Revista Científica da Universidade Eduardo Mondlane, Série: Letras e Ciências Sociais ISSN 2307-3918

Pecha Kucha

MOTOR MULTIFUNÇÕES: pesquisa terminológica bilíngue e assistente de escrita académica com base em dados científicos abertos

Micaela Aguiar, José Monteiro e Sílvia Araújo

Universidade do Minho, Portugal

RESUMO: Neste artigo, exploraremos o processo de construção de um motor multifunções que está a ser desenvolvido no âmbito do projeto de investigação PortLinguE (ref. PTDC/LLT-LIG/31113/2017) e que parte da reutilização de dados científicos disponíveis em regime de acesso aberto. Daremos conta da arquitetura geral do motor que assenta numa framework Django e do modelo lógico do motor que funcionará em modelos de BERT, pois permitem efetuar pesquisas que consideram o contexto e as semelhanças semânticas. O motor tem duas funções principais que apresentamos em detalhe: (1) a função de pesquisa bilíngue de terminologia, capaz de identificar equivalentes de tradução de textos comparáveis retirados de repositórios científicos (útil a tradutores e investigadores que trabalhem com línguas de especialidade) e (2) a função de assistente de escrita académica, que parte da constituição de um banco de frases para o português académico europeu, através da recolha, anotação e análise de artigos científicos retirados de repositórios nacionais (útil a estudantes que procurem melhorar a sua escrita em contextos académicos).

Palavras-chave: Dados científicos abertos, literacia académica, motor de pesquisa, repositórios

MULTIFUNCTION ENGINE: bilingual terminology search and academic writing assistant stemming from open scientific data

ABSTRACT: In this paper, we will give an overview of the creation process of a multi-function engine that is being developed within the PortLinguE research project (ref. PTDC/LLT-LIG/31113/2017) and reuses scientific data available in open access regime. We will describe the general architecture of the engine, which is based on a Django framework, and the logical model of the engine that will work with BERT machine learning models, as it enables searches that consider context and semantic similarities. The engine has two main functions that are presented in detail: (1) the bilingual terminology search function, capable of identifying translation equivalents of comparable texts taken from scientific repositories (useful to translators and researchers working with specialized languages) and (2) the academic writing assistant function, which relies on the constitution of a phrase bank for European academic Portuguese, through the collection, annotation and analysis of scientific articles taken from national repositories (useful to students seeking to improve their writing in academic contexts).

Keywords: Open scientific data, academic literacy, search engine repositories,

Correspondência para: (correspondence to:) maguiar60@gmail.com

INTRODUÇÃO

Atualmente, o conhecimento cresce mais rápido do que qualquer bem produzido pelo ser humano, contudo é estimado que metade dos artigos científicos produzidos globalmente nunca serão lidos e que 90% nunca chegam a ser citados (CRIBB e SARI, 2010). No âmbito do projeto financiado por fundos europeus, PortLinguE, que visa a criação de um portal de recursos de apoio à literacia e ao trabalho científico, académico e técnico, estamos a desenvolver um motor multifunções que faça o aproveitamento da quantidade inumerável de conteúdo científico em regime de acesso aberto disponível em repositórios científicos. O objetivo é criar um motor com duas funções principais: a pesquisa bilíngue de terminologia, capaz de identificar equivalentes de tradução de textos comparáveis e um assistente de escrita académica, que parte da constituição de um banco de frases para o português

académico europeu, através da recolha, anotação e análise de artigos científicos.

Construção do Motor Multifunções

A arquitetura e funcionamento geral do motor são apresentados de seguida. O sistema assenta na *framework* Django (https://www.djangoproject.com), uma *framework python* para desenvolvimento web que é altamente escalável, fácil de utilizar e que possui uma vasta quantidade de pacotes, facilitando assim a qualidade do desenvolvimento. Com esta framework conseguimos estabelecer o tipo de pedidos que o utilizador poderá realizar (*search query*) mas também os resultados que são devolvidos ao utilizador (*search results*), no nosso sistema (Django API).

A lógica do motor funcionará com modelos *machine learning* BERT (DELVIN *et al*., 2019). BERT é um modelo de Processamento de Linguagem Natural que analisa *corpus* de texto em termos das semelhanças ao nível do significado das palavras, das colocações e das frases e distribui os dados processados com base na similaridade semântica, gerando assim vetores semânticos. Para a função de pesquisa terminológica bilíngue, são usados dois modelos, um pré-treinado com um *corpus* português e outro com um *corpus* inglês, que permitirá transformar as pesquisas dos utilizadores e os artigos extraídos dos repositórios científicos em vetores semânticos. Para a função do assistente de escrita, apenas o modelo português será necessário para transformar em vetores semânticos as pesquisas do utilizador e as frases-modelo, que serão disponibilizadas no nosso assistente. Os artigos e as frases-modelo, bem como os respectivos vetores semânticos, serão armazenados numa base de dados *ElasticSearch*.

