A bilingual semantic engine to help search specialized fields of knowledge and translation equivalents from comparable scientific texts: design and implementation

Diogo Monteiro, Sílvia Araújo [1,2]

[1,2] *University of Minho, Rua da Universidade, Braga, 4710-057, Portugal*

Abstract

The technological revolution of the last decades has contributed to the consolidation of a new social paradigm known as knowledge society or information society [1]. This paradigm is reflected in a globalized and multilingual world, full of economic, commercial, political, social and cultural relations, where professional specialization is a necessity. In this context, specialized languages, with their specific vocabulary, structures and grammatical functions, have an important role to play. As a result of this new paradigm, we have a wealth of text in several areas of specialization, such as the academic field due to the promotion of Open Science [2], which has received heavy national and European investment for the setting up of scientific repositories [3]. Despite the abundance of such material, however, the format and structure of these data have limitations in terms of textual processing. We therefore believe that it is important to repurpose the wealth of open access texts on the web, enhancing their usability and exploring their potential to a greater extent. It is in this context that the PortLinguE project emerges, which aims to create a portal for specialized languages. It is our objective to reuse the material available in these academic repositories to create a bilingual search engine capable of identifying translation equivalents from comparable texts. The design aims at a Google style use, that is, very intuitive for the user. The user will only have to enter the desired term and can have access to the translation equivalents of the searched term/expression. For the moment, the search is limited to Portuguese and English and to medical texts, but the integration of other languages and scientific areas is planned. In combination with the reutilization of open access texts, this is a pioneer idea, which will pose a significant and stimulating challenge from a technical point of view. As a brief overview, our search engine will work with two major and innovative machine learning technologies, both of which will focus on a different problem.

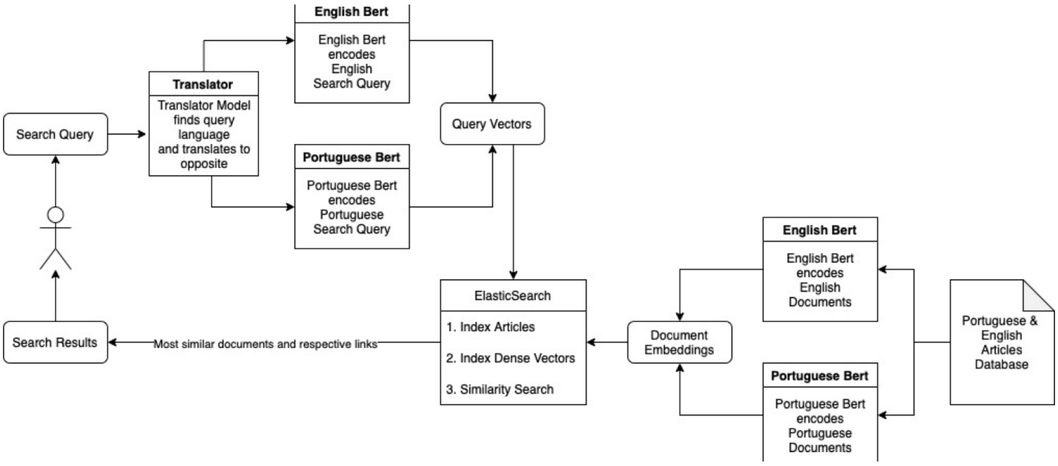
English Bert
English Bert encodes English Search Query
Translator
Translator Model finds query language and translates to opposite
Search Query
Query Vectors
Portuguese Bert
Portuguese Bert encodes Portuguese Search Query
English Bert
English Bert encodes English Documents
ElasticSearch
1. Index Articles
2. Index Dense Vectors
3. Similarity Search
Document Embeddings
Portuguese & English Articles Database
Search Results
Most similar documents and respective links
Portuguese Bert
Portuguese Bert encodes Portuguese Documents

