europeana

Europeana DSI 4

Report: Experiments on an Information Retrieval cross-lingual approach with the eTranslation service

Revision History

Revision No.	Date	Author	Organisation	Description
0.1	14.10.2019	Mónica Marrero	Europeana Foundation	Draft
0.2	25.10.2019	Antoine Isaac, Nuno Freire	Europeana Foundation, INESC-ID	Review
1.0	25.11.2019	Antoine Isaac, Mónica Marrero	Europeana Foundation	Refinements and first version
1.1	06.10.2020	Mónica Marrero, Antoine Isaac	Europeana Foundation	Corrections in results, update introduction.
1.2	09.09.2021	Mónica Marrero, Antoine Isaac, Nuno Freire	Europeana Foundation	Improve structure, add section state of the art, add figure 3.

Table of Contents

Introduction

Currently Europeana Collections contains more than 60M million objects (paintings, textual documents, like archives or newspapers, multimedia objects like audio and videos) which are primarily associated with 38 different languages[1]. Europeana relies on a search engine that indexes the data for these objects (i.e. content and metadata) in order to provide users with a search functionality over the collection.

In most cases, only one language is available in the data (i.e., content and metadata) contributed to Europeana for those objects. But users from all around the world access our portal and issue queries in their native language, expecting to find any type of objects, whose contents and/or metadata are not necessarily present in that specific language.

Europeana performs data enrichment, adding to its metadata records contextual (named) entities (persons, locations and concepts) described in multiple languages. Yet the coverage of this post-processing is incomplete: there is no wide-spread translation of metadata, content and/or queries. In addition, queries are issued against the whole data (metadata or full-text separately) instead of routing them to specific languages. This results in missing relevant objects in other languages, but also noisy results when one word exists in more than one language with different meanings, or when different words in different languages are reduced to one common string after the normalization process (e.g., stemming).

[1] In reality metadata about these objects is much more varied: we have counted over 400 language tags can be used for metadata values.

In order to tackle those problems, and provide users with relevant documents independently of the language in which the queries are issued, we need to implement a multilingual information retrieval approach. This type of systems cover the retrieval of documents written in languages different from the language used for query formulation (cross-lingual retrieval), as well as within-language retrieval (monolingual retrieval). Two main strategies can be found in the scientific literature: translating and indexing the data in all the possible languages of the queries, or translating the queries to the different languages of the data of objects. The first approach suffers from efficiency and scalability issues, while for the second we have to deal with the usually poor translation quality of the queries (Peters et al. 2012). An additional problem in both cases could be how to merge the results in one single ranked list, although this will ultimately depend on the user experience design.

We have run an experiment using part of Europeana's collections to see the effectiveness of a MLIR system in this domain. For this first experiment, we have focused only on the content, not the metadata, and we have adopted a mixed approach where queries and object data (transcriptions) are translated to English as a pivot language (see Figure 1). This is in line with the Europeana Multilingual Strategy[2] and findings from the related work (see section 2) as well as the fact that English is the most present language in the Europeana collections. For the translations of queries and transcriptions, we have used the CEF Automated Translation service (eTranslation)[3].

We will analyze how different are the results obtained when we translate query and transcriptions to English, compared to the results retrieved with the original query in the transcriptions in that language. In addition, as the quality of the translation plays a big role in the experiment, we have additionally manually assessed the translation of the queries. The evaluation of the quality of the translated content is left for a future work[4].

The repository with the data and results of this experiment can be found at https://doi.org/10.5281/zenodo.5045066, while the online version of the Google excel sheet used to annotate the translation quality of the queries is accessible here.

Related work

MLIR is still a challenge in the area of Information Retrieval, and it is one of the most relevant topics addressed in academia that have a big impact in digital libraries (Agosti et al. 2019). However, in practice, the descriptions of digital objects in the digital collections, as well as the tools for their access and retrieval, remain largely monolingual, despite the evidence that the audience is multilingual (Matusiak et al. 2015) and thus could be interested in results beyond their native language. This is also supported by Stiller et al. (2013), who analyzed 31 CH digital libraries and observed that MLIR is rarely implemented beyond the interface language. Only a few practical cases

