Explorando Hierarquias Conceituais para a Seleção de Conteúdo na Sumarização Automática Multidocumento

Andressa C. I. Zacarias[1,2], Ariani Di Felippo[1,2]

[1] Programa de Pós Graduação em Linguística (PPGL)
Centro de Educação e Ciências Humanas (CECH)
Universidade Federal de São Carlos (UFSCar)
Caixa Postal 676 –13.565-905 – São Carlos – SP – Brasil

[2]Núcleo Interinstitucional de Linguística Computacional (NILC)
Universidade de São Paulo (USP) – São Carlos – SP – Brasil

{andressacizacarias,arianidf}@gmail.com

Abstract. *Based on the formal representation of a cluster of related texts in a conceptual hierarchy, we explore statistical measures to determine the most relevant concepts of the cluster. Then, the most relevant measures can be used in deep methods of Automatic Multi-document Summarization based on lexical-conceptual knowledge.*

Resumo. *Partindo-se de uma representação hierárquica dos conceitos de uma coleção de textos sobre determinado assunto, exploram-se medidas estatísticas para detectar os conceitos mais relevantes da coleção. Com isso, as medidas mais pertinentes podem ser usadas em métodos profundos de Sumarização Automática Multidocumento baseados em conhecimento léxico-conceitual.*

1. Introdução

Diante da enorme quantidade de informação textual disponível na *web* e do pouco tempo que se têm para assimilá-la, o interesse por aplicações de Sumarização Automática Multidocumento (SAM), desenvolvidos no âmbito das pesquisas sobre o Processamento Automático de Língua Natural (PLN), intensificou-se nos últimos anos. Essas aplicações buscam gerar, a partir de uma coleção de dois ou mais textos (cada um advindo de um jornal distinto) sobre um mesmo tópico, um sumário coeso e coerente [Mani 2001]. Tais sumários são comumente extratos informativos compostos por sentenças extraídas integralmente dos textos-fonte por veicularem a ideia central da coleção.

Assim, a questão central na SAM extrativa tem sido selecionar as sentenças relevantes para compor o sumário. No geral, a seleção segue 2 etapas. Primeiro, as sentenças são pontuadas e ranqueadas por um critério de relevância que busca capturar a redundância da informação na coleção, pois esse é comprovadamente o principal critério utilizado pelos humanos [Mani 2001]. Em seguida, as sentenças no topo do ranque são selecionadas para o sumário, buscando-se eliminar a redundância entre elas, até que se atinja a taxa de compressão (tamanho desejado do sumário).

Para o português, há métodos desenvolvidos segundos os 3 paradigmas de sumarização automática (SA), são eles: (i) métodos superficiais, que usam pouco conhecimento linguístico ou estatística para selecionar as sentenças; (ii) métodos profundos, que fazem uso massivo de conhecimento linguístico e (iii) métodos híbridos, que unem conhecimento linguístico e estatístico. Os métodos profundos, em especial, são mais os caros e têm aplicação mais restrita que os superficiais, pois dependem de recursos (p.ex.: gramáticas, léxicos e modelos de discurso) e ferramentas linguístico-computacionais auxiliares (p.ex.: *parser* discursivo), porém geram sumários mais coerentes, coesos e informativos.

Os métodos profundos para o português pautam-se majoritariamente em conhecimento discursivo, já que os pesquisadores dispõem da CST (*Cross-document Structure Theory*) [Radev 2000], que é uma teoria (e modelo) multidocumento robusta e computacionalmente tratável para representar os textos de uma coleção em nível discursivo. Aliás, o melhor método para o português, o RC-4 [Cardoso 2014], recebe esse nome porque seleciona as sentenças com base em informações advindas da anotação dos textos-fonte de acordo com a CST e também a RST (*Rhetorical Structure Theory*) [Mann e Thompson, 1987].

