

Accessing and Understanding contents in Portuguese by foreigners in scientific digital libraries: can this methodology be generalized to other languages?

Cláudio MENEZES

University of Brasília

Department of Foreign Languages and Translation (LET)

Campus Universitário Darcy Ribeiro

70919-900 Brasília, DF, Brasil

claudiomenezes@unb.br

Abstract

In digital libraries, remote access to documents has become frequent. Some examples: <http://www.ndltd.org/> and <http://bdt.d.ibict.br/vufind/> in Brazil which allow access to the text of original documents. Since the number of foreign students in universities has increased, there is a need for a service for them. However, the full translation of these documents would be a herculean task. This research identified some obstacles encountered by foreigners to access and understand scientific content and offers a methodology supported by a computer application facilitating its understanding by Francophone students. It can be adapted to any language pairs, including sign languages and braille.

Keywords: automatic summarization, digital libraries, translation, scientific contents

Résumé

Em bibliotecas digitais, o acesso remoto a documentos tem sido a regra. Alguns exemplos: <http://www.ndltd.org/> and <http://bdt.d.ibict.br/vufind/> no Brasil permitindo acesso ao texto dos documentos originais. Como o número de estudantes estrangeiros nas universidades tem crescido, há a necessidade de lhes oferecer um serviço específico. No entanto, a tradução completa desses documentos seria uma tarefa hercúlea. Esta pesquisa identificou alguns obstáculos encontrados por estrangeiros para acessar e compreender conteúdo científico e oferece uma metodologia baseada em uma aplicação computacional facilitando a compreensão por estudantes francófonos. Pode ser adaptada para quaisquer pares de línguas, incluindo a língua de sinais e o braille.

1. Introduction

This paper presents a research aimed to identify the obstacles encountered by foreigners to access and understand scientific content in Ph. D. thesis and M. Sc. Dissertations in digital libraries and shows a methodology supported by a computer application that could improve their understanding by Francophone students. The proposed methodology can be adapted to any language pairs. It can also be considered for adaptation to sign languages, braille and oral communication.

2. Research synthesis

The identification of barriers to access and comprehension of scientific texts by French-speaking students at the University of Brasília and the University of Lille 3 (Charles de Gaulle University) was collected through questionnaires¹, which led to obtain data about linguistic expertise and knowledge and use of automatic language processing software.

As a general outcome, two situations have been identified: 1) the simple withdrawal of the use of scientific bibliography in Portuguese by the foreign user due to insufficient knowledge of the Portuguese language; 2) lack of knowledge of natural language processing tools to facilitate access to and understanding of scientific texts available in digital libraries with texts written in Portuguese.

Based on this observation, we have begun an exhaustive search of computer tools that can help the foreign user to access and understand scientific contents available in a *corpus* of theses and dissertations of the University of Brasília, part of the Digital Library of Theses and Dissertations (BDTD), project coordinated by the IBICT (Brazilian Institute of Information Science and Technology).

3. Description of the methodology

The key idea of the proposal is to build a semantic representation of scientific texts in order to avoid the need for a complete translation of original theses and dissertations. To do this, four computing technologies were used: filtering, automatic summarization, machine translation and sentence alignment, as further explained.

Since the documents available in digital libraries are generally in **pdf** format and contain chapters or sections without major semantic interest, the first computing resource used is a **filter**. This feature produces a new document in **txt** format, debugging sections without semantic interest (acknowledgements, presentation, index, bibliography for example). This new document contains only the chapters of the thesis or dissertation to be studied. The conversion of the format to **.txt** is due to the need to use programs in which input files are required in this format.

¹ The pre-test was conducted with French-speaking students from the Portuguese Teaching and Research Center (NEPPE), UnB. A

second data collection was carried out with students in Information Sciences at Charles de Gaulle University (Lille 3)

The second component of the computer tool - the **automatic summarizer** - makes it possible to produce a smaller text formed by the most relevant sentences of the original document. There are several techniques and criteria for creating relevant abstracts. To provide greater flexibility to the user, the input parameters are the start and end page numbers of the text to be summarized. However, it is recommended to choose as parameters the first page of the first chapter and the last page of the last chapter of a thesis. The user must also provide the desired **compression rate**, which will indicate the size of the summary to be created and translated. Two automatic Portuguese automatic summarizers were tested: GISTSUMM and GENSIM. Note, however, that in one of the experiments performed, accuracy and coverage are calculated using the “ROUGE” (Recall-Oriented Understudy for Gisting Evaluation)² program to get an idea of the quality of the summary produced by GISTSUMM.

The third component of the application, **machine translation**, provides the translated text in the target language. In our case, we work with the language pair (FR, PT), but the chosen software has the ability to translate into six languages: Portuguese, French, Spanish, German, English and Japanese.