[2] https://pro.europeana.eu/post/europeana-dsi-4-multilingual-strategy
[3] https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation
[4] Evaluation of quality of text translations has also recently been carried out by Europeana, in the context of virtual exhibitions. While a different case, this can be combined with the insights from query evaluation to have an idea of the possible issues that would apply to metadata translation.

have been reported in the literature (see extensive reviews in Vassilakaki and Garoufallou (2013), Diekema (2012), and Chen (2004)), and most of them use human translations and specialized multilingual CH vocabularies — or aligned monolingual ones. This is the case for example of the World Digital Library (Oudenaren 2012), which provides access to its contents through item-level metadata records manually translated into seven languages, or the International Children's Digital Library[5], where the interface and contents are translated by volunteers. Matusiak et al. (2015) reports an experiment using Google Translate to translate to English a collection of Chinese artworks, but they finally opted for human translation given the limitations found with that approach. Also in the CH domain, Kools et al. (2013) obtained satisfactory results with the machine translation of queries in/to English, German and French. In other domains machine translation seems to work well. It has been reported that when the pair of languages includes English and one of the most widely spoken languages (such as Spanish, German, or French), currently available machine-translation systems offer high effectiveness from an IR point of view (Dolamic and Savoy 2010). An analysis of the multilingual ad-hoc retrieval task at CLEF (Conference and Labs of the Evaluation Forum) between 2000 and 2009 shows that there is only a decrease of performance of 5-12% compared to the monolingual setting when using machine translation in such circumstances (Savoy and Braschler 2019). Bonet et al. (2018) obtained good results in an academic search engine in psychology using specialized dictionaries to automatically translate the queries (average adequacy of 1.4 out of 2), but the experiments showed that machine translation applied to the documents obtained better results in retrieval (Stiller et al. 2019).

The lack of use of machine translation in digital libraries could be explained by the translation ambiguity and the insufficient lexical coverage, which are considered to be among the most prominent problems in MLIR (Peters et al. 2012). These issues are especially relevant in digital libraries, where queries and metadata are usually very limited in terms of context — which also makes language detection, when required, harder. The domain is also very specific to apply general translation tools, and the resources are limited to create the appropriate language resources required. As noted by Stiller (Stiller and Petras 2016), machine translations should be used with care especially for highly specialized and curated content.

There are also discrepancies in the literature regarding the best strategy to adopt — document translation, query translation or both. The literature shows that document translation is consistently more effective (Savoy and Braschler 2019; Stiller et al. 2019; Peters et al. 2012), however Savoy and Braschler (2019) report that the performance differences between query translation and document translation approaches vary greatly depending on the query. They suggest taking advantage of both translation models in a hybrid approach: in an experiment run using English as a pivot language, the results obtained outperformed other strategies. This strategy is also more scalable than document translation when the number of different languages is considerable, as is the case in Europeana. On the other hand, in the Cultural Heritage in CLEF (CHiC) lab

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

(Petras et al. 2013), where the Europeana metadata was used, the best result in the multilingual task was obtained by manually translating the topics (the translations were provided by the conference). The organizers suggested though that more participants are necessary in order to provide comparative analyses.

Approach

The pivot strategy in a cross-lingual system requires the translation of queries and documents to a common language. As mentioned in the introduction, we have chosen a pivot approach using English, in line with the Europeana Multilingual Strategy and findings from the related work (see section 2) as well as the fact that English is the most present language in the collections we have worked with. The approach is the one shown in fig. 1, where the documents are translated and indexed offline, while the queries are translated in real time. If not provided, it is necessary to identify first the source language of queries and documents using language detection tools and/or other signals (e.g., language of the interface of the user issuing a query). The original query and its translation is then used to search the part of the collection in the original language and the translation of other parts to English, respectively. It is also possible to search using only the English translations of queries and documents, in a regular monolingual set up.