Além desses, destacam-se os métodos de SAM mutilíngue (português-inglês) de Tosta (2014) que, para gerar extratos em português, baseiam-se em conhecimento léxico-conceitual. Tais métodos partem de coleções compostas por 1 texto em português e 1 em inglês. Na sequência, os nomes que ocorrem nos 2 textos-fonte são indexados à WordNet de Princeton [Fellbaum 1998] e, em seguida, as sentenças dos textos-fonte são pontuadas e ranqueadas com base na frequência de ocorrência de seus conceitos constitutivos na coleção. A partir do ranque, um dos métodos seleciona apenas as sentenças em português com pontuação mais alta para compor o sumário, até que a taxa de compressão desejada seja atingida. Outro método seleciona as sentenças mais bem pontuadas independentemente de sua língua-fonte e, caso sentenças em inglês sejam selecionadas, faz-se a tradução destas para o português. Segundo Tosta (2014), tais métodos se mostraram muito promissores, gerando extratos com boa qualidade linguística e informatividade.

Além de terem sido testados somente no cenário multilíngue, os métodos de Tosta (2014) utilizam apenas a frequência de ocorrência de conceitos na coleção como critério para capturar a redundância e, por conseguinte, selecionar as sentenças para o sumário. Na literatura, no entanto, tem-se o método de Hennig *et al* (2008) para o inglês que, a partir da indexação das palavras de conteúdo das sentenças de uma coleção de textos-fonte a uma hierarquia de conceitos, utilizam informações estruturais da hierarquia para delimitar os conceitos mais relevantes e, por conseguintes, as sentenças que os veiculam.

Assim, diante desse cenário, apresenta-se aqui uma investigação sobre a pertinência das propriedades hierárquicas mais difundidas na literatura para a identificação dos conceitos ou tópicos mais relevantes de dada coleção de textos-fonte. Com isso, as mais relevantes poderão subsidiar à seleção de sentenças em métodos extrativos de SAM.

Na Seção 2, apresentam-se os principais métodos de sumarização da literatura baseados na representação dos textos-fonte em hierarquias conceituais e demais

trabalhos relacionados. Na Seção 3, delimita-se o conjunto de propriedades hierárquicas investigadas. Na Seção 4, apresentam-se o *corpus*, a hierarquia conceitual e o processo de indexação léxico-conceitual do *corpus*. Na Seção 5, apresenta-se o processo de descrição das propriedades da hierarquia e a avaliação da sua pertinência. E, por fim, na Seção 6, tecem-se algumas considerações finais e trabalhos futuros.

2. Trabalhos relacionados

Os métodos profundos de SA baseados em conhecimento conceitual englobam uma fase de análise dos textos-fonte em que as palavras de conteúdo são indexadas a uma hierarquia conceitual, resultando em uma representação léxico-conceitual do conteúdo da coleção de textos. Essas hierarquias são compostas basicamente por conceitos e relações de subsunção (*is-a* ou é-um) entre os conceitos e, uma vez concebidas como árvores, os conceitos são representados por folhas (ou nós) e as relações por galhos.

No cenário monodocumento, tem-se, por exemplo, o trabalho de Reimer e Hahn [1988, *apud* Mani, 2001], em que se descreve o Topic, ou seja, uma espécie de sumarizador para o alemão que identifica os trechos de um texto-fonte do domínio "computador" que veiculam seu conteúdo principal[1]. Para tanto, o Topic indexa o núcleo dos sintagmas nominais do texto-fonte a conceitos de uma hierarquia conceitual de domínio construída manualmente por especialistas. A cada indexação a um conceito *x*, este é pontuado. Ao final, a sub-hierarquia que engloba os conceitos mais pontuados representa o conteúdo principal do texto-fonte. Consequentemente, as sentenças que expressam os conceitos da sub-hierarquia são as mais relevantes do texto.

