	90	90	7	2	2	2	2	1	**13592**
Queries	12	0	13	29	2	2	0	0	1	0	0	0	3	6	0	0	0	**68**
Translated	12	0	13	29	2	2	0	0	1	0	0	0	3	6	0	0	0	**68**

of our future work). As a result, we obtained 13,592 transcriptions translated to English from 15 different languages (see Table 1).

Regarding queries, we successfully translated a small sample of 68 queries issued in languages other than English from the logs of Europeana's 1914–1918 collection between January and August 2019.

We manually assessed the quality of translation of the queries, as they play a major role in the cross-lingual system. We also conducted a quantitative evaluation to answer the following research question: *is it possible to obtain similar results as those obtained with the original query, when searching on the same collection using translations?* Our assumption is that the results obtained in a monolingual system for a specific query and collection in that language, should also appear when searching with the translated query in the same collection translated to English. In order to answer this question, we compare two lists of retrieval results per query q in original language l: a) the set s_{qo} obtained when searching with the original query q_o in the transcriptions in l, and b) the set s_{qt} obtained when searching with the English translation of q_o, q_t, in the transcriptions in l translated to English. The precision and recall of s_{qt} with respect to s_{qo} is then computed. Finally, we calculate the additional number of transcriptions retrieved when using q_t in the whole corpus of English transcriptions (translated or not).

3 Results

After a manual assessment of the queries, we discovered that in a number of cases the input to the translation tool was wrong because the queries contain typos or have the wrong language assigned (i.e., our assumption that its language is the language of the portal is wrong). The first issue happened 6 times, while the second happened in 18 queries, with two of them having both issues at the same time. After removing them, and an additional 3 for which the user's intention was not clear to us, we manually analyzed the translation of the remaining 43 queries. In 37 cases the query was an entity that had to be left unchanged (e.g., 'Bernhard Stiens' is to be left unchanged, while 'Italia' must be translated to 'Italy'). The service correctly translated (that is, left unmodified) 20 of those entities (54%). In the remaining 6 cases, where the translation was supposed to be different from the original, the translation service did it correctly in 5 cases (83%).

The incorrect translation of named entities is the main source of problems as, setting aside other issues, there are more queries with entities than without: 42 of the 68 queries are (or include) named entities (62%). The problem is especially hard to solve as the named entities present and queried in the World War I context are very specialized (less-known authors, small villages) and sometimes incompletely referred to (e.g., 'Tonale' refering to 'Passo del Tonale'), or are formulated with typos (e.g., 'san elia' refering to Antonio Sant'Elia). In some other cases they include common nouns that are not correctly disambiguated (e.g.,'Antonio Sordi' and 'Fogliano' are translated from Italian as 'Antonio Deaf' and 'sheet' respectively). This ambiguity issue is also observed in queries not involving named entities. For example, 'carnet de route' is correctly translated from French as 'journey log' in the transcriptions, however the query 'carnet' is translated as 'notebook', so no relevant results are retrieved.

For the quantitative evaluation, we obtained precision, recall, and new translations found for the queries with search results, that is, 31 queries out of the 68 originally considered (see Table 2). The recall indicates that 67% of the objects in s_{qo} are retrieved when using the translations. As a negative counterpart, we have on average 49% of results that are not in s_{qo}. Given the poor quality of the translation of the queries, we would have to assume that those results are more likely to be noisy: in our case, on average 337 of those new transcriptions retrieved are less likely to be relevant. This could be however compensated in some cases by the new transcriptions found. When using q_t in the whole corpus of English transcriptions we retrieve an average of 687 new transcriptions per query. A quick review shows that some of those new results are relevant. For example, for the query 'domov' in Czech ('home' in English) we only retrieve 2 results, however if we search by 'home' in the English translations we retrieve more than 1500 transcriptions in 9 additional languages.

Table 2. Precision and recall obtained when comparing s_{qt} and s_{qo} per language, as well as additional transcriptions retrieved when searching on the translations of any language.

Language tag	cs	de	it	fr	ro	Average
Queries	1	8	16	4	2	**6.2**
Precision	0.15	0.57	0.44	0.74	0.5	**0.51**
Recall	1	0.87	0.57	0.70	0.5	**0.67**
New transcriptions	1527	397	823	851	1	**687**

4 Conclusions and Future Work

This experiment in a real scenario shows (or confirms) some of the benefits and the challenges of deploying MLIR systems in this specific domain. Albeit focused on a rather small set of queries, our case illustrates the problem of performing query translation in the CH context: the number of queries that we are sure the service should actually translate is way smaller than the number of queries that it should leave unmodified, so the selection of a high quality translation service is important. Additional techniques like controlled vocabularies and named entity recognition tools are also needed [16], although they need to be adapted to the specific domain and updated regularly.

