Enriquecendo o Córpus CSTNews – a Criação de Novos Sumários Multidocumento

[1]Márcio S. Dias, [1]Alessandro Y. Bokan Garay, [2]Carla Chuman, [3]Cláudia D. Barros, [1]Erick G. Maziero, [1]Fernando A. A. Nobrega, [2]Jackson W. C. Souza, [1]Marco A. Sobrevilla Cabezudo, [2]Marina Delege, [1]Maria Lucía R. Castro Jorge, [2]Naira L. Silva, [1]Paula C. F. Cardoso, [1]Pedro P. Balage Filho, [1]Roque E. López Condori, [2]Vanessa Marcasso, [2]Ariani Di Felippo, [1]Maria das Graças V. Nunes, [1]Thiago A. S. Pardo

Núcleo Interinstitucional de Linguística Computacional (NILC)
[1]Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
[2]Departamento de Letras, Universidade Federal de São Carlos
[3]Instituto Federal de Educação, Ciência e Tecnologia de São Paulo

Resumo. Relata-se, neste artigo, o processo de criação de novos sumários multidocumento – extrativos e abstrativos – para o córpus CSTNews, que é um córpus voltado para o processamento multidocumento, em especial, a sumarização automática para a língua portuguesa. Com isto, tem-se mais dados para subsidiar novas pesquisas na área, tanto no desenvolvimento quanto na avaliação de métodos e sistemas de sumarização.

Keywords: Sumarização Multidocumento, Córpus

1 Introdução

Considerada uma subárea de pesquisa de Processamento de Linguagem Natural (PLN) [4], a Sumarização Automática Multidocumento (SAM) é a tarefa de se produzir sumários (comumente chamados resumos) de maneira automática a partir de um conjunto de textos-fonte sobre um mesmo assunto.

Na SAM, assim como em diversas tarefas de PLN, os córpus exercem um papel muito importante, por exemplo, para se fazer uma análise linguística do material textual e modelar processos computacionais de sumarização, além de treinar e avaliar sumarizadores automáticos.

Para o português, tem-se conhecimento de apenas dois córpus disponíveis para a SAM: o córpus CSTNews[1] [1, 2], para o português brasileiro e já amplamente utilizado na área, e o recente *Priberam Compressive Summarization Corpus*[2] (PCSC) [7], para o português europeu. O CSTNews, foco deste trabalho, é composto por textos e seus sumários extrativos (i.e., formado por sentenças extraídas dos textos-fonte e justapostas) e abstrativos (com material reescrito) manualmente construídos, além de anotações linguísticas variadas.

[1] Disponível em http://www.icmc.usp.br/pessoas/taspardo/sucinto/cstnews.html
[2] Disponível em http://labs.priberam.com/Resources/PCSC.aspx e em http://www.linguateca.pt/Repositorio/PCSC/

Para algumas tarefas citadas anteriormente, ter um córpus com mais sumários humanos (considerados como referência) é muito importante. Apesar do CSTNews ter subsidiado a pesquisa em SAM para o português (veja, por exemplo, [3], [5], [6] e [8]), sua limitação no número de sumários de referência impede que avaliações mais consistentes/confiáveis sejam feitas. Por essa razão, foi desenvolvida uma tarefa de produção de sumários multidocumento extrativos e abstrativos, a qual é relatada neste artigo.

A seguir, o processo de criação de sumários e uma caracterização quantitativa dos resultados é apresentada.

2 O Córpus CSTNews

O córpus CSTNews foi criado para fins de sumarização multidocumento, no âmbito do projeto SUCINTO[3], sendo atualmente composto por 50 coleções de textos jornalísticos de domínios variados. Cada coleção possui 2 ou 3 textos sobre o mesmo assunto, provenientes de diferentes agências de notícias on-line, como Folha de São Paulo, Estadão, O Globo, Jornal do Brasil e Gazeta do Povo. Os textos foram coletados manualmente das páginas das agências por um período de 2 meses, entre Agosto e Setembro de 2007.

A escolha das agências citadas deve-se à popularidade que as mesmas possuem no meio e por trazerem as principais notícias do dia corrente, fato mais relevante para o córpus multidocumento, ou seja, uma mesma notícia publicada em fontes diferentes. Os textos jornalísticos foram escolhidos por possuírem uma linguagem clara e do dia a dia, além da facilidade de serem encontrados na web. Acredita-se, com isso, que métodos e modelos linguístico-computacionais para língua geral possam ser desenvolvidos com base nesse córpus.