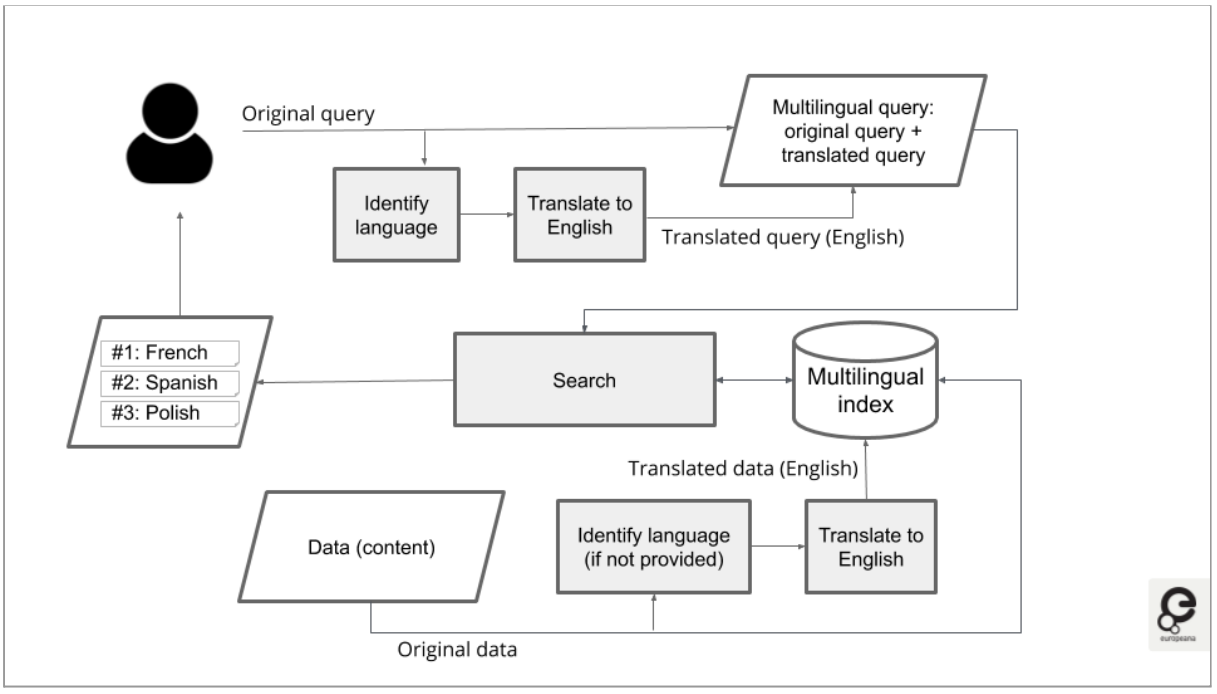


Figure 1. Cross-lingual search approach using English as pivot language.

For the purpose of the experiment we have translated offline a set of documents and queries described in the next section. All the documents were already provided with tags indicating their original language. We set up an Apache Solr search engine

following the same configuration used in Europeana, with the standard BM25 search algorithm, and with standard term normalization processes for the different languages (i.e. tokenizing, lower case normalization, and stemming). We indexed the documents using fields with language tags (e.g. fulltext de, fulltext en), keeping original and translation in the same record. We have assumed that the language of the portal used by the user reflects the language of the query, for example Spanish for https://www.europeana.eu/portal/es. The original query and its translation is then used to search in the collection in the original language and its translation respectively. The translated query is also used to search in the whole collection of translated documents.

As indicated earlier we have used the eTranslation service provided by the European Commission. Other well-known services can be found online, and actually eTranslation is not (yet) fully fit for synchronous translation, as translating queries live on a portal would require. Yet, the European Commission claims it achieves good performance and is committed to its development. In addition, it is intended as a free, secure service for public bodies, which can be appealing for CH institutions, especially in Europe. In order to get the translations (which are received asynchronously), we have used a Java client available at: https://github.com/nfreire/europeana-etranslation-research

The use of a cross-lingual approach, compared to a monolingual one, is expected to improve the search experience by offering more results that are relevant to the query in languages different from the language in which the query is written. In order to assess this, as well as its impact on precision, we need to compare the results obtained in terms of relevance. Directly evaluating them by human assessment would be the preferred option, but it requires extensive resources. Therefore, for this preliminary experiment, we have defined a less ambitious, but testable, 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 (be they or not actually relevant for the user). If that is not the case, the risk of applying a multilingual system with our current configuration and experiment set up is high. 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. We then computed the precision and recall of s_{qt} with respect to s_{qo} , defined as:

