

REFINAMENTO DE FILTROS DE CLASSIFICAÇÃO DE SENTENÇAS DE ARTIGOS TÉCNICOS EM UM TEMA ESPECÍFICO

Herlon Augusto De Aguiar Lima (herlon214@gmail.com)

Joinville Batista Junior (joinville@ufgd.edu.br)

Em um trabalho anterior, foi desenvolvido um protótipo para aplicar padrões de classificação, que foram gerados manualmente a partir de sentenças de um artigo técnico em inglês. Para especificar tais padrões foi definida uma sintaxe, capaz de capturar trechos de interesse para caracterizar a semântica de sentenças, denotando: trechos de texto original; trechos opcionais; trechos genéricos (qualquer sequência de palavras e símbolos de pontuação, até encontrar o próximo elemento definido na sintaxe do padrão); listas de elementos entre vírgulas finalizadas por “or” ou “and”; e parâmetros que podem assumir um elemento de um conjunto de valores. O propósito primário destes padrões é a classificação das sentenças de um artigo técnico, do ponto de vista de seu papel semântico no contexto do artigo. No entanto, os padrões também podem ser utilizados para delimitar relações de associação, de interesse para construir uma ontologia para apoiar prospecção tecnológica a partir de textos técnicos. Como o objetivo de refinar padrões para classificar sentenças, e delimitar relações de associação vinculadas à classificação, foi utilizada uma metodologia inicial de refinamento dos padrões: (a) obtenção de sentenças de um artigo técnico, a partir da aplicação de ferramenta desenvolvida anteriormente para extrair automaticamente sentenças de um formato TXT, gerado a partir de um formato PDF; (b) aplicação automática dos padrões às sentenças do artigo; (c) generalização manual dos padrões a partir da captura de trechos menores da sentença e de maior utilização de trechos opcionais; e (d) repetição da generalização dos padrões, se necessário, e enquanto for viável aumentar a generalidade dos padrões. No entanto, apesar dos esforços de generalização, o reuso resultante demonstrou ser insignificante. Adicionalmente, a utilização de generalização deteriora a capacidade do padrão de delimitar as relações de interesse para a construção da ontologia. Com os resultados obtidos, ficou evidente que a única forma de utilizar os padrões, preservando suas características de delimitar as relações de interesse, seria gerá-los automaticamente para cada sentença de interesse, de forma a não depender do reuso dos mesmos em sentenças distintas das suas sentenças de origem. Foi então realizada a especificação e a prototipação de uma ferramenta, com interface gráfica, para a criação iterativa de padrões. A conclusão é que este resultado servirá de base para a posterior especificação e prototipagem de heurísticas para automatizar a geração de padrões. A continuidade deste trabalho deverá englobar: (a) a utilização da ferramenta para auxiliar na prospecção de um conjunto de heurísticas necessário para automatizar a geração de padrões; (b) a prototipagem de uma ferramenta para automatizar as heurísticas de extração de padrões; e (c) a utilização dos padrões para guiar a extração de relações de interesse para a construção da ontologia para apoiar a prospecção tecnológica a partir de artigos técnicos.

Palavras-chave: Automação de Ontologias, Classificação de Sentenças, Processamento de Linguagem Natural.