Além de cada coleção de textos possuir um sumário abstrativo e um extrativo, os textos e sumários são acompanhados de várias anotações linguísticas, como relacionamento discursivo mono e multidocumento entre os segmentos textuais dos textos-fonte, alinhamento sentencial entre textos e sumários, segmentação dos textos-fonte em subtópicos e suas correlações, delimitação de expressões temporais nos textos-fonte, identificação de aspectos informativos nos sumários abstrativos, desambiguação lexical de sentidos de substantivos (mais frequentes) e verbos, e anotação morfossintática automática dos textos-fonte, dentre outras.

3 Metodologia de Criação de Sumários

Para que o córpus tivesse uma boa quantidade de sumários de referência para subsidiar pesquisas em SAM, foi conduzida a produção de mais 5 extratos e 5 *abstracts* para cada coleção de textos do CSTNews, totalizando 250 extratos e 250 *abstracts*, portanto. Para tal finalidade, foram reunidos 20 pesquisadores de PLN – alunos e docentes das áreas da Linguística e da Ciência da Computação – sendo que

[3] Mais informações em http://www.icmc.usp.br/pessoas/taspardo/sucinto/

cada pesquisador teria a incumbência de produzir 25 sumários entre extratos e *abstracts*. A atribuição das coleções e do tipo de sumário a cada um dos pesquisadores foi feita de forma balanceada, já que cada coleção do CSTNews possui diferentes tamanhos.

A tarefa de criação de sumários foi realizada diariamente, sendo que, a cada dia, os pesquisadores deveriam criar dois sumários, um extrato e um *abstract* de coleções diferentes, com o intuito de deixarem os sumários tão diversificados quanto possível na sua construção. A tarefa foi realizada em aproximadamente 1 mês, sendo que, inicialmente, foi realizada uma reunião com todos os pesquisadores para que as instruções para a realização da tarefa fossem passadas e explicadas. Como não havia necessidade de reunir todos os pesquisadores no mesmo local para a criação dos sumários, já que a mesma necessitava apenas da subjetividade de cada pesquisador e de sua capacidade de resumir, decidiu-se que os sumários poderiam ser feitos a distância, desde que a entrega fosse feita por e-mail em, no máximo, 24 horas depois do prazo estipulado para cada coleção de textos. Esse tipo de restrição é importante para manter o comprometimento dos participantes e o controle sobre os prazos da tarefa.

A tarefa teve algumas restrições que todos os pesquisadores deveriam respeitar para manter a uniformidade dos sumários. Uma delas foi a limitação de tamanho dos sumários, já que, nesta tarefa, utilizou-se uma taxa de compressão de 70% em relação ao tamanho do maior texto da coleção em análise. Por exemplo, a coleção 23 do CSTNews possui 2 textos e o maior deles possui 405 palavras. Com a taxa de compressão de 70% sobre o maior texto, os extratos e os *abstracts* dessa coleção deveriam ter aproximadamente 122 palavras. Foi permitida uma tolerância de 10 palavras para mais ou para menos em relação ao tamanho especificado. Assim, para a coleção 23, os pesquisadores poderiam criar sumários com tamanhos que poderiam variar entre 112 e 132 palavras.

Outra restrição importante foi que cada pesquisador deveria evitar ao máximo copiar qualquer parte do texto-fonte quando o sumário em foco era do tipo *abstract*. No caso do extrato, os sumarizadores tiveram que selecionar sentenças completas para formar o sumário, incluindo, ao final de cada uma, sua identificação de origem, isto é, sua numeração no texto-fonte. Essa identificação já estava associada a cada sentença de todos os textos fornecidos aos pesquisadores. Tal identificação, ajudará os pesquisadores na recuperação de informações presentes no córpus CSTNews, caso necessário.

4 Caracterização Quantitativa dos Sumários

A Tabela 1 mostra, para cada coleção (Col.), os tipos de sumários (TS) construídos, o tamanho médio (TM) em número de palavras dos sumários obtidos, a variação da quantidade de palavras (VP) utilizadas pelos pesquisadores (que corresponde à diferença de tamanho entre o maior e o menor sumário), a quantidade de sentenças (QS) dos textos-fonte que mais foram utilizadas na construção dos extratos de sua respectiva coleção, a porcentagem de extratos (%Ext) em que ocorrem a(s)

sentença(s) de maior uso (dada pela coluna QS), o número médio de sentenças dos sumários (NMS), e a variação da quantidade de sentenças (VS) utilizadas pelos pesquisadores (também correspondente à diferença entre o maior e o menor sumário).