$$\text{Precision} = |s_{qo} \cap s_{qt}| \; / \; |s_{qt}|$$

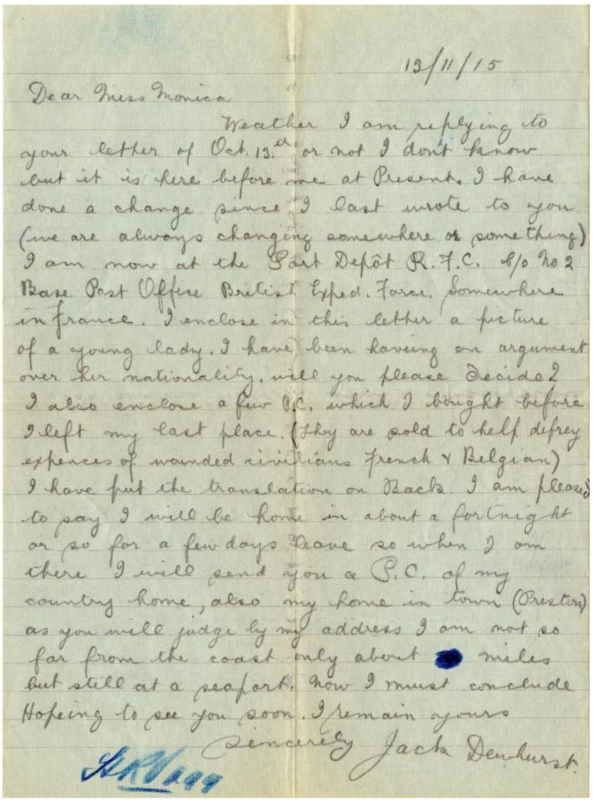
$$\text{Recall} = |s_{qo} \cap s_{qt}| \; / \; |s_{qo}|$$

We additionally calculated the number of new transcriptions retrieved when using q_t in the whole corpus of English transcriptions (translated or not). Finally, we manually assessed the quality of translation of the queries, as they play a major role in the cross-lingual system.

Data Acquisition and Processing

Contents

The collection used is a sample of 18,257 transcriptions of documents[6] from the Europeana 1914-1918 thematic collection[7], obtained from the Transcribathon crowdsourcing platform[8] as part of the Enriching Europeana project[9]. This collection includes many World War I related objects contributed by members of the public all over Europe, like soldiers' diaries or letters, for which manual transcription is needed to facilitate access. Each transcription is done in the original language of the web resource (see example in fig. 2).



13/11/15
Dear Miss Monica
Weather I am replying to your letter of Oct. 13th or not I don't know but it is here before me at Present. I have done a change since I last wrote to you (we are always changing somewhere or something) I am now at the Port Depôt R.F.C. C/o No 2 Base Post Office British Exped. Force. Somewhere in France. I enclose in this letter a picture of a young lady. I have been having an argument over her nationality. will you please Decide? I also enclose a few P.C. which I bought before I left my last place. (They are sold to help defrey expences of wounded civilians French & Belgian) I have put the translation on Back. I am pleased to say I will be home in about a fortnight or so for a few days leave so when I am there I will send you a P.C. of my country home, also my home in town (Preston) as you will judge by my address I am not so far from the coast only about [illegible] miles but still at a seaport. Now I must conclude Hoping to see you soon. I remain yours
Sincerely Jack Dewhurst

Transcription: 13/11/15 Dear Miss Monica Weather [whether] I am replying to your letter of Oct. 13th or not I don't know but it is here before one at Present. I have done a change since I last wrote to you (we are always changing somewhere or some thing) I am now at the Port Depôt R.F.C. C/o No 2 Base Post Office British Exped. Force. Somewhere in France. I enclose in this letter a picture of a young lady. I have been having an argument over her nationality. will you please Decide ? I also enclose a few P.C. which I bought before I left my last place. (They are sold to help defrey expences of wounded civilians French & Belgian) I have put the translation on Back. I am pleased to say I will be home in about a fortnight or so for a few days leave so when I am there I will send you a P.C. of my country home, also my home in town (Preston) as you will judge by my address I am not so far from the coast only about [number blotter out] but still at a seaport. Now I must conclude Hoping to Jack Dewhurst [includes envelope]

Letter DCLA/RDFA1.05.036 from Jack Dewhurst. 1915-11-13. Dublin City Library and Archive.

Figure 2. Transcription of a 1915 letter from the Europeana 1914-1918 collection.

