Núcleo Interinstitucional de Linguística Computacional - NILC

Anotação de subtópicos do córpus multidocumento CSTNews

Relatório Técnico

NILC-TR-12-07

Paula C. F. Cardoso, Amanda P. Rassi, Erick G. Maziero, Fernando A. A. Nóbrega, Jackson W. C. Souza, Márcio S. Dias, Maria Lúcia R. Castro Jorge, Pedro P. Balage Filho, Renata T. Camargo, Verônica Agostini, Ariani Di Felippo, Lucia H. M. Rino, Thiago A. S. Pardo

Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional
NILC-ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil

Resumo

A segmentação topical visa a dividir um texto em segmentos topicalmente coerentes. Esse procedimento pode ser muito útil para aplicações de Processamento de Linguagem Natural, tais como recuperação de informação, sumarização automática e sistemas de perguntas e respostas. Assume-se que um texto tem um tópico principal, que é o assunto sobre o qual se escreve ou discute, e que esse assunto pode ser descrito em uma sequência de subtópicos. Tais subtópicos podem mudar continuamente, sendo que algumas mudanças são mais sutis do que outras. Visando criar uma segmentação de subtópicos de referência, este relatório descreve o processo de anotação de subtópicos do córpus CSTNews, um córpus multidocumento de notícias jornalísticas em português do Brasil. As diretrizes de anotação e seus resultados são apresentados e discutidos. Esta segmentação foi desenvolvida para fins de investigação na área de Sumarização Automática de textos.

O trabalho relatado contou com o apoio da CAPES, da FAPESP e do CNPq.

ÍNDICE

1. INTRODUÇÃO

Koch (2009) afirma que um texto compõe-se de segmentos topicais, direta ou indiretamente relacionados com o tema geral ou tópico discursivo. Um segmento topical, quando introduzido, mantém-se por um determinado tempo, após o qual, com ou sem intervalo de transição, ocorre a introdução de um novo segmento topical. O processo de segmentação topical, portanto, visa dividir um texto em segmentos topicalmente coerentes. A segmentação topical pode ser muito útil para várias aplicações de Processamento de Linguagem Natural (PLN), por exemplo, recuperação de informação, sistemas de perguntas e respostas e sumarização automática.

Prince e Labadié (2007) explicam que, para a recuperação de informação, que busca documentos relevantes para uma dada consulta do usuário, além desses documentos, podem ser fornecidos fragmentos de texto que são semanticamente e topicalmente relacionados a uma dada consulta. Isso facilita para que o usuário encontre rapidamente a informação de seu interesse. Para Oh et al. (2007), um sistema de perguntas e respostas, cujo objetivo é responder uma pergunta/consulta submetida pelo usuário, pode mapear essa consulta para os subtópicos, de forma a facilitar a localização da resposta de forma mais rápida. A sumarização automática, que visa produzir uma representação concisa de um ou mais textos, também pode se beneficiar da segmentação topical. Sabe-se que uma coleção de textos contém uma variedade de informações que cobrem diferentes aspectos de um tópico principal. A partir da segmentação topical, os sistemas de sumarização podem criar sumários que selecionem diferentes aspectos da coleção de textos.

Na Figura 1, apresenta-se um exemplo de texto segmentando topicalmente. O tópico principal do texto é “Cielo leva ouro nos 100m nos Estados Unidos’. Pode ser observado que o texto traz distintas informações que sugerem a segmentação indicada na figura. O subtópico que fala de “Cielo” é apresentado nos parágrafos 1 e 5, ou seja, “Cielo” aparece no início e no fim do texto. Isso indica que um subtópico que já foi descrito pode voltar após um determinado tempo. O subtópico “Equipe brasileira” é formado por mais de um parágrafo, o que indica que a granularidade de um subtópico é variada.

Subtópico	Texto
Cielo	**[P1]** O nadador brasileiro César Cielo confirmou o favoritismo neste domingo e conquistou o segundo ouro no Grand Prix de natação do Missouri, nos Estados Unidos. O campeão olímpico venceu a final dos 100 m livre fazendo 50s57, à frente do canadense Richard Hortness e do americano Matt Grevers.
Nicolas	**[P2]** Nicolas Oliveira foi outro brasileiro a competir na final dos 100 m livre, mas ficou longe do pódio: terminou apenas no sexto lugar, à frente de Jason Lezak e de Colin Russell.
Equipe brasileira	**[P3]** A vitória encerra a participação brasileira na competição. Ao todo, foram conquistadas 10 medalhas - Cielo também venceu os 50 m livre. Além dele, ficaram com o ouro Felipe Lima, nos 100 m peito; Thiago Pereira, nos 200 m medley; e Joanna Maranhão, nos 200 m borboleta. O país ainda conquistou duas de prata e três de bronze. **[P4]** Os nadadores brasileiros encerram, assim, o período de treinamento no exterior de olho na Olimpíada de Londres 2012. Os atletas permaneceram trabalhando em La Loma, localizado em San Luis Potosí, no México, onde fizeram a preparação para os Jogos Pan-Americanos de Guadalajara, em 2011.
Cielo	**[P5]** "Essa competição era um desafio a mais para a cabeça dele, o de tentar levar o corpo brigando pelo primeiro lugar em todas as provas. Está muito cansado, era mesmo um desafio", exaltou Albertinho, técnico de Cielo no Projeto Rumo ao Ouro 2016.

