EXTENSÃO DOS PADRÕES DE CLASSIFICAÇÃO COM PADRÕES DE EXTRAÇÃO DE RELAÇÕES VINCULADAS ÀS SENTENÇAS CLASSIFICADAS

Jeovano De Oliveira Coutinho (jeovanocoutinho@gmail.com)
Joinvile Batista Junior (joinvile.batista@gmail.com)

A classificação das sentenças, utilizando filtros baseados em padrões, foi desenvolvida em um trabalho anterior, para ser utilizada na construção de uma ontologia para caracterizar o papel de cada sentença relevante no artigo técnico como, por exemplo: objetivos, contribuições, principais resultados, vantagens, desvantagens. O objetivo deste trabalho é o de agregar padrões de extração de relações, aos padrões de classificação, para que seja possível construir uma ontologia para representar artigos técnicos em uma dada área do conhecimento, dado que a ontologia interliga entidades através de relações. A metodologia utilizada neste trabalho foi baseada nas seguintes etapas: (a) escolha de padrões de classificação representativos, para selecionar as sentenças do corpus (texto de entrada) a ser utilizado; (b) extrações das sentenças selecionadas, com base nos padrões de classificação; (c) definição de uma notação para representação de padrões de extração; e (d) implementação e teste de uma ferramenta para gerar a extração das relações a partir dos padrões de extração. A escolha dos padrões de classificação representativos, foi baseada no ponto focal de cada padrão (elementos determinantes da semântica de classificação do padrão), para servir de direcionador da geração de extrações na sentença a partir do qual cada padrão de classificação foi concebido. Baseado no ponto focal de cada padrão e na premissa de evitar perda da semântica da sentença original, foram definidas extrações para cada uma das sentenças selecionadas. A partir dos textos das extrações foi definida uma notação capaz de representar padrões de extração baseados em regras, com antecedente e consequentes, a ser representada em XML. A notação especificada foi concebida para ser a mais genérica possível, priorizando a utilização de categorias de frases (nominais, verbais, adverbiais, etc) em relação a utilização de léxicos (substantivo, verbo, adjetivo, etc), e estes por sua vez priorizados em relação à utilização de texto original (com algumas exceções, como pontuações). Adicionalmente foram introduzidos outros elementos genéricos na notação para agrupar categorias de frases distintas. Finalmente foi implementada e testada uma ferramenta que automatiza a geração das extrações de relações, a partir do texto de uma sentença e de sua regra de extração descrita em XML. A utilização desta ferramenta simplifica significativamente a geração de um conjunto maior de padrões de extração, dado que o acréscimo de um novo padrão de extração não implica na alteração do código da ferramenta. A partir do resultado deste trabalho, poderá ser gerado um conjunto significativo de padrões de extração, a partir de padrões de classificação, para servir como base para a especificação de um processo automático de geração de padrões de extração.