Universidade: presente!

XXXI SIC

21. 25. OUTUBRO • CAMPUS DO VALE

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

Autora: Mariana Almeida Collin | Bolsista de Iniciação Científica PROPESQ-UFRGS
Orientadora: Profa. Dra. Ana Eliza Pereira Bocorny | Universidade Federal do Rio Grande do Sul

Introdução

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.

Justificativa

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.

Objetivo

O objetivo do presente estudo é a coleta de dados que auxiliem na construção de um algoritmo computacional de extração de informação terminológica que utilizará o glossário do portal LÚMINA Idiomas como banco de dados de treino.

Metodologia

1. Revisão da literatura sobre construção de algoritmos e sobre ferramentas de extração de informação da Web;
2. Construção de instruções: como extrair informações terminológicas da Web;
3. Coleta de dados realizada por humanos;
4. Análise dos dados coletados por humanos com aprendizagem guiada/supervisionada (instrução aos humanos), identificação de problemas frequentes;
5. Estabelecimento de parâmetros de busca de informação;

Referências selecionadas

Batista, A. H. (2011). Extração automática de definições: um estudo de caso em textos legislativos.
Biber, D., Egbert, J., & Davies, M. (2015). Exploring the composition of the searchable web: A corpus-based taxonomy of web registers. Corpora, 10(1), 11-45
Kilgarriff, A., & Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. Computational linguistics, 29(3), 333-347.
Oliveira, C., & de Freitas, M. C. (2006). Classes de palavras e etiquetagem na Lingüística Computacional. Calidoscópio, 4(3), 179-188.

Resultados

A partir da revisão dos 112 verbetes contendo informações extraídas pelos humanos após o aprendizado supervisionado foi possível dividir as ocorrências de inadequações em 4 categorias (gráfico 1). Após análise de cada categoria observou-se que:

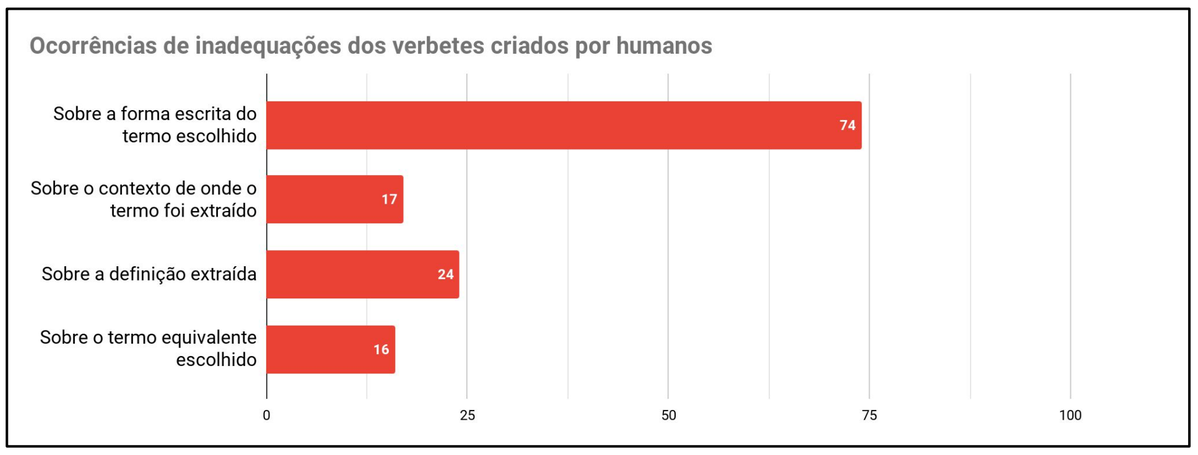

Gráfico 1: visão geral dos resultados

Análise e Conclusões

1) As inadequações sobre a forma escrita do termo ocorreram em maior número, mas as ocorrências sobre letras maiúsculas e minúsculas não seriam um problema para a máquina, já que as ferramentas de busca não são sensíveis a maiúsculas e minúsculas no sintagma terminológico; No entanto, em relação aos acrônimos e siglas é importante que se adicione o acrônimo/sigla em seguida do termo para que se tenha melhores resultados da busca pela definição do termo.

2) Foi possível perceber que algumas das ocorrências que apresentaram categorias de erro feita pelos humanos, mesmo após a orientação dada, seriam fáceis de serem evitadas a partir de uma instrução feita pela máquina. Como em "UAPS" cuja definição inserida pelo humano foi "Unlicensed Assistive Personnel", que não caracteriza uma definição, sendo apenas a expansão da sigla;

3) Por outro lado, algumas categorias de erro realizadas pelos humanos provavelmente seriam também realizados pela máquina. Como os observados em sentenças com fraseologia e estrutura de definição terminológica, mas que são não-definições informativas (Navigli et al, 2010 citado por Batista, A. H., 2011). Como, por exemplo, a definição do termo "*Sample*" foi minerada como "*Sample can be defined as well as various material properties*", onde "*can be defined*" é fraseologia de definição, mas a sentença em questão apenas relaciona o termo com uma característica.