Wu e Liu (2003), por sua vez, utilizam uma hierarquia conceitual, manualmente construída, composta por 142 conceitos do domínio *Sony Corporation* (Fig.1). Com base nela, os autores identificam os principais conceitos (tópicos) de um único texto-fonte e, na sequência, os parágrafos que os veiculam para compor o sumário.

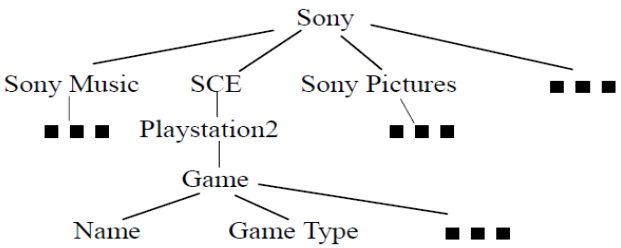

Figura 1 – Hierarquia parcial do domínio *Sony Corporation* [Wu e Liu 2003].

Para tanto, as palavras de conteúdo do texto são indexadas aos conceitos da hierarquia. A cada indexação, o conceito *x* é pontuado, juntamente com seus conceitos superordenados. Por exemplo, se *Game* da Figura 1 for pontuado, pontuam-se também os conceitos *Playstation2*, *SCE* e *Sony*. Por consequência, os conceitos superordenados acumulam a pontuação de todos os seus subordinados. Ao final, os autores consideram que os conceitos de maior pontuação do segundo nível da hierarquia (sentido *top-down*) correspondem aos tópicos do texto e, por isso, os parágrafos que os contêm são selecionados para compor o sumário. Método semelhante foi proposto por Silva (2006) para o português.

[1] O TOPIC não gera sumários, apenas indica os trechos do texto que expressam seu conteúdo central.

No cenário multidocumento, Hennig *et al.* (2008) utilizaram uma hierarquia de língua geral (em inglês) composta por 1036 conceitos. Cada conceito é representado por um rótulo simples, como *health* ("saúde"), e por um "saco de palavras" (do inglês, *bag-of-words*), ou seja, um conjunto de palavras de conteúdo semanticamente relacionadas ao conceito rotulado, como *well-being* ("bem-estar") e *life* ("vida") no caso de *health.* Com base na similaridade lexical entre uma sentença S dos textos-fonte e os "sacos de palavras" da hierarquia, o método indexa S a um ou mais ramos da árvore. Após a indexação de todas as S de uma coleção, os autores determinam as sentenças mais representativas da coleção com base em 3 métricas, 2 delas relativas à hierarquia, a saber: (i) número de ramos indexados e (ii) profundidade dos ramos indexados. A propriedade (i) busca capturar a especificidade do conteúdo de S. A propriedade (ii) busca capturar a quantidade de informação distinta que S expressa. Com base em critérios como esses, a relevância das sentenças é calculada e as de maior pontuação selecionadas para o sumário.

Além dos métodos profundos de SA mono e multidocumento baseados em conhecimento léxico-conceitual, destacam-se os conduzidos na área de pesquisa denominada Sumarização de Ontologias[2] (SO) [Zhang *et al.* 2007]. Segundo Zhang *et al*. (2007), o objetivo da SO é desenvolver métodos automáticos para produzir uma versão reduzida de uma ontologia, composta pelo subconjunto dos conceitos mais representativos da ontologia original. Tais trabalhos são relevantes porque exploram diferentes estratégias que capturam a relevância dos conceitos a partir de uma representação ontológica do conhecimento. Tais estratégias, mesmo baseadas na representação das ontologias em grafos[3], podem ser aplicadas a representações arbóreas ou árvores[4] (ou hierarquias). Segundo Sousa (2011), há uma série de medidas utilizadas para determinar a relevância de um conceito nas aplicações de SO. Dentre as mais eficazes, tem-se (i) centralidade, que é número de relacionamentos (arestas) do conceito *c* em uma ontologia O, (ii) frequência, ou seja, o número de ocorrência de *c* em ontologias que deram origem a O[5], (iii) simplicidade do nome, que define a relevância do conceito com base na complexidade ou simplicidade de sua expressão linguística, e (iv) proximidade, que define a distância entre os conceitos de maior relevância de O.

