Can machine translation assist scholars who have limited English proficiency with searching?

Lynne Bowker, University of Ottawa, Canada

Indexing languages constitute formalized languages designed and used to describe the subject content of documents for information retrieval purposes. In addition many scientific databases include a less controlled means of describing the subject content of documents: author-supplied key words. Gil-Leiva and Alonso-Arroyo (2007) conducted a detailed study of 640 scientific articles that possess author keywords and are indexed in databases. They found that author-supplied keywords have an important presence in the database descriptors studied: nearly 25% of all keywords appeared in exactly the same form as descriptors, while another 21% have undergone a normalization process but are still detected in the descriptors. In total, about 46% of the author keywords appeared in the same or a normalized form as descriptors leading Gil-Leiva and Alonso-Arroyo (2007, p. 1181) to posit that author keywords are a valuable source of information for indexing articles and for information retrieval.

In the last half-century, English has emerged as the dominant language of scholarly communication despite the fact that only about 6% of the world's population are Anglophones (Corcoran, 2015). What does this mean for scholars who are not native English speakers and may have only Limited English Proficiency (LEP)? Many study to achieve a high level of proficiency, while others may engage professional translation or editing services. Such options may be viable for well-funded scholars but may pose a challenge to LEP scholars from developing countries. In such cases, these LEP scholars may turn to cheaper alternatives, such as machine translation (MT), to help them engage in the scholarly communication process. In its 2013 *Trend Report,* the International Federation of Library Associations and Institutions (IFLA) lists MT as one of five key high-level trends in the global information environment. We are currently investigating the role of MT in various aspects of the scholarly communication process, including its potential for helping LEP scholars to understand scientific articles written in English, as well as to disseminate the results of their own research in English. However, our goal in this paper is more modest; we undertake a small pilot study to examine whether MT can help LEP scholars at the stage of searching. Kit and Wong (2008) observe that "MT quality is far from publishable," but they go on to suggest that MT may be good enough to serve many translation demands for the purposes of information access. While they did not test this application of MT, we will do so here in a small pilot experiment.

If we assume that many LEP scholars first learn about their domain through their own language and begin by accessing scholarly articles in that language, how can they take the next step of looking for comparable or related material in English? Can free online MT systems (e.g. Google Translate) help with searching?

Corpus and Methodology

Some non-English-language journals provide abstracts and keywords in English. However, others, such as national journals or journals run by individual university departments – often by and for their own graduate students who are therefore new to the domain – may not. For this pilot study, we identified two Information Science journals in which the articles, abstracts and keywords, are provided only in Spanish. The first, entitled *e-Ciencias de la Información*, is a biannual online academic journal launched in 2011 and published by the School of Library Studies and Information Sciences at University of Costa Rica. It aims to promote research in a wide range of LIS-related areas including information systems, library studies, information policy, research methodology, and archival studies. The second is entitled *Métodos de Información*. Online since 2010, this biannual journal of the School of Library and Documentation Studies of Valencia addresses subjects related to libraries, archives and documentation centres.

For each journal, we randomly selected one article from each issue published in the past five years (2013-2017) for a total of twenty articles. From each article, we extracted the list of the author-supplied keywords, which were copied into a spreadsheet and sorted alphabetically. After eliminating duplicates, we were left with a list of 71 Spanish keywords which we then translated into English using Google Translate. While other free online MT systems are available, we selected Google Translate because Buitrago Ciro (2018), who surveyed 48 LEP scholars at a university in Colombia, found that of the 42 who actively use MT, over 97% use Google Translate.

Next, we used the translated keywords to search the Library, Information Science & Technology Abstracts (LISTA) database. Although there are other bibliographic databases available for the LIS-domain, Vinson and Welsh (2014) report that LISTA has one of the broadest ranges, covering a wide variety of subjects pertaining to library and information science such as librarianship, bibliometrics, cataloging and classification, reference, online information retrieval, and information management. As Vinson and Welsh (2014, 124-125) emphasize, resources are a crucial consideration for libraries, and not every library can afford multiple databases. At institutions with limited means, which may be the case in developing countries, LISTA may well be the database of choice for its breadth of coverage and access to a variety of full-text materials. In addition, LISTA, without full-text availability, is offered free by EBSCO (www.libraryresearch.com). We therefore elected to use the LISTA database for this pilot study.

Using the advanced search option, we restricted our searches in the following ways:

- Publication type: Academic Journals
- Document type: Article
- Language: English
- Field: SU Subject Terms (Performs a keyword search of subject headings, companies, people, and author -supplied keywords for terms describing a document's contents).

Results

Of the 71 translated keywords, 37 (52%) returned relevant search results (i.e., articles on a similar topic to the corresponding Spanish article from which the original keywords were taken), while 34 (48%) did not. These 37 productive keywords appear to have been well translated.[1] Among the 34 translated keywords that did not return any results, 23 (68%) appear to be appropriately translated but simply not in alignment with the descriptors used in LISTA. See Table 1 for examples.

Keywords that returned results	Keywords that did not return results
• *acceso a la información*: access to information • *alfabetización informacional*: information literacy • *Internet de las Cosas*: Internet of Things • *minería de datos*: data mining • *recuperación de la información*: information retrieval	• *comunicación científica*: scientific communication • *historia de la lectura*: history of reading • *libro impreso*: printed book • *patrimonio documental*: documentary heritage • *políticas de conservación*: conservation policies

Table 1. Examples of translated keywords that did and did not return results.

