Proposta de Delineamento Conceitual de *Corpus* Via Indexação Ontológica

Andressa C. Inácio Zacarias[1,2], Ariani Di Felippo[1,2], Thiago A. S. Pardo[2]

[1] Departamento de Letras (DL) – Centro de Educação e Ciências Humanas (CECH)
Universidade Federal de São Carlos (UFSCar)
Caixa Postal 676 – 13.565-905– São Carlos – SP – Brazil

[2] Núcleo Interinstitucional de Linguística Computacional (NILC)
Inst. de Ciências Matemáticas e de Computação (ICMC) – Universidade de São Paulo (USP)
Caixa Postal 668 – 13.560-970 - São Carlos, - SP – Brazil

andressa.caroline.z@bol.com.br,arianidf@gmail.com,taspardo@icmc.usp.br

Abstract. *Taking into consideration the semantic processing of documents in Brazilian Portuguese (BP) language, specially in the automatic summarization task, we present in this paper a proposal for investigating the conceptual delineation of corpus in BP using ontological indexation. Based on that, we intend to identify strategies to automate the processes of ontological indexing and conceptual delineation of textual corpora.*

Resumo. *Tendo em vista o processamento automático do português do Brasil em nível semântico, especialmente na tarefa de sumarização automática, apresenta-se, neste artigo, uma proposta de investigação do processo de delineamento conceitual de corpus em BP por meio de indexação ontológica. Com base nessa investigação, possíveis métodos ou estratégias de indexação léxico-conceitual e de delineamento conceitual de corpus podem ser estabelecidos de modo a subsidiar a automatização dos mesmos.*

1. Introdução

Para a adequada interpretação das línguas naturais no cenário do Processamento Automático das Línguas Naturais (PLN), é preciso que a máquina identifique os conceitos subjacentes às unidades lexicais e as relações que se estabelecem entre esses conceitos [Dias-da-Silva *et al.* 2007], estabelecendo uma espécie de delineamento conceitual de um texto ou *corpus* sob análise. Esse delineamento pode ser organizado formalmente, dando origem a uma *ontologia* do texto ou *corpus*. Na sumarização automática multidocumento (SAM) que envolve o português do Brasil (PB), a situação é bastante curiosa. Atualmente, há dois *sumarizadores multidocumento*, ou seja, sistemas que produzem um sumário a partir de uma coleção de textos-fonte que abordam um mesmo tópico: o GIST SUMMarizer [Pardo 2005], que é baseado em conhecimento linguístico superficial, e o CST SUMMarizer [Jorge e Pardo 2010], que produz sumários com base na identificação de relações discursivas entre as sentenças dos textos de uma coleção. Vê-se, dessa forma, que, de um lado, tem-se um sistema baseado em conhecimento superficial e, de outro, um sistema baseado em conhecimento discursivo, considerado mais abstrato e complexo que o semântico. Não há, nesse cenário, trabalhos que se baseiam no tratamento léxico-conceitual do texto ou *corpus*. Sabendo-se da relevância da aplicação dos atributos de uma ontologia na interpretação dos

textos durante a sumarização automática (SA) [Hennig *et al.* 2008], propõe-se investigar o processo de delineamento conceitual de *corpus* em PB por meio da indexação de suas unidades lexicais a uma ontologia. Dessa forma, métodos de indexação léxico-conceitual e de delineamento conceitual de *corpus* poderão ser estabelecidos de modo a subsidiar a automatização dessas tarefas. Na próxima Seção, apresenta-se uma breve revisão sobre a questão do delineamento conceitual de *corpus* na SA. Na Seção 3, apresentam-se as etapas de execução da pesquisa. Por fim, na Seção 4, algumas considerações finais sobre o projeto são feitas.

2. Breve Revisão Bibliográfica

O sumarizador automático monodocumento TOPIC [Raimer e Hahn 1988, *apud* Mani 2001] é um exemplo de sistema de PLN que realiza o delineamento conceitual dos textos-fonte. Para a sumarização de textos em alemão do domínio "computador", o TOPIC inicialmente identifica o núcleo dos sintagmas nominais (SN) do texto sob análise e, na sequência, indexa as unidades nucleares a uma ontologia do referido domínio em alemão. A seguir, o sistema aumenta o peso do conceito à medida que ele ocorre no texto e, consequentemente, é indexado/ "ativado" na ontologia. Ao final, a saliência dos conceitos do texto-fonte indexados é calculada, por exemplo, com base em: (i) frequência de ativação de um conceito *x* em relação à frequência de ativação dos demais conceitos indexados à ontologia e (ii) número de conceitos subordinados ao conceito *x* que foram ativados em relação ao número total de conceitos subordinados. Por fim, o sistema gera uma "subontologia" composta apenas pelos conceitos do texto mais indexados à ontologia, a qual pode ser utilizada para guiar o processo de SA. Wu e Liu (2003), por exemplo, selecionam o conteúdo que irá compor o sumário por meio do ranqueamento de cada parágrafo em função da saliência que seus conceitos constitutivos apresentam na ontologia. Hennig *et al*. (2008), de forma semelhante, ranqueiam as sentenças do texto-fonte em função dos conceitos da mesma na ontologia. Na próxima Seção, apresentam-se as etapas de realização da investigação proposta.