Com base nos trabalhos desenvolvidos em SA e SO, delimitou-se um conjunto de 4 propriedades. Na sequência, apresenta-se cada uma delas, enfatizando a hipótese sobre a sua relevância para a identificação dos conceitos mais centrais de uma coleção de textos-fonte, a partir de sua representação em uma árvore conceitual.

3. Propriedades Hierárquicas e Respectivas Métricas

Com base nos trabalhos de SA e SO, delimitara-se, como mencionado, 4 propriedades dos conceitos em uma representação hierárquica, as quais foram codificadas em

[2] No caso, as ontologias são objetos formais, ou seja, inventários de conceitos e relações diversas entre conceitos (não somente *is-a*) descritos de forma explícita por um formalismo, como RDF (*Resource Description Framework*) ou OWL (*Ontology Web Language*).

[3] As ontologias na SO são visualmente representadas em grafos direcionados, em que os conceitos são codificados em vértices e as relações em arestas com direção [Sousa 2011].

[4] Na teoria dos grafos, uma árvore é um grafo simples, no qual não existem ciclos. As hierarquias se caracterizam pela relação de subsunção (*is-a* ou *é um*).

[5] Essa medida é usada quando O resulta de um processo de integração de outras ontologias (O^1, O^2, $O^{n)}$.

medidas estatísticas[6], a saber: (i) frequência, (ii) centralidade, (iii) proximidade e (iv) nível. A medida frequência, em especial, foi especificada em 2: (i) frequência simples e (ii) frequência acumulada, resultando no total de 5 atributos.

a. *Frequência*: no caso da SAM, esse atributo é relevante porque captura a redundância, que é o critério usado na seleção de conteúdo/sentenças, pois os conceitos mais redundantes são considerados os mais importantes dada uma coleção de textos. Aqui, foram consideradas as frequências *simples* e *acumulada* para pontuar os conceitos na árvore. Na indexação com base na *frequência simples,* a pontuação de um conceito *x* reflete unicamente a frequência de ocorrência de *x* na coleção. Utilizando a *frequência acumulada,* a pontuação de conceitos superordenados acumulada a frequência de todos os seus conceitos subordinados. Com isso, busca-se privilegiar os conceitos mais genéricos.
b. *Centralidade*: esse atributo é definido pelo número de ligações que um conceito possui com outros conceitos da hierarquia, sendo os relacionamentos codificados em arestas. Selecionou-se esse atributo porque ele busca definir o quão um conceito está relacionado a outros, o que pode ser relevante para selecionar sentenças que veiculam informações relacionadas, contribuindo para a coerência do sumário.
c. *Proximidade*: essa propriedade identifica os conceitos relevantes que estão próximos a outros conceitos relevantes. Em outras palavras, essa medida não determina a relevância de um conceito *x* de forma isolada, mas sim em relação a relevância de outros conceitos. Essa medida pode ser importante para garantir o relacionamento entre os conceitos de maior importância de uma representação conceitual.
d. *Nível*: esse atributo determina a localização de um conceito *x* em uma representação conceitual. No caso de um modelo hierárquico, o nível expressa a generalidade ou especificidade de *x.* Assim, há conceitos genéricos, intermediários e específicos. Segundo estudos de diferentes áreas, os conceitos intermediários costumam ser os mais representativos, posto que não são tão genérico e nem específicos.

Para testar a relevância das medidas, selecionou-se uma das coleções do CSTNews *corpus* multidocumento de referência em português para a SAM [Cardoso *et al*. 2011]. A seguir, o CSTNews e a representação hierárquica de sua coleção C1 são descritos.