[6] These correspond to web resources (digital representations) in the Europeana Data Model. Note that one object in the Europeana Collection may be represented by more than one web resource, but in the context of this work, a document corresponds to the transcription of an image (web resource)

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

[8] https://transcribathon.com

[9] https://pro.europeana.eu/project/enrich-europeana

From this sample we removed 18 transcriptions that lacked indication of the original language[10], so we eventually used for the experiment 18,239 transcriptions in 17 different languages (see Table 1).

Language tag	Transcriptions	Translated to en
de	9300	9151
fr	1669	1659
it	992	973
ro	578	577
nl	455	454
el	364	356
lv	226	226
bs	215	0
cs	90	90
da	90	90
sl	7	7
hu	3	2
es	2	2
pl	2	2
sk	2	2
hr	1	1
TOTAL (non-en)	**13996**	**13592**
en	4243	0
TOTAL	**18239**	**13592**

Table 1. Original language of the transcriptions, and number of transcriptions successfully translated to English.

The contents were pre-processed to unescape HTML characters and to change the original language tag from 'gr' to 'el', in order to be conformant with the ISO standard and recognized by the eTranslation service[11]. The carriage returns contained in the text were not removed as in some cases they are used to split different sections. Future work should be done to assess the impact of the use of these carriage returns in the quality of the translations obtained.

After removing the transcriptions already in English as its original language, 13,996 transcriptions in different languages were sent to the eTranslation service to translate to English. We received errors for 404 of them (2.9%), 215 because the language (bs, i.e.

[10] These are caused by the data coming from an older version of the transcription platform. The issue has been solved already in the transcription platform.
[11] This issue is also being solved now in the transcription platform, which generated the wrong tag.

Bosnian) is not supported by the service, and 189 because the text is too long[12]. The latter problem can probably be solved by sending the requests to the service in binary mode, which is planned for future work. As a result, we obtained 13,592 translations for the experiment, with an average response time of 1.11 seconds per translation.

Queries

Regarding the queries, we have used a sample of queries issued on the 1914-1918 collection from the Europeana Portal. Using Google Analytics Reporting API[13], we obtained a total of 90 queries from those issued by users in that collection between the 1st of January 2019 and the 1st of September 2019. Due to technical limits of the Google Analytics API, we could not obtain a larger sample of queries. We extracted the largest amount of usage data available ('large' sampling option) from what is available in Google Analytics (which gathers all usage of Europeana Collections) but the limit in the Standard Analytics API is 500k sessions, and apparently most of those sessions do not contain queries at all, or they do not contain queries to the specific collection we are interested in. As an alternative, we could work with several smaller samples to avoid the automatic sampling made by Google as much as possible. That approach is left for future work.

From the 90 queries, 22 were issued from the portal in English, so only 68 of them were sent to the eTranslation service. All of them were successfully translated, and therefore they were included as part of the experiment (see Table 2).

Language tag	Queries	Translated to English
it	29	29
fr	13	13
de	12	12
pl	6	6
es	3	3
nl	2	2
ro	2	2
cs	1	1
TOTAL (non-en)	**68**	**68**
en	22	0
TOTAL	**90**	**68**

Table 2. Original language of the queries (assuming it is the same as the language of the portal), and number of queries successfully translated to English.

[12] eTranslation makes available another interface that allows long texts to be submitted for translation, however at the time of this experiment, we did not have an implementation for invoking it. This implementation is part of our future work.

[13] https://developers.google.com/analytics/devguides/reporting/core/v4

Results

From the 68 original queries considered, 37 of them had zero results so we could not use them to evaluate. For the remaining 31 queries (in 5 languages), we obtained the precision and recall of s_{qt} compared to s_{qo} for the results retrieved. The results are shown in table 3.

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

Table 3. Precision and Recall obtained when comparing the list of retrieved results for a query and its translation, and additional results retrieved when searching with the translated query in all the translated transcriptions in the corpus.