Tabela 1. Dados dos sumários produzidos

Col	**TS**	**TM**	**VP**	**QS**	**%Ext**	**NMS**	**VS**
C1	*Abstract*	58	13	-	-	3,2	1
	Extrato	57	9	1	60	3,0	2
C2	*Abstract*	128	12	-	-	5,6	1
	Extrato	129	12	8	40	5,2	2
C3	*Abstract*	182	14	-	-	8,4	3
	Extrato	174	12	2	80	7,8	2
C4	*Abstract*	106	17	-	-	4,6	3
	Extrato	105	8	1	100	4,0	0
C5	*Abstract*	132	14	-	-	5,8	3
	Extrato	130	16	1	100	4,4	1
C6	*Abstract*	108	10	-	-	4,0	2
	Extrato	107	12	3	60	5,0	2
C7	*Abstract*	116	18	-	-	4,8	2
	Extrato	117	7	3	60	4,2	1
C8	*Abstract*	78	14	-	-	4,0	2
	Extrato	78	10	1	60	3,2	1
C9	*Abstract*	127	12	-	-	4,6	3
	Extrato	130	14	1	60	4,8	4
C10	*Abstract*	142	18	-	-	7,2	3
	Extrato	138	11	1	80	4,6	1
C11	*Abstract*	161	16	-	-	6,6	3
	Extrato	160	17	5	60	8,4	4
C12	*Abstract*	102	15	-	-	4,6	2
	Extrato	106	10	1	60	3,8	2
C13	*Abstract*	109	14	-	-	4,8	2
	Extrato	111	9	1	80	4,2	1
C14	*Abstract*	86	18	-	-	4,8	3
	Extrato	77	14	1	80	2,4	1
C15	*Abstract*	83	19	-	-	4,4	1
	Extrato	78	15	1	60	3,6	1
C16	*Abstract*	154	9	-	-	6,6	5
	Extrato	151	11	2	80	6,4	1
C17	*Abstract*	175	19	-	-	6,6	3
	Extrato	177	15	3	80	6,6	3
C18	*Abstract*	206	10	-	-	11,4	5
	Extrato	202	17	1	80	9,8	4
C19	*Abstract*	53	11	-	-	3,2	1
	Extrato	49	2	1	100	2,0	0
C20	*Abstract*	138	15	-	-	6,6	3
	Extrato	133	14	1	100	5,2	2
C21	*Abstract*	146	20	-	-	7,4	3
	Extrato	145	9	2	80	6,6	1
C22	*Abstract*	119	13	-	-	6,6	5
	Extrato	113	9	1	80	5,8	4
C23	*Abstract*	128	5	-	-	6,2	4
	Extrato	123	14	1	100	4,0	0
C24	*Abstract*	88	10	-	-	4,0	2
	Extrato	83	9	2	60	3,6	1
C25	*Abstract*	169	20	-	-	9,4	3
	Extrato	170	13	4	80	8,2	2
C26	*Abstract*	182	18	-	-	8,8	2
	Extrato	178	17	4	60	7,2	3
C27	*Abstract*	191	18	-	-	9,4	3
	Extrato	182	17	1	60	9,0	5
C28	*Abstract*	86	13	-	-	4,0	2
	Extrato	88	16	1	60	3,4	1
C29	*Abstract*	154	8	-	-	6,4	4
	Extrato	145	9	7	40	5,2	2
C30	*Abstract*	134	10	-	-	6,2	5
	Extrato	131	14	2	60	4,0	2
C31	*Abstract*	46	9	-	-	2,4	1
	Extrato	45	10	1	100	2,0	0
C32	*Abstract*	153	11	-	-	7,2	3
	Extrato	152	16	2	80	8,6	3
C33	*Abstract*	285	12	-	-	10,8	9
	Extrato	282	16	2	80	11,0	9
C34	*Abstract*	164	14	-	-	8,0	5
	Extrato	162	20	2	80	6,8	2
C35	*Abstract*	134	18	-	-	7,0	3
	Extrato	132	9	3	60	5,0	0
C36	*Abstract*	227	20	-	-	14,0	7
	Extrato	228	11	4	60	11,8	4
C37	*Abstract*	82	13	-	-	5,6	2
	Extrato	85	13	2	80	4,0	2
C38	*Abstract*	89	11	-	-	4,2	3