4. O *Corpus* e a Hierarquia Conceitual

O CSTNews está organizado em 50 coleções, distribuídas nas categorias "esporte" (10), "mundo" (14), de "dinheiro" (1), "política" (10), "ciência" (1) e "cotidiano" (14). Cada coleção é composta por: (i) 2 ou 3 notícias sobre um mesmo assunto, coletadas de diferentes jornais; (ii) 5 *abstracts* multidocumento manuais e 5 extratos multidocumento manuais; (iii) sumários automáticos multidocumento, (iv) anotações linguísticas diversas.

Para este trabalho, selecionou-se a coleção C1, composta por 3 textos da seção "mundo", os quais relatam a "queda de um avião no Congo". Nela, selecionaram-se as 38 palavras da categoria dos nomes, as quais foram manualmente indexadas aos seus respectivos conceitos da WordNet de Princenton (WN.Pr.) [Fellbaum 1998]. Apesar de ser uma ontologia em inglês, a WN.Pr. foi escolhida devido a acessibilidade, pertinência linguística e abrangência. Na WN.Pr, as palavras ou expressões do inglês estão

[6] As fórmulas matemáticas para cada uma delas podem ser encontradas em Souza (2011).

divididas nas categorias de nome, verbo, adjetivo e advérbio. As unidades de cada categoria estão codificadas em *synsets* (*synonym sets*) (ou seja, conjuntos de formas sinônimas ou quase-sinônimas como {car; auto; automobile; machine; motorcar}), sendo que cada *synset* representa um único conceito lexicalizado. Os *synsets* estão conectados entre si pela relação léxico-semântica da antonímia e pelas relações semântico-conceituais de hiperonímia/ hiponímia, holonímia/ meronímia, acarretamento e causa.

A indexação dos 38 nomes seguiu 4 passos. Para ilustrar, descreve-se a indexação de um deles, "acidente": (i) tradução de "acidente" para o inglês "accident", o que foi feito pela consulta a vários recursos (p.ex.: dicionários bilíngues e serviços de tradução *online*); (ii) identificação dos *synsets* da WN.Pr em que "accident" ocorre; no caso, {accident} ("*a mishap; especially one causing injury or death*") e {accident, fortuity, chance event}; (iii) identificação do *synset* que representa o conceito subjacente à unidade lexical em C1; no caso, escolheu-se {accident} com base em seus hiperônimos (superordenados), e (iv) seleção dos *synsets* hiperônimos relativos ao *synset* {accident}, resultando em uma hierarquia para {accident}. As hierarquias resultantes da indexação de cada um dos nomes foram unificadas com o auxílio da ferramenta gráfica *Cmap Tools* (http://ftp.ihmc.us/). A hierarquia final, que codifica o conteúdo de C1, possui 12 níveis, cujos conceitos estão organizados pela relação de hiponímia. A Figura 2 ilustra simplificadamente essa hierarquia. Nela, os conceitos/*synsets* em negrito são oriundos da coleção e os demais foram herdados da WN.Pr para construção da hierarquia.