Figure 1: Summary of data flow

Firstly, two BERT models [4], one for each language, will allow us to transform user queries and article texts into sentence vectors. The articles and respective vectors will be stored in a database and the vectors from the user queries will be used to check the semantic similarity [5] between the request and the data available in our database. At the end of this pipeline, we will be able to supply users with articles whose content is more relevant to their needs and, because the search is based on semantic context, we will be able to find adequate results even if the user has little to no knowledge of the specific vocabulary used in the area. A translation model is also used given the bilingual nature of the user request. The previous figure depicts a summary of every interaction where the Bert models are used. Secondly, we will implement a recommendation engine that will make use of the data we are able to extract from the articles retrieved in the previous BERT interaction. This engine will involve topic modeling with pre-trained machine learning models [6] and the identification of comparability metrics such as the number of views, citations or references. In turn, these metrics are used to help retrain the engine, in order to optimize the information retrieval process, i.e., to make the contents that the engine returns to the users as relevant as possible to their search. As for the project itself, the innovative aspect in terms of technology is the search engine, which manages to take comparable texts and make them parallel. It aims to solve three basic limitations: 1) in many scientific areas, it is difficult to find online texts aligned with their translations; 2) when such material exists, the translations are often of poor quality. It is therefore not only important but also useful to have a search engine which is capable of performing queries on a bilingual lexicon in an originally non-parallel corpus of academic texts; and 3) the amount of specific lexicon is too vast, which is why we chose to focus on a "semantic" search engine able to find results by context and not only specific keywords. This search engine is different from other platforms such as Google or Linguee in that it allows users to perform queries (mono and / or bilingual) in a more efficient and reliable environment in terms of documentary sources. Once processed, the massive pool of open text data provides an excellent opportunity to extract specialized information in different languages [7]. This project, which brings together a multidisciplinary team (specialists in terminology, corpus linguistics, natural language processing and artificial intelligence), aims to offer linguists a large-scale database of specialized language in use, teachers and their students an interesting didactic tool for teaching and learning languages for specific purposes [8], and translators a useful cross-language search tool that makes it easy to find translation equivalents [9] [10]. In view of the above, we firmly believe that this project would make an important contribution not only for academia but also for companies. In fact, we are in contact with several companies, in connection with the Multilingual Translation and Communication Master's degree, which have already shown interest in having such a project, as it would help to streamline their language services (e.g. translation, technical writing, glossary building). Finally, we would like to stress that PortLinguE could also be a dissemination platform for emerging scientific areas, bringing the general public, and in particular schoolchildren, closer to science.

Keywords

comparable academic texts, translation, bilingual semantic search engine, BERT, semantic similarity

Acknowledgements

The PortLinguE project (ref. PTDC/LLT-LIG/31113/2017) is funded by ERDF under Portugal 2020 and by national funds through the Portuguese Foundation for Science and Technology.

References

1. B. Wessels, R.L. Finn, K. WadhwMots.machinesa, T. Sveinsdottir, Open data and the knowledge society. Amsterdam University Press, 2017.

Words/Machines #4: Text Simplification and Adaptation, March 25, 2022, Brest Cedex, France
EMAIL: jdiogoxmonteiro@gmail.com (A. 1); saraujo@elach.uminho.pt (A. 2)
ORCID: 0000-0003-4321-4511 (A. 2)

2. OECD, Making open science a reality. OECD Science, Technology and Industry Policy Papers, No. 25, OECD Publishing, 2015. doi: 10.1787/5jrs2f963zs1-en
3. N. Rettberg, B. Schmidt, OpenAIRE; Supporting a European open access mandate. College & Research Libraries News, 76(6), pp. 306–310, 2015. doi:10.5860/crln.76.6.9326
4. J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019. URL: ArXiv, abs/1810.04805.
5. M. Walia, Measuring Text Similarity Using BERT, in: Analytics Vidhya, 2021. URL: https://www.analyticsvidhya.com/blog/2021/05/measuring-text-similarity-using-bert/
6. N. Seth, Part 2: Topic Modeling and Latent Dirichlet Allocation (LDA) using Gensim and Sklearn, in: Analytics Vidhya, 2021. URL: https://www.analyticsvidhya.com/blog/2021/06/part-2-topic-modeling-and-latent-dirichlet-allocation-lda-using-gensim-and-sklearn/
7. M. Pecman, Étude lexicographique et discursive des collocations en vue de leur intégration dans une base de données terminologiques. The Journal of Specialised Translation (JoSTrans). Issue 18, Special issue on Terminology, Phraseology and Translation, pp. 133-138, 2012. URL: https://jostrans.org/issue18/art_pecman.pdf
8. Z. Yin, E. Vine (Eds.), Multifunctionality in English: Corpora, language and academic literacy pedagogy. London, UK: Routledge, 2022. doi:10.4324/9781003155072
9. S. Bernardini, Discovery learning in the language-for-translation classroom: corpora as learning aids. Cadernos de Tradução, 36, pp. 14-35, 2016. doi:10.5007/2175-7968.2016v36-nesp1p14.
10. C. Frérot, Corpora and corpus technology for translation purposes in professional and academic environments. Major achievements and new perspectives. Cadernos de Tradução, 36, pp. 36-61, 2016. doi:10.5007/2175-7968.2016v36nesp1p36.