	Extrato	84	17	1	80	3,2	1
C39	*Abstract*	95	18	-	-	3,4	2
	Extrato	93	19	2	40	2,6	1
C40	*Abstract*	108	14	-	-	4,2	1
	Extrato	101	18	1	40	4,6	1
C41	*Abstract*	131	18	-	-	5,8	5
	Extrato	137	13	1	100	5,4	2
C42	*Abstract*	186	19	-	-	7,4	4
	Extrato	184	14	4	60	5,8	1
C43	*Abstract*	168	18	-	-	7,2	2
	Extrato	167	13	1	80	5,2	3
C44	*Abstract*	156	15	-	-	7,6	5
	Extrato	156	15	1	100	6,0	2
C45	*Abstract*	161	15	-	-	8,6	4

	Extrato	168	8	1	80	7,2	2
C46	*Abstract*	89	10	-	-	6,4	2
	Extrato	82	13	1	100	3,4	1
C47	*Abstract*	162	15	-	-	7,0	3
	Extrato	158	10	5	40	4,2	1
C48	*Abstract*	135	20	-	-	6,8	4
	Extrato	133	18	1	100	6,6	2
C49	*Abstract*	127	18	-	-	5,0	2
	Extrato	131	11	3	60	5,0	4
C50	*Abstract*	186	16	-	-	8,8	3
	Extrato	181	10	7	40	7,2	2
Mé-dia	***Abstract***	**134**	**14,5**	**-**	**-**	**6,3**	**3,1**
	Extrato	**133**	**12,4**	**2,2**	**72**	**5,4**	**2,0**

De acordo com a tabela, em 35 coleções, o tamanho médio dos *abstracts* foi maior do que os extratos. Tal resultado advém da maior liberdade na construção dos *abstracts*. Entretanto, na média geral, o tamanho dos *abstracts* foi similar ao dos extratos. Devido a tolerância de 10 palavras no tamanho dos sumários, calculamos a variação média do tamanho dos sumários. Podemos observar que houve variação alta de tamanho tanto para *abstracts* quanto para extratos (conforme coluna VP na tabela). Tal dado mostra a importância do uso da tolerância no tamanho dos sumários, principalmente para a criação dos extratos (mais restrição do que os *abstracts*), já que a maioria dos sumários extrativos tiveram seus tamanhos acima da taxa de compressão utilizada em cada coleção. Isso mostra que houve uma certa dificuldade por parte dos pesquisadores em produzir bons extratos informativos dentro de um espaço reduzido.

Observa-se, nas colunas QS e %Ext, que todos os extratos produzidos para 10 coleções tiveram 1 sentença em comum, e, na maioria desses casos, foi a primeira sentença de um dos textos-fonte de suas respectivas coleções. Vê-se também que não há casos de 2 ou mais sentenças em comum em todos os sumários. Há também extremos, em que 7 ou 8 sentenças são comuns a uma parcela (não todos) dos sumários (veja, por exemplo, as coleções 2 e 50). Esses dados indicam que grande parte da informação principal estava contida no início do texto-fonte e foi utilizada para compor o extrato.

Outro dado interessante, representado pela coluna NMS, é que a maioria das coleções (42) tiveram os *abstracts* com uma média de sentenças superior aos extratos (6), e em apenas 2 coleções o número médio de sentenças tanto dos *abstracts* quanto dos extratos foi igual. O comportamento é similar quando analisamos a variação do número de sentenças (coluna VS), sendo a coleção 33 a que teve a maior variação de sentenças para *abstracts* e extratos (variação de 9 sentenças). Esses dados já eram esperados, uma vez que os pesquisadores tiveram uma liberdade maior na criação dos *abstracts*, consequentemente podendo produzir sentenças mais curtas e com altas variações na quantidade das mesmas entre os sumarizadores.

5 Discussão e Considerações Finais

Com a criação dos novos sumários, cada coleção de textos do CSTNews contém, agora, 6 sumários abstrativos e 6 sumários extrativos, o que constitui um aumento significativo na quantidade de dados de referência em relação ao que se tinha anteriormente. Esses dados devem subsidiar novas pesquisas na área de SAM e permitir maior acurácia na medição dos resultados das pesquisas atuais.