1. This assessment was made by the present author, who is a certified translator.

For the remaining 11 (32%) keywords that did not return results, it appears that translation-related problems stemming from orthographic variation, synonymy, or differing syntactic preferences and semantic field coverage have interfered with the information retrieval process.

Discussion

From our list of translated keywords, we can see that 'ebook' (*libro electrónico*) and 'bibliographic data bases' (*bases de datos bibliográficas*) use different orthographic variants than the LISTA descriptors, which are the full-form term 'electronic book' and 'bibliographic databases' (where 'database' is written as a single word). If the knowledge organization system (KOS) were more robust and able to handle such variants, then these translated keywords would have generated relevant results also.

Synonymy exists when two or more terms refer to the same concept. There were three cases where Google Translate translated a Spanish term into English using a synonym for the descriptor, rather than the descriptor term. For instance, *competencias informacionales* was translated as 'information competences' rather than as its synonym 'information skills' (which corresponds to a LISTA descriptor). Again, if the KOS could be expanded to better handle potential synonymic translations, then machine translated keywords could potentially generate more positive results.

Four of the translation problems result from the different syntactic structures most commonly used by Spanish and English. In all four cases, Google Translate has produced a literal translation that mirrors the underlying Spanish preference for prepositional phrases. The resulting translations are all grammatically and semantically correct; however, the more idiomatic way of expressing these structures in English is to use pre-modification. Moreover, in all four cases, if pre-modification had been used, the resulting term would have returned results from LISTA, as illustrated in Table 2. If a KOS could be augmented to recognize these very common and often predictable types of MT "errors" that arise from differing syntactic preferences between languages, then machine translated keywords could be more productive for information retrieval.

Original Spanish keyword	English translation by Google	Preferred English structure contained in LISTA descriptor
arquitectura de la biblioteca	architecture of the library	library architecture
representación del conocimiento	representation of knowledge	knowledge representation
sociedad de la información	society of information	information society
utilización del espacio de bibliotecas	use of library space	library space utilization

Table 2. Examples of differing syntactic structures

Finally, it is well known that languages divide the world up differently such that the semantic space referred to by a single term in one language (L1) might be covered by two different terms in another language (L2). In such a case translating from L2 to L1 is simple, but translating from L1 to L2 requires making a choice. For instance, the Spanish term *revista* can be translated as either 'journal' or 'magazine', and Google Translate chose 'magazine', which is the wrong choice in this particular context.

Similarly, the Spanish term *deontología* can be translated as 'deontology' or as 'ethics', and Google Translate chose to translate the keyword *deontología profesional* as 'professional deontology' rather than as 'professional ethics', which corresponds to a descriptor in LISTA. Because these choices are so often context dependent, it would be challenging for a KOS to deal with this issue. However, in our experiment, fewer than 20% of the translation issues fell into this category.

Concluding remarks

While keeping in mind that this pilot experiment was conducted on a very small scale—using just two journals, 71 keywords, one MT system, one language pair, and one bibliographic database—the results nonetheless appear promising. Globally, only 11 out of 71 (15%) of the author-supplied keywords were translated in a way that led to no results being retrieved from the LISTA database. A much higher proportion of the keywords did not generate results because they did not align with the descriptors used in LISTA, rather than because the keywords were poorly translated. For LEP scholars, MT appears to be a viable tool for helping with information retrieval, and if KOSs could be augmented to better handle "predictable" translation-related challenges such as orthographic variation, synonymy and some types of syntactic variation, then MT could prove to be even more useful in this context. However, more work is needed to see whether MT can help LEP scholars to access and understand the content of the retrieved documents.

References

Buitrago Ciro, Jairo (2018). La littérature des sciences de la santé et la traductique: Quel est le rôle des bibliothécaires? Presented at the 2018 Canadian Health Libraries Association Annual Conference, St. John's, Canada, 15-18 June 2018.

Corcoran, J. 2015. *English as the International Language of Science: A Case Study of Mexican Scientists' Writing for Publication*. PhD dissertation, University of Toronto, Canada.

Gil-Leiva, I. & Alonso-Arroyo, A. (2007). Keywords Given by Authors of Scientific Articles in Database Descriptors. *Journal of the American Society for Information Science & Technology*, 58(8), 1175–87.

International Federation of Library Associations and Institutions (IFLA). (2013). *Riding the Waves or Caught in the Tide? Navigating the Evolving Information Environment. Insights from the IFLA Trend Report*. The Hague: IFLA.

Kit, C. & Wong, T. M. (2008). Comparative Evaluation of Online Machine Translation Systems with Legal Texts. *Law Library Journal*, 100(2), 299–321.

Vinson, T. C. & Welsh, T. S. (2014). A Comparison of Three Library and Information Science Databases. *Journal of Electronic Resources Librarianship*, 26(2), 114–126.

Journals used as a source for the corpus

e-Ciencias de la Información: https://revistas.ucr.ac.cr/index.php/eciencias

Métodos de Información: http://www.metodosdeinformacion.es/mei/index.php/mei