Delineamento Conceitual de Corpus Via Indexação Léxico-conceitual: Primeiros Resultados

Andressa C. I. 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

1. Introdução

Na Sumarização Automática Multidocumento (SAM), produzem-se sumários a partir de coleções de textos que tratam de um mesmo assunto. A maioria dos métodos de SAM produz sumários a partir da seleção das sentenças dos textos-fonte que veiculam o conteúdo principal da coleção. Uma das estratégias consiste na seleção das sentenças que contêm os conceitos lexicalizados mais relevantes da coleção [Li *et al.,* 2010]. Para tanto, as unidades lexicais dos textos-fonte são mapeadas aos conceitos de uma ontologia de domínio previamente construída de forma manual para a pesquisa e, diante da identificação dos conceitos mais representativos, as sentenças que os contêm são selecionadas para compor os sumários. Por ontologia, entende-se um conjunto de tipos de objetos, eventos e propriedades, organizados em função de certos relacionamentos [Gruber, 1995].

Visto que a construção de ontologias é uma tarefa cara, pois demanda tempo e equipe especializada, investigou-se o delineamento conceitual de *corpus* multidocumento por meio da indexação de suas unidades lexicais a uma ontologia de língua geral construída previamente e subsequente delimitação da região da ontologia que engloba os conceitos mais representativos da coleção.

Para apresentar a investigação, organizou-se este artigo em 5 Seções. Na Seção 1, apresenta-se a indexação das unidades lexicais à ontologia. Na Seção 2, descreve-se a investigação de diferentes critérios para a delimitação da subontologia. Na Seção 4, apresenta-se a avaliação das várias subontologias delimitadas. Por fim, na Seção 5, algumas considerações finais são apresentadas.

2. Indexação das Unidades Lexicais à Ontologia

Neste trabalho, utilizou-se o CSTNews, *corpus* multidocumento em português composto por 50 coleções de textos jornalísticos [Cardoso *et al*., 2011]. Especificamente, selecionou-se a coleção C1, composta por 3 textos da seção "mundo" dos jornais *on-line A Folha de São Paulo, Estadão* e *Jornal do Brasil* que relatam a "queda de um avião no Congo". Dessa coleção, 38 unidades lexicais da categoria dos nomes foram selecionadas para a indexação. A categoria dos nomes foi escolhida por ser responsável pela veiculação de grande parte do conteúdo textual.

Após o cálculo da frequência de ocorrência na coleção, as 38 unidades foram manualmente indexadas à WordNet de Princenton (WN.Pr) [Fellbaum, 1998], ontologia construída para o inglês. A indexação englobou os seguintes passos:

Anais do III Workshop de Iniciação Científica em Tecnologia da Informação e da Linguagem Humana, páginas 37–39, Fortaleza, CE, Brasil, Outubro 21–23, 2013.

(i) tradução da unidade lexical do português para o inglês com base no *Michaelis: Dicionário de Inglês Online*[1], *Linguee*[2] e *Google Translator*[3];
(ii) identificação dos *synsets* (isto é, conjuntos de sinônimos que codificam um único conceito) da WN.Pr que continham a tradução;
(iii) identificação do *synset* que representa o conceito subjacente à unidade lexical em C1;
(iv) associação da frequência da unidade lexical ao *synset* selecionado em (iii);
(v) seleção dos *synsets* hiperônimos (isto é, que codificam os conceitos mais genéricos) e hipônimos (ou seja, que representam os conceitos mais específicos) relativos ao *synset* selecionado em (iii), e
(vi) propagação da frequência do *synset* selecionado aos hiperônimos e hipônimos.

As hierarquias parciais herdadas da WN.Pr em função da indexação de cada uma das 38 unidades lexicais foram unificadas por meio da ferramenta *Cmap Tools*[4], que forneceu o ferramental para a representação gráfica (árvore conceitual) da hierarquia final.

3. Delimitação da Subontologia

Diante da hierarquia final, passou-se à fase de investigação, também manual, de estratégias de poda que permitissem delimitar a região da hierarquia ou árvore conceitual que englobasse os conceitos mais representativos da coleção C1. Para tanto, optou-se pela estratégia geral de poda no sentido *top-down*, com base na qual se partiu dos conceitos (*synsets*) mais genéricos em direção aos conceitos mais específicos.

Para identificar os conceitos mais representativos e podar os de menor importância, tomou-se como ponto de partida um dos critérios adotados por Raimer e Hah [1988, *apud* Mani, 2001], a saber: a frequência relativa do conceito na ontologia.

Tendo em vista a estratégia *top-down*, iniciou-se a poda pelos níveis mais genéricos. O conceito que compõe o primeiro nível da hierarquia unificada, representado pelo *synsets* {entity} e os conceitos do segundo nível, codificados nos *synsets* co-hipônimos {physical entity} e {abstract entity} não foram submetidos à poda porque capturam os tipos conceituais básicos. Assim, a poda foi realizada a partir dos conceitos/*synsets* que compõem o terceiro nível. O procedimento de poda consistiu em: (i) identificar a média das frequências dos conceitos do nível; (ii) identificar uma porcentagem da média; (iii) podar os conceitos/*synsets* que apresentavam frequência menor que a obtida em (ii). No caso, foram testadas, na etapa (ii), 5 diferentes porcentagens sobre a média da frequência (30%, 40%, 50%, 60% e 70%), gerando-se 5 subontologias distintas.