O *design* visa uma utilização intuitiva para o utilizador, à semelhança de motores de pesquisa, como o Google. O utilizador realiza a pesquisa e, mesmo que tenha um conhecimento limitado sobre o que está a pesquisar, os modelos BERT serão capazes de apresentar resultados relevantes, uma vez que estes modelos são capazes de inferir contexto semântico (Pogiatzis, 2019), através de mecanismos de similaridade semântica (VARUN, 2020).

A função de pesquisa bilíngue procura dar resposta a três obstáculos que estudantes e profissionais, como tradutores e intérpretes, enfrentam quando trabalham em ambientes de linguagem de especialidade: em muitas áreas científicas, é difícil encontrar textos *online* alinhados com as suas traduções; quando tal material existe, as traduções são muitas vezes de má qualidade e a quantidade de léxico específico é demasiado extensa. Nesse sentido, o motor de pesquisa será capaz de realizar consultas em léxico bilíngue num *corpus* de artigos científicos originalmente não paralelos, retirados de repositórios científicos. Para identificar textos comparáveis nos *corpus* de artigos em português e em inglês, utilizaremos modelos de *sentence transformer*, que são treinados com pares de frases, facilitando assim a identificação de frases semelhantes nas duas línguas e a extração de equivalentes de tradução. Como o nosso motor está restrito a repositórios científicos, ao contrário de outras plataformas como o Google ou o Linguee, garantimos que os resultados provêm sempre de fontes fidedignas.

Para dar conta do vasto léxico terminológico que existe, optámos por construir um motor de pesquisa "semântico", uma vez que é capaz de encontrar resultados por contexto e não apenas por palavras-chave específicas. A pesquisa semântica será proveniente da utilização de modelos BERT pré-treinados.

O assistente de escrita assenta na mesma tecnologia e é sobretudo direcionado a estudantes do ensino superior, que procuram ajuda no processo de escrita em contextos académicos. A falta de literacia académica está bem documentada: os professores queixam-se da falta de qualidade da escrita dos alunos (ESTRELA

e SOUSA, 2011) e, para os alunos que não conseguem dominar este tipo de discurso, a falta de literacia é um fator de desigualdade (PRETO-BAY, 2004).

O objetivo do assistente de escrita é, assim, auxiliar os estudantes na escrita académica. O assistente vai partir da constituição de um banco de frases do discurso académico em Português Europeu. Este banco de frases vem na linha do Academic PhraseBank da Universidade de Manchester desenvolvido para o inglês académico. Morley (2004) analisou dissertações universitárias e determinou um conjunto de expressões, usadas para diversas funções, como introduzir textos, referir fontes, escrever conclusões, entre outros. Na construção do banco de frases para o português, partiremos destas funções para a anotação de 40 artigos científicos disponíveis em acesso aberto retirados de repositórios nacionais. O objetivo é obter como resultado frases-modelo, que poderão ser reutilizadas pelos alunos, sem o risco de plágio.

O motor será capaz de, a partir do banco de frases-template que será armazenado na base de dados ElasticSearch, analisar semanticamente o *input* do utilizador (termos, expressões ou frases) e sugerir as frases-modelo mais semelhantes ou com funções similares. Tal permitirá ao utilizador enriquecer a qualidade da sua escrita, em contextos académicos.

CONCLUSÃO

A construção deste motor multifunções visa oferecer um recurso gratuito de apoio à literacia académica e científica, que parte do aproveitamento de dados abertos, e que poderá ser útil a um leque variado de utilizadores, desde tradutores, intérpretes, investigadores, professores e estudantes. Este motor será disponibilizado, no âmbito do projeto PortLinguE, na plataforma *online* Lang2Science, que integrará ainda outros recursos de apoio às línguas de especialidade e à literacia académica.

REFERÊNCIAS

CRIBB, J.; SARI, T. **Open Science: Sharing Knowledge in the Global Century**. Collingwood: Victoria, 2010. DOI: 10.1071/9780643097643

DEVLIN, J. *et al.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. **ArXiv:1810.04805 [Cs]**, 2019. https://doi.org/10.48550/arXiv.1810.04805

ESTRELA, A.; SOUSA, O. C. Competência textual à entrada no Ensino Superior. **Revista de Estudos da Linguagem**, v,19 (1), pp. 247-267, 2011.

MORLEY, J. **Academic Phrasebank.** 2004. Disponível em: https://www.phrasebank.manchester.ac.uk/about-academic-phrasebank/. Acesso: 4 abril de 2023

POGIATZIS, A.: **NLP**: Contextualized word embeddings from BERT. 2019. Disponível em: https://towardsdatascience.com/nlp-extract-contextualized-word-embeddings-from-bert-keras-tf-67ef29f60a7b. Acesso: 4 abril de 2023

PRETO-BAY, A. M. The Social-Cultural Dimension of Academic Literacy Development and the Explicit Teaching of Genres as Community Heuristics. **The Reading Matrix**, v. 4, n. 3, 2004. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.621.8717&rep=rep1&type=pdf.

VARUN. **Calculating Document Similarities using BERT, word2vec, and other models.** 2020. Disponível em: https://towardsdatascience.com/calculating-document-similarities-using-bert-and-other-models-b2c1a29c9630. Acesso em: 4 abril de 2023