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}. In this scenario, the strategy adopted to merge the results, those in the original language and those coming from translations, should be carefully designed in order to prevent a negative impact on the user experience: in our case, on average 337 of the transcriptions found are not retrieved when using the original query. Even though some of those results could be actually relevant due to the existence of synonyms that are not matched by (non-semantic) search in a single language, the following subsection gives some insights, which shed a rather negative light on its impact: these results are more likely to be noisy. This could be however compensated in some cases by the new transcriptions found. When using s_{qt} 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 (see fig.3). Of course, the real impact on the user should be properly tested by using relevance assessment, which is planned as future work.

original query	language	translated query	results original query	results translated query	new docs retrieved thanks to translation
domov	cs	home	2	1529	1527
Bernhard Stiens	de	Bernhard Stiens	16	21	8
cimitero	de	ciemitero	0	0	0
eastern front	de	Eastern front	345	1272	955
lagazuoi	de	lapiönoi	0	0	0
letters	de	letters	25	1935	1913
nova vas	de	Nova vas	4	31	29
Pinsk	de	Pinsk	1	1	0
podgora	de	podgora	1	7	6
Rokitno	de	Roitno	0	0	0
san elia	de	San elia	40	49	16
Talies	de	Talies	0	2	2
women	de	women	4	255	251
antonio sordi	it	Antonio Deaf	12	25	14
Asiago	it	Asiago 1)	4	2552	2548
avion	it	Avion	0	4	4
bini cima	it	Bini top	3	837	835
celle lager	it	lager cells	2	56	56

Fig. 3. Example of original and translated queries, and number of results retrieved. Noticeable examples of good and bad translations are highlighted in green and red respectively. The right column shows the impact of translation on the number of results retrieved.

Issues

One of the first issues we found is that, in a considerable amount of cases, the language of the query did not match the language of the portal (i.e., our assumption that its language is the language of the portal is wrong). For example, the query 'Eisenbahn', which is German, was found among the French queries, and the English query 'women' was found among German, French, Italian and Dutch queries. That happened in 17, maybe 18[14], of the 68 queries. In most of the cases the queries were written in English while the language of the portal was different. The eTranslation service did not complain at all about those translation requests, and in most of the cases it kept the same text than in the original query, except for two cases where it changed the translation to non existing words. The eTranslation service also kept the typos found in such 'foreign languages', hence practically leaving these queries untouched. However, even when the resulting query in English may be correct, this issue negatively affects the evaluation conducted, as the language of the transcriptions searched to build s_{qo} does not match with the language of q_o.(we only obtained search results in the original language in 4 of them). Note that in a real scenario, even if the translation is correct, the wrong identification of the language in which the query is issued is likely to also have a negative impact in the search because the original query may return no results sincethe query term does not exist in the target language, or even wrong results from the set of documents in the original language (because of say, wrong normalization process being applied to the query).

The next issue is that of the quality of the translation, which of course has an impact on the search results. In some cases the translated text contains spurius characters (e.g. 'Asiago' translated to 'Asiago 1)') or is wrong (e.g. 'Trăiscă Averescu' from Romanian,

[14] Note of course that we encountered here the usual difficulties in judging the intention of queries, as 'avion' queried on the Italian portal may well be an Italian user using a French word to higher the odds of finding documents about planes in collections from France, or a user interested in finding documents in any language about the Avion village in Northern France, where intense WWI action took place.

translated to ‘Torisca’Averscu’). A more subtle problem is the ambiguity of the language and its impact in translations, which is especially relevant in the queries because of the lack of context. For example, ‘carnet de route’ is correctly translated from French as ‘journey log’ in the transcriptions, however the query ‘carnet’ is translated as ‘notebook’, which is not wrong but misses the information need that was only partially expressed in the query, so no results are retrieved in that case. The selection of a high quality translation service is important to reduce these errors, but additional techniques would be needed in order to deal with the disambiguation of the queries.

A special case of wrong translations are those where named entities are involved. It has been previously estimated that around half of all the queries in the Europeana Portal are about named entities, and 42 of the 68 queries in this experiment are (or include) named entities (62%), including those with typos (e.g.'san elia' probably comes from Antonio Sant'Elia), and those composed of common and proper nouns (e.g. 'Celle lager', where 'lager' in German is camp, and 'Front Strzelno'). The translation of those entities as common words will lead to bad search experiences because the user can not find what she is looking for. This is the case for example for the queries ‘Antonio Sordi’ and ‘Fogliano’ translated from Italian to English as ‘Antonio Deaf’ and ‘sheet’ respectively[15]. The problem seems especially hard to resolve as the named entities present and queried in the World War I context are very specialized (less-known authors, small villages) and sometimes referred to incompletely, such as "Tonale" in Italian, which in general would belong to the music domain but here refers to a pass (Passo del Tonale). Handling such cases would require getting access to very fine-grained data and careful consideration of the context of the query and the collections.