Como ilustração, o Quadro 1 mostra os 5 novos *abstracts* produzidos para a coleção 1 do córpus, com 3 textos sobre um acidente aéreo, exibidos no Quadro 2. Pode-se ver, nos *abstracts*, a variação em se descrever as informações relacionadas ao acidente, sendo que todos os *abstracts* tiveram em comum a informação da quantidade de vítimas e o local do acidente. O *abstract* 3 foi o único que não trouxe de forma mais clara a causa do acidente. Nota-se que a informação sobre a carga de minerais que o avião também transportava ocorre em 3 *abstracts* (1, 3 e 5). Além disso, o *abstract* 4 é o único que faz menção ao histórico de acidentes aéreos na região. Assim, mesmo sendo *abstracts* pequenos, vê-se a pluralidade de formas de sintetização de informação.

Quadro 1. Sumários abstrativos para a coleção 1 do córpus CSTNews

Abstract 1. Nesta quinta-feira aconteceu um acidente aéreo na região de Bukavu, na República Democrática do Congo, que acabou com a vida de 17 pessoas, entre elas 14 passageiros e três membros da tripulação. Uma porta-voz da ONU informou que o avião não conseguiu aterrissar no aeroporto por causa da tempestade, caindo na floresta e explodindo. O avião também levava uma carga de minerais.
Abstract 2. A queda de um avião no Congo matou as 17 pessoas a bordo, sendo 14 passageiros e 3 tripulantes. De fabricação russa e propriedade de uma empresa congolesa, o voo saiu da cidade de Lugushwa com destino a Bukavu. Por causa do mau tempo, a aeronave caiu em uma floresta localizada a 15 km do aeroporto de destino, explodindo e se incendiando.
Abstract 3. Após a queda e explosão de um avião na localidade de Bukavu, no leste da República Democrática do Congo, todos os passageiros e tripulantes morreram, totalizando 17 mortes. O avião caiu em uma floresta a 15 km do aeroporto. Além de passageiros, o avião também levava uma carga de minerais.
Abstract 4. A queda de um avião em Bukavu, na República Democrática do Congo, resultou na morte de 17 pessoas, sendo 14 passageiros e 3 membros da tripulação. O avião, da companhia congolesa Trasept Congo, foi prejudicado pelo mau tempo e não conseguiu chegar à pista de aterrisagem, caindo em uma floresta a 15km do aeroporto. O Congo tem um grave histórico de acidentes aéreos.
Abstract 5. Um acidente aéreo causou a morte de 17 pessoas, na República Democrática do Congo. Segundo um porta-voz da ONU, o avião que tentava aterrissar durante uma tempestade, explodiu ao se chocar com uma montanha, a 15 Km do aeroporto de Bukavu. Operado pela Trasept Congo, o avião também levava uma carga de minerais. Não houve sobreviventes.

Quadro 2. Textos da coleção 1 do córpus CSTNews

Texto 1

Ao menos 17 pessoas morreram após a queda de um avião de passageiros na República Democrática do Congo.

Segundo uma porta-voz da ONU, o avião, de fabricação russa, estava tentando aterrissar no aeroporto de Bukavu em meio a uma tempestade.

A aeronave se chocou com uma montanha e caiu, em chamas, sobre uma floresta a 15 quilômetros de distância da pista do aeroporto.

Acidentes aéreos são freqüentes no Congo, onde 51 companhias privadas operam com aviões antigos principalmente fabricados na antiga União Soviética.

O avião acidentado, operado pela Air Traset, levava 14 passageiros e três tripulantes.

Ele havia saído da cidade mineira de Lugushwa em direção a Bukavu, numa distância de 130 quilômetros.

Aviões são usados extensivamente para transporte na República Democrática do Congo, um vasto país no qual há poucas estradas pavimentadas.

Em março, a União Européia proibiu quase todas as companhias aéreas do Congo de operar na Europa. Apenas uma manteve a permissão.

Em junho, a Associação Internacional de Transporte Aéreo incluiu o Congo num grupo de vários países africanos que classificou como uma vergonha para o setor.

Texto 2

Um acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo (RDC), matou 17 pessoas na quinta-feira à tarde, informou nesta sexta-feira um porta-voz das Nações Unidas.

As vítimas do acidente foram 14 passageiros e três membros da tripulação. Todos morreram quando o avião, prejudicado pelo mau tempo, não conseguiu chegar à pista de aterrissagem e caiu numa floresta a 15 quilômetros do aeroporto de Bukavu. Segundo fontes aeroportuárias, os membros da tripulação eram de nacionalidade russa.