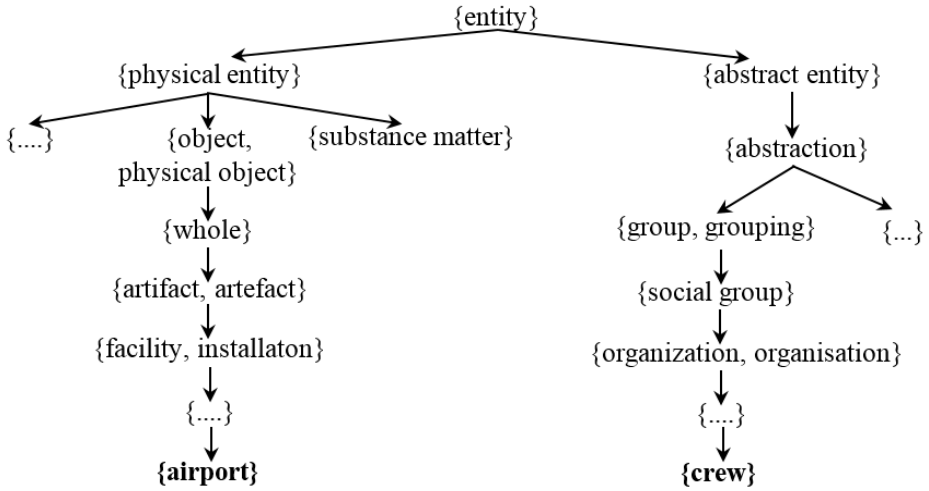

Figura 2 – Exemplo simplificado da hierarquia conceitual de C1 a partir da WN.Pr.

5. Cálculo das Métricas e Avaliação da Pertinência

As 5 medidas foram calculadas para os 38 nomes/conceitos da C1 com base em Sousa (2011). Objetivando investigar a pertinência das medidas para a identificação dos conceitos relevantes, os 13 nomes do sumário humano de C1 foram considerados os conceitos relevantes de referência, já que tal sumário é informativo e genérico e, por isso, veicula o conteúdo principal da C. Os valores das métricas para os 38 nomes/conceitos e a relevância de cada um estão descritos na Tabela 1. A análise manual da Tabela 1 visou à correlação entre as métricas e a relevância dos conceitos. Para tanto, calculou-se a média simples dos valores de cada métrica em função dos

conceitos. Na sequência, verificou-se o número de conceitos que obteve valor igual ou superior à média. Por exemplo, a média da *frequência simples* para os 38 conceitos foi 2,578947368. Dos 13 conceitos "sim", 6 apresentaram *frequência simples* igual/superior à média. A única exceção a esse cálculo com base na média diz respeito ao Nível, pois se identificou na Tabela 1 o número de conceitos intermediários (nível 5 a 8) de cada classe. Os resultados da análise manual da Tabela 1 estão na Tabela 2.

Tabela 1 – Cálculo das métricas de relevância à árvore conceitual da C1 do CSTNews.

Nome/ conceito	Ocorrência no sumário (relevância)	Métricas				
		Freq. simples	Freq. acumulada	Centralidade	Proximidade	Nível
avião	Sim	11	11	1	8,230659721	1
carga	Sim	2	2	1	7,712585958	6
floresta	Sim	3	3	1	6,956477585	6
membro	Sim	4	4	1	7,463956661	4
mineral	Sim	2	2	1	7,876829102	8
montanha	Sim	1	1	1	7,719872133	7
nacionalidade	Sim	2	2	1	6,844506727	6
passageiro	Sim	5	5	1	7,810380995	5
pessoa	Sim	3	23	5	8,811776484	7
queda	Sim	1	6	1	6,795889943	2
tempo	Sim	2	2	1	7,11956514	6
tripulação	Sim	4	5	1	6,949137169	5
vítima	Sim	2	2	1	7,575455097	5
acidente	Não	5	6	2	6,821435922	3
aeronave	Não	1	12	2	8,465048425	3
aeroporto	Não	4	4	1	7,562850767	5
aterrissagem	Não	2	2	1	6,367248415	4
chama	Não	1	1	1	6,539658371	4
cidade	Não	1	1	1	6,690983796	4
companhia	Não	4	4	1	6,633386858	3
distância	Não	2	2	1	6,579270984	5
estrada	Não	1	1	1	7,645909791	6
fabricação	Não	3	3	1	6,359046992	3
fonte	Não	2	2	1	6,636039068	5
leste	Não	2	2	1	6,422047258	4
localidade	Não	2	2	1	7,062993835	5
país	Não	2	2	1	7,029656151	5
permissão	Não	1	1	1	5,973810688	3
pista	Não	3	3	1	7,470468024	5
porta-voz	Não	7	7	1	7,955764756	5
propriedade	Não	2	2	1	7,181055684	6
quilômetro	Não	4	4	1	0,763595	5
setor	Não	1	1	1	0,747029	5
sobrevivente	Não	2	2	1	0,859697	5
tarde	Não	2	2	1	0,750951	5
tempestade	Não	1	1	1	0,802491	6
transporte	Não	1	13	2	0,979921	6
tripulante	Não	1	5	1	0,788619	5

Tabela 2 – Correlação entre as métricas e a relevância dos conceitos.