In the end, we analyzed the quality of the translation for 43 queries: out of the total 68 queries, we removed those were the user's intention was not clear to us (4) and those with typos or wrong language assignment (21), for which the service was thus not given appropriate input. In 37 cases the query was an entity that had to be left unchanged (e.g. 'Bernhard Stiens' is supposed to be left unchanged, while 'Italia' must be translated to 'Italy'). The translation 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 query, the translation service did it correctly in 5 cases (83%).

Finally, we also found categories of issues where quality of translation plays a role but is combined with the impact of other processes in the whole indexing and search workflow. For example, the bad results with the named entities can be worse when those entities are not analyzed differently from the regular text by the text processing components of the search engine. In those cases, (and in fact even when the translation is correct), the search may retrieve non-relevant results because those words are treated differently in the original and translated language. This happens for example for the query in Italian ‘Italia’ translated to ‘Italy’: ‘Italia’ (Italy) and ‘italiano’ (Italian) are reduced to one common string after stemming, however in English ‘Italy’ and ‘Italian’ are reduced to different strings. As a consequence, the query may retrieve different information depending on the language.

[15] Note that the 'Fogliano' case combines two issues: it has been translated to 'sheet' in the context of transcriptions, but the query translation has left it unchanged (rightly).

Conclusions

With relatively low effort we could set up an experiment to evaluate a basic cross-lingual approach where we use English as pivot language. In order to build a real cross-lingual system though, we would still need to include the automatic redirection of the queries to the proper contents according to the language, and, depending on the UX design, merge or sort the retrieved results coming from different languages, both feasible things to do. A synchronous translation service would make the process easier, and it is definitely required to translate the queries in real time, but independently of that feature, we can rely on the eTranslation service in general as most transcriptions and all the queries were translated with no technical issues.

The experiment conducted is limited for two main reasons: the evaluation without human assessments of relevance, and the small number of queries and languages used. Therefore the quantitative results obtained should be only considered as a measure of the impact of translations in a cross-lingual system when compared with the monolingual version. Still, the experiment already shows the risk of adopting such a system, and, combined with the qualitative results, contributes with a preliminary measurement (under relevance assumptions not yet fully tested) of search effectiveness . However, the analysis of the results shows already some of the benefits and issues we face with this type of systems.

The experiment 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 well the problem of performing query translation in a context like Europeana: 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, and smaller than the number of cases in which it is given misleading input.

Our analysis of the quality of translations mainly focused on the translation of queries. We anticipate that most issues we found, however, apply to the translation of transcriptions as well. For example, issues related to the wrong translation or normalization of entities will also apply when those entities are in the transcriptions. Translation services may work better for transcriptions, as there is more context available. However this is likely not to solve our case in general, as it relies on query translation to bring more results for all non-English queries. Only English queries would be better served, if we disable query translations. This work shows then that, without addressing the issues found, especially the issues related to the identification of the query language and the handling of entities, the drawbacks of a multilingual system in a CH domain could easily exceed its benefits.

Future work

The first option to continue the work presented here would be to confirm the observations obtained by testing with more queries, as the sample we have used is relatively small (and limited in the languages represented). We could source more queries from Europeana's Google Analytics report or our own logs.

The second regards the approach to cross-lingual search. In order to assess the impact of translation, we have compared the results obtained in a monolingual system with those obtained in a cross-lingual one that relied on translating queries and collections into a common language: English. Different approaches and techniques could be adopted for cross-lingual search though. We should conduct similar experiments to assess the feasibility and effectiveness of other approaches in Europeana, like translating queries or contents to all or some of the official languages (even though document translation is not easily scalable, it could compensate for the need to translate the documents in real-time for displaying purposes). We could also use different tuning for the search algorithm, for example testing exact full string matching between queries and text (removing part of the normalization process), which would have removed a lot of the noise coming for all the translations that added "Mr" to the query. Once we have progressed on the inclusion of a cross-lingual search routine in a more complete cross-lingual system (e.g. with merging and ranking of search results coming from different languages), we should proceed with a more holistic evaluation of the effectiveness of that system. Additionally, the experiments should be extended not only to the retrieval of transcriptions, but also to other content (for example newspapers' OCR) and to the retrieval of metadata in order to assess its applicability to all the data in our collection. In the latter case, additional issues could arise due to the usual lack of context in the metadata, the possibility of having metadata in multiple languages in the same record, the lack of language tag in some cases, and the existence of multiple entity-based fields (e.g. creator).