O avião explodiu e se incendiou, acrescentou o porta-voz da ONU em Kinshasa, Jean-Tobias Okala. "Não houve sobreviventes", disse Okala. O porta-voz informou que o avião, um Soviet Antonov-28 de fabricação ucraniana e propriedade de uma companhia congolesa, a Trasept Congo, também levava uma carga de minerais.

Texto 3

Um acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo, matou 17 pessoas na quinta-feira à tarde, informou hoje um porta-voz das Nações Unidas.

As vítimas do acidente foram 14 passageiros e três membros da tripulação. Todos morreram quando o avião, prejudicado pelo mau tempo, não conseguiu chegar à pista de aterrissagem e caiu numa floresta a 15 Km do aeroporto de Bukavu.

O avião explodiu e se incendiou, acrescentou o porta-voz da ONU em Kinshasa, Jean-Tobias Okala. "Não houve sobreviventes", disse Okala.

O porta-voz informou que o avião, um Soviet Antonov-28 de fabricação ucraniana e propriedade de uma companhia congolesa, a Trasept Congo, também levava uma carga de minerais.

Segundo fontes aeroportuárias, os membros da tripulação eram de nacionalidade russa.

Por fim, em relação ao córpus recentemente divulgado PCSC [7], para o português europeu, é interessante observar as diferenças existentes. Apesar do CSTNews ser consideravelmente menor (tendo 140 textos frente aos 801 disponíveis no PCSC), ele é ricamente anotado e, agora, passa a ter 12 sumários de referência por coleção (frente aos 2 sumários disponibilizados pelo PCSC). Também é importante notar a diferença na forma de produção dos sumários: no PCSC, os sumários devem ser produzidos por compressão, realizada necessariamente pela supressão de sentenças e palavras. Acredita-se que, juntos, o CSTNews e o PCSC atendam melhor a comunidade de pesquisa e possam servir a fins variados.

Agradecimentos

À FAPESP, à CAPES, ao CNPq e à Universidade Federal de Goiás, pelo apoio a este trabalho.

Referências

1. Aleixo, P.; Pardo, T.A.S. *CSTNews: um córpus de textos jornalísticos anotados segundo a teoria discursiva multidocumento CST (Cross-Document Structure Theory)*. Relatório Técnico, NILC-ICMC-USP. 12p. (2008)
2. Cardoso, P.C.F.; Maziero, E.G.; Jorge, M.L.C.; Seno, E.M.R.; Di Felippo, A.; Rino, L.H.M.; Nunes, M.G.V.; Pardo, T.A.S. 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*, pp. 88-105. October 26, Cuiabá/MT, Brazil. (2011)
3. Castro Jorge, M.L.R.; Pardo, T.A.S. Experiments with CST-based Multidocument Summarization. In the *Proceedings of the ACL Workshop TextGraphs-5: Graph-based Methods for Natural Language Processing*, pp. 74-82. July 16, Uppsala/Sweden. (2010)
4. Mani, I. *Automatic Summarization*. John Benjamins Publishing Co. (2001)
5. Ribaldo, R.; Akabane, A.T.; Rino, L.H.M.; Pardo, T.A.S. Graph-based Methods for Multi-document Summarization: Exploring Relationship Maps, Complex Networks and Discourse Information. In the *Proceedings of the 10th International Conference on Computational Processing of Portuguese* (LNAI 7243), pp. 260-271. April 17-20, Coimbra, Portugal. (2012)
6. Cardoso, P.C.F. *Exploração de metódos de sumarização automática multidocumento com base em conhecimento semântico-discursivo*. Tese de Doutorado. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos-SP. (2014)
7. Almeida, M.B.; Almeida, M.S.C.; Martins, A.F.T.; Figueira, H.; Mendes, P.; Pinto, C. A New Multi-Document Summarization Corpus for European Portuguese. In the *Proceedings of the Language Resources and Evaluation Conference*, pp.146-152. May 26-31, Reykjavik, Iceland. (2014)
8. Silveira, S.; Branco, A. Combining a double clustering approach with sentence simplification to produce highly informative multi-document summaries. In the *Proceedings of the International Conference on Information Reuse and Integration*, pp. 482-489. August 8-9, Las Vegas, USA. (2012)