Machine translation and author keywords: A viable search strategy for scholars with limited English proficiency?

Lynne Bowker
University of Ottawa
lbowker@uottawa.ca

ABSTRACT
Author keywords are valuable for indexing articles and for information retrieval (IR). Most scientific literature is published in English. Can machine translation (MT) help researchers with limited English proficiency to search for information? We used two MT systems (Google Translate, DeepL Translator) to translate into English 71 Spanish keywords and 43 French keywords from articles in the domain of Library and Information Science. We then used the English translations to search the Library, Information Science & Technology Abstracts (LISTA) database. Half of the translated keywords returned relevant results. Of the half that did not, 34% were well translated but did not align with LISTA descriptors. Translation-related problems stemming from orthographic variation, synonymy, differing syntactic preferences, and semantic field coverage interfered with IR in just 16% of cases. Some of the MT errors are relatively "predictable" and if knowledge organization systems could be augmented to deal with them, then MT may prove even more useful for searching.

Keywords
Machine translation, author keywords, information retrieval, knowledge organization systems.

INTRODUCTION
Knowledge organization systems (KOSs) aid information discovery by modeling the underlying semantic structure of a domain, providing a semantic road map of individual fields and the relationships among and across fields, and relating concepts to terms. Indexing languages are formalized languages used to describe the subject content of documents for information retrieval (IR). Many scientific databases also include a less controlled means of describing document content: author keywords.

81st Annual Meeting of the Association for Information Science & Technology | Vancouver, Canada | Nov. 10 - 14, 2018

Gil-Leiva and Alonso-Arroyo (2007) studied 640 scientific articles that have author keywords and are indexed in databases and found that author 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. Overall, about 46% of the author keywords appeared in the same or a normalized form as descriptors, leading the researchers to posit that author keywords provide valuable information for indexing articles and for IR.

English has become the language of scholarly communication even though only 6% of the world's population is Anglophone (Corcoran, 2015). What does this mean for scholars with Limited English Proficiency (LEP)? Many study to achieve proficiency, while others hire professional translators/editors. These expensive options may pose challenges for LEP scholars from developing countries who may seek cheaper alternatives, such as machine translation (MT), to help them engage with scholarly literature. The International Federation of Library Associations and Institutions (IFLA, 2013) identifies MT as a key high-level trend in the global information environment. Anazawa et al. (2013) report that MT is useful for "gisting" scientific literature (i.e., getting a general understanding of a text), especially if the reader is a domain expert. Similarly, Kit and Wong (2008) note that while "MT quality is far from publishable," it can be used for gisting specialized literature. They also suggest it may be good enough for database access purposes, but they do not test this idea. MT systems may translate entire sentences more easily than isolated words presented out of context (i.e., because parsing is difficult). Therefore, while MT is useful for gisting scientific articles, it is unclear whether it can usefully translate individual keywords for database searching. This is an important question, however, because scholars need to be able to locate articles before they can (gist) read them.

We undertook a pilot study to see if MT can help with database searching. If we assume that many LEP scholars first learn about their domain through their own language (e.g. as students) and begin by accessing scholarly articles in that language, how can they take the next step of looking for pertinent material in English? Can online MT systems help with searching?

KEYWORDS THAT RETURNED RESULTS	KEYWORDS THAT DID NOT RETURN RESULTS
Spanish-English (Google Translate)	
• *acceso a la información*: access to information • *alfabetización informacional*: information literacy • *Internet de las Cosas*: Internet of Things • *minería de datos*: data mining	• *historia de la lectura*: history of reading • *libro impreso*: printed book • *patrimonio documental*: documentary heritage • *políticas de conservación*: conservation policies
French-English (DeepL Translator)	
• *collection numérique:* digital collection • *données ouvertes:* open data • *inclusion numérique:* digital inclusion • *système d'information:* information system	• *cycle de vie de la donnée:* data life cycle • *données de recherche:* research data • *gouvernance des données:* data governance • *risque numérique:* digital risk

Table 1. Examples of well-translated keywords that did and did not return results in the LISTA database.

CORPUS AND METHODOLOGY

Some non-English journals provide abstracts and keywords in English. However, others may not, such as national journals or those run by university departments (often for graduate students new to the field). For this pilot study, we identified two Library and Information Science (LIS) journals that provide the articles, abstracts and keywords only in Spanish, and one that provides them only in French. For the Spanish journals, we randomly selected one article from each issue published in the past 5 years (20 articles). The French journal is new with just four issues, so we randomly selected 3 articles from each issue (12 articles). From each article, we extracted the author keywords to a spreadsheet and sorted them alphabetically. After eliminating duplicates, we had 71 Spanish keywords and 43 French keywords. We translated the Spanish keywords into English using Google Translate and the French keywords using DeepL Translator; both use neural networks (machine learning).