We have observed a significant number of cases where the queries had typos or there was a mismatch between the language of the query and the language assigned according to the language of the portal. These cases are especially harmful as the translation service was not given appropriate input. A spelling-correction system could mitigate the first problem, while for the second, language detection based on various signals [14] could improve the results.

This work shows that without addressing these issues, the drawbacks of a multilingual system in a CH domain could easily exceed its benefits. The next step will be to address those challenges and complement the evaluation conducted with a more balanced sample of queries in terms of languages to see its impact in the results. A qualitative analysis of the retrieval results is also due to better account for additional benefits of the translation (e.g. synonyms).

References

1. CEF automated translation service: etranslation. https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation
2. Chen, H.: Digital library research in the US: an overview with a knowledge management perspective. Prog. Electr. Libr, Inf. Syst. **38**(3), 157–167 (2004)
3. Diekema, A.R.: Multilinguality in the digital library: a review. Electr. Libr. **30**(2), 165–181 (2012)
4. Dolamic, L., Savoy, J.: Retrieval effectiveness of machine translated queries. J. Am. Soc. Inf. Sci. Technol. **61**, 2266–2273 (2010)
5. España-Bonet, C., Stiller, J., Ramthun, R., van Genabith, J., Petras, V.: Query translation for cross-lingual search in the academic search engine PubPsych. In: Research Conference on Metadata and Semantics Research, pp. 37–49. Limassol, Cyprus (2018)
6. Freire, N.: GitHub repository: europeana-etranslation-research. https://github.com/nfreire/europeana-etranslation-research
7. Kools, J., Lagos, N., Petras, V., Stiller, J., Vald, E.: GALATEAS project (Generalized Analysis of Logs for Automatic Translation and Episodic Analysis of Searches). D7.4 Final Evaluation of Query Translation (2013), version 2.0
8. Marrero, M., Isaac, A., Freire, N.: Automatic translation and multilingual cultural heritage retrieval: a case study with transcriptions in Europeana [Dataset], June 2021. https://doi.org/10.5281/zenodo.5045066

9. Matusiak, K.K., Meng, L., Barczyk, E., Shih, C.J.: Multilingual metadata for cultural heritage materials: the case of the Tse-Tsung Chow collection of Chinese scrolls and fan paintings. Electr. Libr. **33**(1), 136–51 (2015)
10. Neale, A., Isaac, A., Manguinas, H., Moskalenko, D.: Multilingual strategy. Tech. rep., Europeana (2020). https://pro.europeana.eu/post/europeana-dsi-4-multilingual-strategy
11. Oudenaren, J.V.: The world digital library. Uncommon Culture **3**(5/6), 65–71 (2012)
12. Peters, C., Braschler, M., Clough, P.: Multilingual Information Retrieval: From Research to Practice. Springer, Heidelberg (2012)
13. Savoy, J., Braschler, M.: Information retrieval evaluation in a changing world: lessons learned from 20 Years of CLEF, chap. Lessons Learnt from Experiments on the Ad Hoc Multilingual Test Collections at CLEF, pp. 177–200. Springer, Cham (2019)
14. Stiller, J., Gäde, M., Petras, V.: Ambiguity of queries and the challenges for query language detection. In: CLEF 2010 LABs and Workshops. Padua, Italy (2010)
15. Stiller, J., Gäde, M., Petras, V.: Multilingual access to digital libraries: the European use case. Inf. Wissenschaft Praxis **64**(2–3), 86–95 (2013)
16. Stiller, J., Petras, V.: Best practices for multilingual access. Tech. rep., Europeana (2016). https://pro.europeana.eu/post/best-practices-for-multilingual-access
17. Stiller, J., Petras, V., Lüschow, A.: CLUBS Project (Cross-Lingual Bibliographic Search). M5.3 Final Evaluation (2019), version 1.0
18. Transcribathon crowdsourcing platform. https://transcribathon.eu
19. Vassilakaki, E., Garoufallou, E.: Multilingual digital libraries: a review of issues in system-centered and user-centered studies, information retrieval and user behavior. Int. Inf. Libr. Rev. **45**, 3–19 (2013)