Figura 1 – Exemplo de texto segmentado topicalmente

Como se pode notar, um texto tem um tópico principal, que é o assunto sobre o qual se escreve ou discute (Biryukov et al., 2005; Hovy e Lin, 1998; Lin, 1995). Um tópico pode ser descrito em uma sequência de discussões divididas em subtópicos (Hennig, 2009; Hearst, 1997). Os subtópicos de um discurso mudam continuamente; algumas mudanças são sutis, outras são mais proeminentes (Kazantseva e Szpakowicz, 2012). Algumas vezes, a estrutura de subtópicos é marcada em textos técnicos por cabeçalhos de seções que dividem o texto em segmentos coerentes. Entretanto, há textos que quase não possuem uma marcação explicita de subtópicos, por exemplo, textos jornalísticos. Nesses casos, os subtópicos podem ser formados por uma ou mais sentenças ou parágrafos. Outro ponto a ser considerado é que os segmentos de um mesmo subtópico nem sempre são adjacentes. Eles são constantemente entremeados por outros segmentos, sejam esses inserções de outros subtópicos ou inserções parentéticas (Pinheiro, 2008).

Outro exemplo de um texto segmentado topicalmente é dado na Figura 2. O tópico principal do texto é “Maradona tem recaída e médico descarta pancreatite”. Além do subtópico que descreve que “Maradona teve uma recaída devido a hepatite”, fala-se

também sobre o “estado atual do jogador” e por último, relata-se que “o jogador recebeu mensagens de apoio por parte dos torcedores”.

Subtópico	Texto
Recaída de Maradona	Maradona voltou a ter problemas de saúde no fim de semana. Internado em um hospital em Buenos Aires, ele teve uma recaída e voltou a sentir dores devido a hepatite aguda que o atinge, segundo seu médico pessoal, Alfredo Cahe.
Estado atual	"Agora está estável. Mesmo com esta melhora, ele continuará internado", disse o médico, que descartou a possibilidade do ex-jogador ter uma pancreatite (inflamação do pâncreas, órgão situado atrás do estômago e que influencia na digestão). Cahe reforçou que Maradona ainda tem problemas. "Os valores hepáticos dele na avaliação não estão equilibrados e ele não está bem. Mas não é nada grave", afirma, em entrevista ao diário La Nación.
Mensagens de apoio	No domingo, Maradona assistiu ao empate por 1 a 1 no clássico Boca Juniors e River Plate pela televisão. Os torcedores do Boca, que compareceram em grande número ao Estádio La Bombonera, levaram muitas faixas e bandeiras com mensagens de apoio ao ídolo argentino. Sua filha, Dalma, foi ao estádio assistir ao jogo.

Figura 2 – Exemplo de texto segmentado topicalmente

Existem vários trabalhos de segmentação topical automática, contudo, sabe-se que a tarefa é pouco definida mesmo do ponto de vista humano. Juízes humanos anotando um mesmo texto utilizam critérios diferentes para identificar as mudanças de subtópicos. Algumas pessoas marcam somente as mudanças mais evidentes, enquanto outras incluem mudanças mais finas.

Neste relatório, apresentamos o processo de anotação topical de um córpus de textos jornalísticos em português do Brasil. O córpus, chamado CSTNews[1] (Aleixo e Pardo, 2008; Cardoso et al., 2011), é composto por 140 textos de diferentes seções e diferentes fontes. A anotação, que consiste na segmentação topical dos textos, visa subsidiar pesquisas tanto do processo humano de segmentação quanto para a produção de ferramentas automáticas de processamento textual.

Em particular, este trabalho faz parte do projeto SUCINTO (*summarization for clever information access*)[2], que visa investigar e explorar técnicas de sumarização

[1]Disponível em http://www.icmc.usp.br/~taspardo/sucinto/cstnews.html
[2]http://www.icmc.usp.br/~taspardo/sucinto

multidocumento, além de tarefas relacionadas como análise do discurso, ordenação temporal de informação, resolução de correferência e processamento multilíngue.

O relatório está organizado em 4 seções. Na Seção 2, é feita uma descrição linguística sobre tópico discursivo. Na Seção 3, descrevem-se as características do córpus CSTNews. Na Seção 4, descreve-se o processo de anotação de subtópicos. Na Seção 5, apresentam-se os resultados da anotação. Por fim, são feitas algumas considerações finais.