Next, we used the translated keywords to search the Library, Information Science & Technology Abstracts (LISTA) database. While other bibliographic databases are available for this domain, Vinson and Welsh (2014) report that LISTA has one of the broadest ranges, covering a wide variety of LIS subjects. As Vinson and Welsh (2014) emphasize, resources are a crucial consideration, and not every library can afford multiple databases. At institutions with limited means, as 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).

Using the advanced search option, we restricted our searches in the following ways: a) Publication type: Academic Journals; b) Document type: Article; c) Language: English; and d) 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

Spanish-English translations using Google Translate

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. The 37 productive keywords were well translated (as assessed by the present author, a certified translator). Of the 34 translated keywords that did not return results, 23 (68%) were well translated but simply not in alignment with the LISTA descriptors. See Table 1 for examples. 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 IR. Table 1

French-English translations using DeepL Translator

Of the 43 translated keywords, 20 (47%) returned relevant search results, while 23 (53%) did not. The 20 productive keywords were well translated. Of the 23 translated keywords that did not return any results, 16 (70%) are appropriately translated but simply not in alignment with the LISTA descriptors (see Table 1). For the remaining 7 (30%) keywords that did not return results, translation problems related to orthographic variation and semantic field coverage interfered with IR.

DISCUSSION

From our list of keywords translated from Spanish, we can see that 'ebook' (*libro electrónico*) and 'bibliographic data bases' (*bases de datos bibliográficas*) use different orthographic variants than the full-form LISTA descriptors 'electronic book' and 'bibliographic databases' (where 'database' is written as a single word). Meanwhile, a spelling error was produced by DeepL Translator, which translated the French *gestion des données* as 'data managment' (with a missing 'e'). DeepL Translator also had difficulty with the adjective ending in *informationnelle*, translating it literally as 'informational' on four occasions, rather than as 'information' (e.g. 'informational governance', 'informational poverty'). While 'informational' is a legitimate translation of *informationelle*, in these cases it was not the correct choice. Interestingly, there were other cases where 'information' was correctly identified as the right translation (e.g. *sources informationnelles*/'information sources'). If the

Original Spanish keyword	English translation by Google	Preferred structure 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.

KOS were more robust and could handle spelling variants or errors, 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 'information skills' (which corresponds to a LISTA descriptor). No translation errors from French were caused by synonymy. 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 from Spanish to English result from the different syntactic structures most commonly used in these two languages. 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 premodification. Moreover, in all four cases, if premodification 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 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 IR.

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 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. Meanwhile, DeepL Translator translated the French keyword *documentation* by the English keyword 'literature' rather than as the expected 'documentation'. Because such choices are often context dependent, they present a true challenge for a KOS. However, in our study, just 3/18 (17%) of the translation issues were in this category.

CONCLUSION

Keeping in mind that this was a small-scale pilot study—using just 3 journals, 114 keywords, 2 MT systems, 2 language pairs, and one research database—the results nonetheless seem promising. Globally, 50% of the translated keywords returned relevant results. Meanwhile, only 18/114 (16%) of the author-supplied keywords were translated in a way that led to no results being retrieved from the LISTA database, and of these only 3/18 (17%) were context-dependent errors caused by a different semantic field coverage. A higher proportion of keywords (39/114 or 34%) did not generate results because they did not align with the LISTA descriptors, rather than because they were poorly translated. For LEP scholars, MT seems to be a viable tool for helping with IR, 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. Given that MT gisting has already proved valuable for enabling LEP scholars to grasp the contents of the articles once located, it is important for KOSs to help them locate these articles effectively and efficiently in the first place.

REFERENCES

Anazawa, R., Ishikawa, H., Park, M. J., & Kiuchi, T. (2013c). Online machine translation use with nursing literature: Evaluation method and usability. *CIN: Computers, Informatics, Nursing 31*(2), 59-65.

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), 1175–1187.

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 of keywords

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

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

Revue COSSI : Communication, Organisation, Société du Savoir et Information: https://revue-cossi.info/la-revue/editorial

mei

Machine translation systems

DeepL Translator: https://www.deepl.com/translator

Google Translate: https://translate.google.com/