3. Metodologia

Além da revisão da literatura sobre SAM e ontologia, as seguintes etapas são propostas para a investigação do processo de delineamento conceitual de *corpus* via indexação ontológica:

(a) Seleção do *corpus*, do tipo de unidade lexical e da ontologia: o *corpus* a ser utilizado será o CSTNews [Aleixo e Pardo 2008], que é composto por 50 coleções de textos jornalísticos de domínios variados, os quais foram anotados em nível discursivo via o modelo CST (*Cross-document Structure Theory*) [Radev 2000]. Pelo menos uma coleção do CSTNews será selecionada e as unidades lexicais de seus textos constitutivos serão indexadas especificamente à WordNet de Princeton [Fellbaum 1998]. As unidades lexicais a serem indexadas poderão ser selecionadas em função da (i) categoria lexical, (p.ex.: substantivos e/ou verbos) e/ou da (ii) frequência (ou seja, seleção apenas das unidades lexicais mais frequentes da coleção) e outros.

(b) Tradução e seleção das unidades para indexação*:* tendo em vista que a ontologia em questão, a WN.Pr, foi feita para o inglês norte-americano, é preciso traduzir as unidades lexicais a serem indexadas para essa língua. Para tanto, duas estratégias serão investigadas: (i) tradução dos textos da coleção do CSTNews por meio de ferramentas automáticas e subsequente seleção das unidades ou (ii) seleção das unidades e subsequente tradução das mesmas por um dicionário eletrônico bilíngue PB-inglês.

(c) Indexação das unidades traduzidas à WN.Pr*:* após a tradução das unidades lexicais, será feita a indexação ou ligação efetiva das mesmas à ontologia. Tal indexação pode requerer uma etapa de desambiguação, pois uma unidade lexical traduzida pode ativar conceitos diferentes, ou seja, conjuntos de sinônimos diferentes (*synsets*); no caso, a desambiguação é necessária para identificar o ramo da ontologia que expressa efetivamente os conceitos presentes nos textos-fonte.

(d) Delineamento conceitual e recorte da sub-rede da WN.Pr*:* após a indexação, esta etapa consiste em delimitar/recortar a região mais densamente ativada da ontologia. O delineamento poderá com base apenas nos *synsets,* levando-se em consideração critérios como: (i) frequência das indexações de um conceito *x*, (ii) o número de hipônimos de *x* e a frequência de ativação dos mesmos; (iii) o número de hiperônimos de *x* e a frequência de ativação, etc. Outra possibilidade é aumentar o peso dos conceitos/*synsets* provenientes de sentenças com relações CST. Para tanto, assume-se a hipótese de que, se há relação CST entre as sentenças, suas unidades lexicais têm alguma sobreposição de sentido, havendo, assim, conceitos/*synsets* em comum, o que ajudaria a identificar a porção da rede que representa adequadamente os tópicos dos textos.

4. Considerações Finais

Os resultados esperados são: (i) indexação de pelo menos uma coleção do CSTNews à WN.Pr; (ii) aquisição de uma sub-rede da WN.Pr que representa o domínio da coleção indexada à WN.Pr; (iii) especificação de métodos de delineamento conceitual de *corpus* via indexação ontológica que possam subsidiar a automatização dos mesmos com vistas à aquisição futura de sub-*wordnets* de forma automática ou semiautomática. Atualmente, a revisão bibliográfica e as tarefas descritas em 3(a) estão em andamento.

5. Referências Bibliográficas

Aleixo, P. e Pardo, T.A.S. (2008) CSTNews: um *corpus* de textos jornalísticos anotados segundo a Teoria Discursiva Multidocumento CST (*Cross-document Structure Theory*). Série de Relatórios Técnicos do ICMC, São Carlos-SP, n. 326, 12p.

Dias-da-Silva, B.C. *et al.* (2007) Introdução ao Processamento das Línguas Naturais e algumas aplicações. Série de Rel. Téc. do NILC, NILC-TR-07-10. São Carlos, 121p.

Fellbaum, C. (1998) (Ed.) Wordnet: an electronic lexical database. Ca, MA: MIT Press.

Henning, L., Umbrath, W. e Wetzker, R. (2008) An Ontology-Based Approach to Text Summarization. In the Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Sydney, Australia, p. 291-294

Jorge,M.L.C. e Pardo,T.A.S.(2010)Experiments with CST-based Multidocument Summarization. In the Proc. of the ACL Workshop Textgraphs,Uppsala,Sweden,p.74-82.

Mani, I. (2001) Automatic Summarization. Amsterdam: John Benjamins Publishing Co.

Pardo, T.A.S. (2005) GistSumm - GIST SUMMarizer: extensões e novas funcionalidades. Série de Relatórios do NILC. NILC-TR-05-05. São Carlos-SP, 8p.

Radev, D. R (2000) "A common theory of information fusion from multiple text sources, step one: cross-document structure". In the Proceedings of the 1st ACL Signal Workshop on Discourse and Dialogue, Hong Kong, p. 74-83.

Wu, C.W. e Liu, C.L. (2003) Ontology-based text summarization for business news articles. In the Proceedings of the 18th International Conference CATA, Hawaii, USA, p. 389−392.