The fourth component – **paragraph alignment** - produces a parallel text (bi-text) in Portuguese and in the target language. In the targeted research, experiments were carried out with the PT - FR pair.

The flow diagram available at

https://github.com/leandro2r/automatic_summarizer

illustrates the relationships between each component of the IT tool used, accessible at

<http://multilingua.cdtc.unb.br:8080/>

4. Extension of the methodology to other linguistic pairs

Current research was conducted with Francophone students. However, the four components of the proposed methodology can be adapted to other language pairs. Naturally, it will be necessary to use filters, automatic summarizers, translation programs and sentence alignment programs able to work with the chosen language pair.

The methodology adaptation to assistive technologies such as sign languages and text-to-voice technologies is also a development to be further explored.

The extension of the proposed methodology to other language pairs is therefore an evolutionary work that can be developed in other similar projects.

5. Conclusion

Whether in libraries or directly by its users via the Web, the use of natural language processing tools to enable access and understanding of content in another language is still very embryonic. It is hoped that the methodological proposal for this research will be adopted and improved by other research groups interested in the subject, with the aim of broadening multilingualism in cyberspace and promoting the linguistic vitality of a greater number of languages. in the digital world.

6. Bibliographical References

- BOJAR, Ondrej et al. Findings of the 2016 conference on machine translation (wmt16). **Proceedings of the First Conference on Machine Translation**, v. 2: Shared Task Papers, p. 131–198, Berlin, Germany, aug. 11-12, 2016.
- BRANCO, António et al. **The Portuguese Language in the Digital Age**. Berlin: Springer, 2012.
- FRIAS-MARTINEZ, E. et al. Automated user modeling for personalized digital libraries. **International Journal of Information Management**, v. 26, n. 3, p. 234-248, 2006.
- GALE, William A.; CHURCH, Kenneth W. A program for aligning sentences in bilingual corpora. **Computational linguistics**, v. 19, n. 1, p. 75-102, 1993.
- GUPTA, Vishal; LEHAL, Gurpreet Singh. A survey of text summarization extractive techniques. **Journal of emerging technologies in web intelligence**, v. 2, n. 3, p. 258-268, 2010.
- LLORET, Elena et al. Compendium: a text summarisation tool for generating summaries of multiple purposes, domains, and genres. **Natural Language Engineering**, v. 19, n. 2, p. 147-186, 2013.
- MARCONDES, Carlos H. et al (Org.). **Bibliotecas digitais: saberes e práticas**. 2. ed. Salvador: Ufba, 2006.
- MÁRDERO, Arellano. Àngel. Serviços de referência virtual. **Ciência da Informação**, Brasília, v. 30, p.1-15, 2001.
- MENEZES, Cláudio; BAPTISTA, Dulce Maria. Metodologia de Acesso a Dissertações de Mestrado de Tradução por Estrangeiros: Uma abordagem preliminar. **Revista Iberoamericana de Ciência da Informação**, Brasília, v.10, n.1, p. 154-163, jan./jul. 2017. Disponível em <http://periodicos.unb.br/index.php/RICI/article/view/16462/18074>. Acesso em 16.10.2017
- MENEZES, Francisco Cláudio Sampaio de. O Multilinguismo e as Novas Tecnologias das Línguas no Século XXI. **Belas Infieis**, Brasília, v. 4, n. 12015, p.85-98, 01 jun. 2015. Disponível em: <<http://periodicos.unb.br/index.php/belasinfieis/issue/view/1175/showToc>>. Acesso em: 15 nov. 2015.
- MIHALCEA, R.; TARAU, P. TextRank: Bringing order into texts. Association for Computational Linguistics. **EECS News**, jul. 2004. Disponível em: <https://web.eecs.umich.edu/~mihalcea/papers/mihalcea.emnlp04.pdf>. Acesso em: 19 jun. 2017.
- RINO, Lúcia Helena Machado et al. Summarizers of Texts in Brazilian Portuguese: Lecture Notes on Artificial Intelligence. In: 17TH BRAZILIAN SYMPOSIUM ON ARTIFICIAL INTELLIGENCE, 1., 2004, São Luis. **Proceedings of the 17th Brazilian Symposium on Artificial Intelligence**. São Luís: Springer-verlag, 2004. v. 1, p.

235 - 244. Disponível em:

<https://www.researchgate.net/publication/220974768_A_Comparison_of_Automatic_Summarizers_of_Texts_in_Brazilian_Portuguese>. Acesso em: 29 set. 2004.

SOUZA, C.F.R.; NUNES, M.G.V. **Avaliação de Algoritmos de Sumarização Extrativa de Textos em Português.**

Relatórios Técnicos do ICMC-USP. NILC-TR-01-09, Novembro 2001

UNESCO, "A Decade in Promoting Multilingualism in Cyberspace", Disponível em

<http://unesdoc.unesco.org/images/0023/002327/232743e.pdf>, Acesso em: 05 out. 2017