Na Figura 1, exemplifica-se o procedimento de poda no nível 4 da hierarquia dos conceitos do tipo {abstract entity}. Nessa Figura, os conceitos/*synsets* estão seguidos por suas respectivas frequências. De acordo a metodologia, calculou-se que a média das frequências dos conceitos desse nível é 6,8, a qual foi obtida pela soma das frequências dos conceitos do nível (5+16+12+42+2=**41**) e divisão do valor obtido pelo número de conceitos do nível (41/**6**=6,8). Especificando-se, por exemplo, 30% da média, obteve-se o valor de 2,04. Dessa forma, todos os conceitos/*synsets* do nível 4 que apresentam frequência igual ou inferior a 2,04 (~2) foram podados, juntamente com seus respectivos hipônimos. No caso, os conceitos podados foram {communication} e {atribute), já que ambos possuem frequência 2. Especificando-se o valor de 70% da média, ou seja, 4,8, podaram-se {communication}, {atribute} e também {relation}. Dessa forma, vê-se que quanto maior a porcentagem da média, mais conceitos são podados.

[1] http://michaelis.uol.com.br/
[2] http://www.linguee.com/
[3] http://translate.google.com/
[4] http://ftp.ihmc.us/

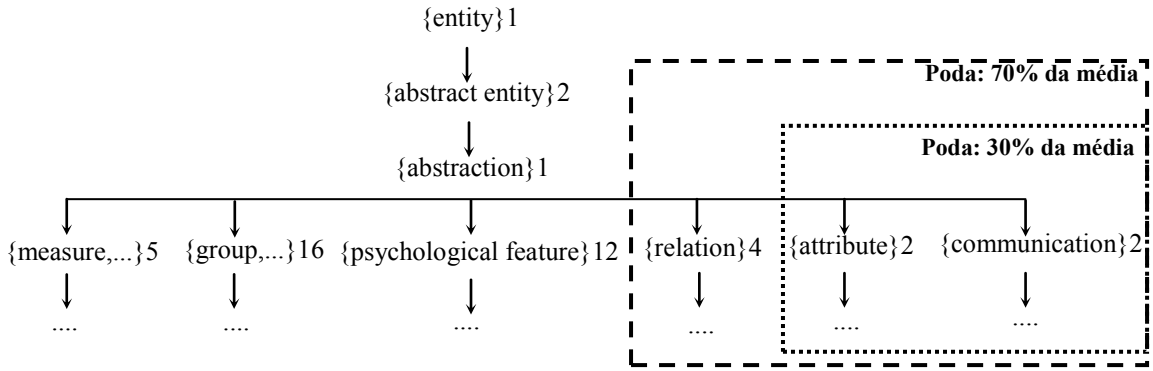

Figura 1. Exemplo de poda com base em diferentes critérios estatísticos.

4. Avaliação das Subontologias

Tendo em vista que os sumários multidocumento do CSTNews são informativos e, por isso, veiculam idealmente os conceitos mais representativos das coleções, verificou-se manualmente qual a menor das subontologias geradas pelas diferentes estatísticas que englobava o maior número de conceitos nominais presentes no sumário humano multidocumento da coleção C1. Os resultados estão sistematizados na Tabela 1.

Tabela 1. Resultados da avaliação das estratégias de poda.

	Quantidade de nomes/conceitos podados					
Nome/conceito do sumário	0%	30%	40%	50%	60%	70%
13	0	3	5	5	6	9

Na Tabela 1, evidencia-se que a estratégia baseada na especificação de 30% da média (da frequência) dos conceitos em cada nível da árvore conceitual podou menos conceitos presentes no sumário de C1; apenas 3 do total de 13. Isso se justifica porque quanto menor a porcentagem da média, menos conceitos são podados.

5. Considerações Finais

Acredita-se que o delineamento conceitual de uma coleção de textos que versam sobre um mesmo assunto é possível por meio da indexação de suas unidades lexicais a uma ontologia. Ademais, acredita-se que uma medida estatística pertinente para a delimitação da região da ontologia que engloba os conceitos mais representativos da coleção esteja em torno de 30% da média da frequência dos conceitos. Como trabalho futuro, pretende-se indexar outras coleções do CSTNews à WN.Pr para verificar se as estatísticas se confirmam pertinentes para o delineamento conceitual.

Referências

Cardoso, P.C.F. *et al* (2011). CSTNews - A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese. In the Proceedings of the 3rd RST Brazilian Meeting, p. 88-105.

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

Gruber, T. (1995). Toward principles for the design of ontologies used for knowledge sharing. *International Journal Human-Computer Studies*, v. 43, n. 5-6, p. 907-928.

Li *et al.* (2010). Ontology-enriched Multi-Document Summarization in disaster management. In the Proceedings of the ACM Special Interest Group on Information Retrieval (SIGIR), Geneva, Switzerland, p. 819-820, 2010.

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