SELEÇÃO AUTOMÁTICA DE EXTRAÇÕES DE RELAÇÕES DE ASSOCIAÇÃO GERADAS DE FORMA AUTOMATIZADA A PARTIR DE TEXTO ABORDANDO UM TEMA DE NUTRIÇÃO

Mateus Trindade Diunisio (mati_diunisio@hotmail.com)

Joinvile Batista Junior (joinvile@ufgd.edu.br)

Extratores de informação suportam a automação da geração de relações, entre conceitos do texto de entrada (corpus), que são utilizadas para a construção automática de ontologias, para representar conhecimento em um dado domínio, partir de fontes textuais. Apesar da evolução contínua de extratores de informação, ainda existem falhas nas extrações geradas. O objetivo deste trabalho foi a especificação e a implementação de um processo automatizado de geração de padrões, que possa servir de base para utilização do extrator OpenIE4. A metodologia utilizada incorpora: (1) escolha da fonte de informação e obtenção das sentenças do corpus para extração automática de relações; (2) especificação, implementação e teste de um processo manual de definição de padrões de seleção das extrações, baseado nas dez primeiras sentenças do corpus; (3) especificação, implementação e teste de um processo para automatizar a geração dos padrões genéricos de seleção, a partir do conjunto de extrações selecionadas manualmente para as dez sentenças iniciais; e (4) ampliação dos conjunto de padrões genéricos a partir da geração automática de padrões para as demais sentenças do corpus. Neste trabalho foi especificada a sintaxe de padrões genéricos de seleção de extrações de relações de associação e implementada uma ferramenta para geração automática dos padrões, a partir da escolha manual das relações de interesse nas sentenças do corpus selecionado. A escolha da utilização de padrões foi baseada na estratégia de suprir as duas lacunas necessárias para utilizar as extrações do OpenIE4 para automatizar a construção de uma ontologia: descartar as extrações não relevantes e gerar automaticamente as extrações não cobertas pelo OpenIE4. Esta segunda lacuna é o maior desafio para a completa utilização do OpenIE4 e os padrões gerados serão muito úteis para atingir esta meta, dado que a representação utilizada para os padrões cobre todos os tokens (palavras e símbolos de pontuação) da sentença, servindo de base para posterior identificação das extrações não cobertas. Como resultado deste trabalho, a partir da seleção manual das extrações relevantes, foi completamente automatizado o processo de geração de padrões para todas as 82 sentenças do corpus, cujo processo foi concebido a partir do processamento manual das dez primeiras sentenças do corpus. Este trabalho é primeiro passo para se atingir um processo automatizado completo, de utilização das extrações de relações geradas pelo OpenIE4, para a construção automática de ontologias. A continuidade desta pesquisa, para atingir a meta de automatizar a utilização das extrações geradas pelo OpenIE, abrange: (a) a especificação e prototipação de heurísticas para seleção das extrações relevantes, com base na experiência adquirida na sua seleção manual, para completar o processo de automação da geração dos padrões genéricos; e (b) a utilização dos padrões para a geração automática das extrações não cobertas pelo extrator de informações OpenIE4.

Palavras-chave: Automação de ontologias, Extração automática de relações, Padrões de seleção.