Métrica	Conceito (Qt. Absoluta e porcentagem)			
	Sim	**Não**	**Sim**	**Não**
Frequência simples	6/13	7/25	46%	28%
Frequência acumulada	6/12	7/25	46%	28%

Centralidade	1/13	3/25	7,6%	12%
Proximidade	8/13	7/25	61,5%	28%
Nível	10/13	16/25	76%	64%

Quanto à Tabela 2, observa-se que:

a. As *frequências* parecem expressar a relevância, pois se destacam em quase metade dos conceitos da categoria "sim" (46%) e em apenas 28% dos da classe "não"; assim, os conceitos do sumário parecem ser os quais mais se repetem nos textos-fonte;
b. A *centralidade* parece não distinguir os conceitos relevantes dos demais, pois pouco se destaca em ambos os conjuntos; isso pode ser explicado pelo fato de que os 38 conceitos dos textos-fonte estão conectados na grande maioria das vezes unicamente ao seu hiperônimo, possuindo, assim, apenas 1 relacionamento na hierarquia;
c. A *proximidade* parece ser um bom indicativo de relevância (61,5% dos casos "sim"); isso indica que os conceitos do sumário são semanticamente relacionados entre si;
d. O *Nível* também parece indicar relevância, já que os conceitos "sim" estão localizados em posições intermediárias da hierarquia em 76% dos casos; tal atributo, no entanto, também é relativamente significativo nos casos da classe "não".

6. Considerações finais

Apesar de preliminares, já que derivam de uma análise manual de um *corpus* pequeno, os resultados do trabalho indicam a pertinência das métricas *frequência* (simples ou derivada), *proximidade* e *nível* na tarefa de identificação de conceitos relevantes nos moldes da Fig. 1. Para refinar a análise, pretende-se submeter os dados da Tabela 1 a algoritmos de Aprendizado de Máquina, os quais podem identificar padrões estatísticos que correlacionam os conceitos às métricas.

Referências

Cardoso, P.C.F. Maziero, E.G.; Castro Jorge, M.L.R.; Seno, E.M.R.; Di-Felippo, A.; Rino, L.H.M.; Nunes, M.G.V.; Pardo, T.A.S. (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.

Cardoso, P.C.F (2014). Exploração de métodos de sumarização automática multidocumento com base em conhecimento semântico-discursivo. Tese de Doutorado. ICMC-USP.

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

Hennig, L., Umbrath, W., Wetzker, R. (2008). An ontology-based approach to Text Summarization. In the proceedings of the Workshop on natural language processing and ontology engineering (NLPOE), Toronto, p. 291-294.

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

Mann, W.C. e Thompson, S.A. (1987). Rhetorical Structure Theory: A Theory of Text Organization. Technical Report ISI/RS-87-190.

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, Canada, p. 74-83.

Silva, P.P. (2006). ExtraWeb: um sumarizador de documentos web baseados em etiquetas html e ontologia. Dissertação de Mestrado. ICMC-USP.

Sousa, P.O.V.Q. (2011). Otimização de uma Ferramenta para Sumarização de Ontologias. Trabalho de Conclusão de Curso. Centro de Informática – UFPE.

Tosta, F. E. S. (2014). Aplicação de conhecimento léxico-conceitual na sumarização multidocumento multilíngue. Dissertação de Mestrado. PPGL-UFSCar.

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.

X. Zhang, G. Cheng, and Y. Qu. (2007). Ontology Summarization Based on RDF Sentence Graph. In the 16th International World Wide Web Conference, Canada, pp. 707-715.