Then, we have compared the cross-lingual and monolingual retrieval in terms of the transcriptions that match the queries, instead of the transcriptions that are actually relevant for them. In our analysis we have seen that the actual relevance of the retrieved results is affected not only by the quality of the translation, but also by the ambiguity of the translations, and the differences in the normalization of the different languages. We should run experiments that evaluate and compare the actual relevance of the retrieved results, given the same search engine and normalization processes, in order to better assess the impact of the intrinsic gap between languages in retrieval. We could also explore different metrics, such as the ones that account for ranking (e.g., nDCG).

In terms of performance (efficiency), in a cross-lingual information retrieval system the collection to search is larger than in the monolingual version, so it is expected that the performance will be affected. Therefore, we should include measures of performance in our experiments in order to assess the impact it has in the user experience.

Finally, the quality of the translation service used also has a big impact on the results obtained in a cross-lingual system. If different translation services can be used, their quality and performance could be analyzed independently given a sample of the queries and data in Europeana. In this analysis, it is important to take into account that it is estimated that half of the queries issued from the Europeana Portal are named entities, and most of them should be left unchanged in the translation. This is where the eTranslation service produced the most incorrect translations. We should therefore carefully consider if the strategy to follow could be different for those queries. Additional techniques like controlled vocabularies and named entity recognition tools may be needed, as indicated by Stiller and Petras (2016), 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, as noted by Stiller et al. (2010), could improve the results. We could also check the input given to the translation service has been applied. In particular we have kept all carriage returns in the transcriptions (reflecting the visual ordering of words in the original handwritten content) but it may have impacted the quality of translations.

Bibliography

Agosti, M., Fabris, E., Silvello, G.: On synergies between information retrieval and digital libraries. In: Proceedings of the Italian Research Conference on Digital Libraries, 2019. pp. 3–17. Pisa, Italia (2019)

Chen, H.: Digital library research in the US: an overview with a knowledge management perspective. Program: Electronic Library and Information Systems 38 (3), 157–167 (2004)

Diekema, A.R.: Multilinguality in the digital library: A review. The Electronic Library 30 (2), 165–181 (2012)

Dolamic, L., Savoy, J.: Retrieval effectiveness of machine translated queries. Journal of the American Society for Information Sciences & Technology 6, 2266–2273 (2010)

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)

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

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. The Electronic Library 33 (1), 136–51 (2015)

Oudenaren, J.V.: The World Digital Library. Uncommon Culture 3 (5/6), 65–71 (2012)

Peters, C., Braschler, M., Clough, P.: Multilingual Information Retrieval: From Research to Practice. Springer, Heidelberg, Germany (2012)

Petras, V., Bogers, T., Toms, E., Hall, M., Savoy, J., Malak, P., Pawłowski, A., Ferro, N., Masiero, I.: Cultural Heritage in CLEF (CHiC) 2013. In: Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages, 2013. pp. 192–211. Valencia, Spain (2013)

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, Switzerland (2019)

Stiller, J., Gade, M., Petras, V.: Ambiguity of queries and the challenges for query language detection. In: CLEF 2010 LABs and Workshops (2010). Padua, Italy (2010)

Stiller, J., Gade, M., Petras, V.: Multilingual access to digital libraries: The Europeana Use Case. Information. Wissenschaft & Praxis 64 (2-3), 86–95 (2013)

Stiller, J., Petras, V.: Best practices for multilingual access. Tech. rep., Europeana (2016), available At https://pro.europeana.eu/post/best-practices-for-multilingual-access

Stiller, J., Petras, V., Luschow, A.: CLUBS Project (Cross-Lingual Bibliographic Search). M5.3 Final Evaluation (2019), version 1.0

Vassilakaki, E., Garoufallou, E.: Multilingual digital libraries: A review of issues in system-centered and user-centered studies, information retrieval and user behavior. The International Information & Library Review 45, 3–19 (2013)