DELIMITAÇÃO DE SUBONTOLOGIAS COM BASE NA INDEXAÇÃO LÉXICO-ONTOLÓGICA: PRIMEIRAS INVESTIGAÇÕES

Zacarias, Andressa C. I.[1, 2] (IC); Di-Felippo, Ariani[1, 2] (O); Pardo, Thiago A. S.[2] (CO)
andressa.caroline.z@bol.com.br
[1]*Departamento de Letras, Universidade Federal de São Carlos;* [2]*Núcleo Interinstitucional de Linguística Computacional*

Para a construção de alguns sistemas computacionais que processam língua natural, é preciso identificar o conceito subjacente às unidades lexicais que ocorrem em um *corpus*. Por exemplo, em certos sistemas de Sumarização Automática (SA) (isto é, sistemas que geram um sumário (resumo) a partir de um ou mais textos-fonte), essa identificação é feita por meio da indexação das unidades lexicais que ocorrem nos textos a uma ontologia. Por ontologia, entende-se uma base de dados que fornece um inventário de conceitos, propriedades e relações entre conceitos que representam "uma interpretação da realidade". Com base nessa indexação, as sentenças dos textos-fontes são pontuadas em função dos conceitos que englobam e as de maior pontuação são escolhidas para compor os sumários. No entanto, a construção dessas ontologias, mesmo que de domínios específicos, é tarefa reconhecidamente custosa. Assim, tem-se investigado o delineamento conceitual de *corpus* por meio da indexação de suas unidades lexicais a uma ontologia geral previamente construída e o subsequente recorte da porção da ontologia que mais adequadamente representa o conteúdo do *corpus*. Ao final dessa investigação manual, objetiva-se propor uma metodologia que possa ser automatizada e inserida em sistemas de PLN. Com isso, dado um *corpus* x, um sumarizador automático, por exemplo, poderá identificar os conceitos subjacentes às unidades lexicais de forma automática. Para investigar o delineamento mencionado, algumas tarefas foram previstas: (i) seleção do *corpus*, das unidades lexicais a serem indexadas e da ontologia geral; (ii) indexação das unidades lexicais à ontologia e (iii) delimitação da subontologia. No caso, selecionou-se o CSTNews (http://caravelas.icmc.usp.br/CSTNews/), *corpus* multidocumento em português do Brasil (PB) composto por 50 coleções de textos jornalísticos. Especificamente, selecionou-se apenas 1 (uma) das coleções, C1, cujas 38 unidades lexicais da categoria dos nomes foram manualmente indexadas à WordNet de Princeton (WN.Pr) (http://wordnetweb.princeton.edu), desenvolvida para o inglês. Essa ontologia foi escolhida devido a sua adequação linguístico-computacional e devido à inexistência em PB de uma ontologia tão robusta quanto a WN.Pr. A indexação englobou a tradução das unidades para o inglês e seleção dos *synsets* (isto é, conjunto de formas sinônimas que representam um conceito lexicalizado) da WN.Pr que codificam os conceitos subjacentes às unidades lexicais. A partir da indexação, investigam-se estratégias que possibilitam a delimitação da estrutura conceitual que engloba os conceitos mais representativos da coleção C1 do *corpus* CSTNews. Em outras palavras, busca-se delinear conceitualmente a coleção C1 por meio da delimitação de uma subontologia da WN.Pr. Alguns critérios para a delimitação de subontologias que estão sendo investigados são: (i) a frequência de ativação de um conceito x em relação à frequência de ativação dos demais conceitos ativados na ontologia; (ii) o número de instâncias de x ativadas e a frequência de ativação das mesmas; (iii) o número de conceitos superordenados de x ativados e a frequência de ativação destes, etc. Sendo assim, a região mais densamente ativada da rede de conceitos, deverá ser delineada, descartado-se o que tiver sido alinhado fora dessa região.

CNPq