Evento	Salão UFRGS 2019: SIC - XXXI SALÃO DE INICIAÇÃO CIENTÍFICA DA UFRGS
Ano	2019
Local	Campus do Vale - UFRGS
Título	Construção de uma ferramenta de extração de informações terminológicas em inglês: a combinação homem/máquina
Autor	MARIANA ALMEIDA COLLIN
Orientador	ANA ELIZA PEREIRA BOCORNY

Universidade Federal do Rio Grande do Sul

Construção de uma ferramenta de extração de informações terminológicas em inglês: a combinação homem/máquina

Autora: Mariana Almeida Collin; Orientadora: Profa. Dr. Ana Eliza Pereira Bocorny

O volume de dados produzidos e disponíveis na internet crescerá exponencialmente nos próximos anos (Allahyari et al, 2017). Isso significa que um grande número de dados não estruturados, em formato de texto, por exemplo, poderão ser pesquisados na WEB e utilizados como corpus para busca de informações. No presente estudo temos como foco a extração de informações terminológicas (contextos definitórios, contextos de uso, equivalentes) que auxiliem alunos de graduação e jovens pesquisadores no entendimento da terminologia de diferentes áreas de especialidade. A evolução do conhecimento conduzida por novas pesquisas gera um grande número de termos e unidades terminológicas que servem para nomear novos processos, substâncias, partes de máquinas, etc. Tal terminologia dificilmente está dicionarizada, tornando difícil o seu entendimento, especialmente por alunos de graduação e jovens pesquisadores, o que justifica nossa pesquisa. Este trabalho é a segunda etapa de um projeto de elaboração de uma ferramenta de extração de informações terminológicas em inglês de textos autênticos do âmbito acadêmico como artigos, teses, dissertações e livros. Utilizando-se de uma combinação de conhecimentos da Linguística de Corpus (Biber et al, 2015) e do Processamento de Linguagem Natural (PLN) (Jurafsky e Martin, 2014) de maneira interdisciplinar com a área de Ciência da Computação, temos como objetivo final a construção de um algoritmo computacional de extração de informação que utilizará o glossário do portal LUMINA Idiomas como banco de dados de treino, empregando técnicas de aprendizado de máquinas e a posterior comparação dos dados obtidos a partir do algoritmo com aqueles produzidos por humanos. Portanto, a metodologia da etapa atual da pesquisa consiste: (i) na revisão e validação de termos incluídos pelos alunos no Glossário do LUMINA Idiomas da UFRGS, (ii) na construção do algoritmo de busca proposto e (iii) na comparação dos dados extraídos pelos humanos e pela máquina. Resultados preliminares indicam que as informações extraídas pelos humanos são mais suscetíveis a variação de qualidade e são mais restritas do que aquelas obtidas pela máquina via o algoritmo proposto. Por outro lado o humano apresenta um “filtro” que diz respeito ao seu conhecimento especializado da área que permite que ele identifique informações mais relevantes. Resultados preliminares indicam que os verbetes de termos e de unidades terminológicas resultantes da interação humano/máquina apresentam melhor qualidade do que aqueles construídos somente pelos humanos ou somente pela máquina.