2. CARACTERIZAÇÃO LINGUÍSTICA DE TÓPICO

A noção de tópico discursivo é descrita nos estudos linguísticos brasileiros inicialmente por Jubran et al. (1992). Conforme os autores, o tópico é uma categoria abstrata, primitiva, que se manifesta na conversação, mediante enunciados formulados pelos interlocutores a respeito de um conjunto de referentes explícitos ou inferíveis, concernentes entre si e em relevância num determinado ponto da mensagem.

Para Jubran (2006), a organização topical é manifestada pela natureza das articulações que um tópico tem com os outros na sequência discursiva, bem como pelas relações hierárquicas entre tópicos mais ou menos abrangentes. Nesse sentido, a organização topical pode ser observada em dois níveis: no plano hierárquico e no plano sequencial. No plano hierárquico, as sequências textuais se desdobram conforme as dependências de super ou subordenação entre tópicos que se implicam pelo grau de abrangência com que são tratados na interlocução. No plano sequencial (progressão topical), dois processos básicos caracterizam a distribuição de tópicos na linearidade discursiva: a continuidade e descontinuidade. A continuidade se caracteriza por uma relação de adjacência entre dois tópicos, com abertura de um tópico subsequente somente quando o anterior é esgotado. A descontinuidade (mudança de tópico) se caracteriza por uma perturbação da sequencialidade linear, causada ou por uma suspensão definitiva de um tópico, ou pela divisão do tópico, que passa a se apresentar em partes descontínuas.

Koch (2009) afirma que a topicalidade constitui um princípio organizador do discurso. Segundo a autora, para que um texto possa ser considerado coerente, é preciso que apresente continuidade topical, ou seja, que a progressão topical em ambos os níveis se realize de forma que não ocorram rupturas definitivas ou interrupções excessivamente longas do tópico em andamento.

De acordo com Jubran et al. (1992), a mudança de tópico pode ocorrer de três formas: após a finalização do anterior, de forma gradativa; por meio de tópicos de transição, que não se encaixam, portanto, em nenhum outro; e pela ruptura, sem que haja, dessa forma, esgotamento do anterior. Entretanto, as partes de um texto são interdependentes, sendo cada uma necessária para a compreensão das demais. Um dos fatores que contribuem para essa interdependência são os mecanismos de sequenciação (coesão sequencial) existentes na língua. Koch (1998) define dois mecanismos de coesão sequencial: sequenciação parafrástica (com procedimentos de recorrência) e sequenciação frástica (sem procedimentos de recorrência estrita).

A sequenciação parafrástica utiliza-se de procedimentos de recorrência ou retomada, tais como: recorrência de termos (por exemplo, repetição, anáfora, sinonímia, hiperonímia e uso de palavra genérica), de estruturas (paralelismo sintático), de conteúdos semânticos (paráfrase), de recursos fonológicos segmentais e/ou suprassegmentais e recorrência de tempo e aspecto verbais. A sequenciação frástica, por sua vez, não se utiliza de procedimentos de recorrência; ocorre mediante procedimentos de manutenção temática (por meio do uso de termos pertencentes ao mesmo campo lexical), progressão temática (concretizada por meio dos blocos comunicativos tema/rema, tópico/comentário, dado/novo) e encadeamento (construído por meio de justaposição ou de conexão).

Carlson e Marcu (2001) definem duas relações discursivas que identificam a mudança ou não de tópico. Quando há fortes indícios de mudança de tópico entre duas sentenças ou dois parágrafos, ocorre a relação *topic-shift*. Por outro lado, quando a mudança de tópicos entre duas sentenças ou dois parágrafos é suave, a relação que ocorre é *topic-drift*. Relacionando ao que foi dito por Jubran et al. (1992), a relação *topic-shift* indica descontinuidade e a relação *topic-drift* indica continuidade.

É interessante notar que muitos desses trabalhos não fazem a distinção entre tópicos e subtópicos, usando os termos de forma intercambiável. Neste relatório, faz-se essa distinção, conforme se apresenta na seção introdutória. Tal distinção está de acordo com o trabalho renomado de segmentação topical de Hearst (1997).

A seguir, apresenta-se brevemente o córpus CSTNews, sobre o qual este estudo é conduzido.

3. O CÓRPUS CSTNews

O CSTNews é o maior córpus multidocumento com textos em português do Brasil que se tem conhecimento. A coleta manual aconteceu durante aproximadamente 60 dias, de Agosto a Setembro de 2007. O CSTNews contém 50 grupos de textos jornalísticos de assuntos variados,coletados manualmente das fontes de notícias Folha de São Paulo, Estadão, O Globo, Jornal do Brasil e Gazeta do Povo. Segundo Aleixo e Pardo (2008), essas fontes foram escolhidas devido a grande popularidade na web e também por trazerem as principais notícias do dia corrente. A escolha de textos do gênero jornalísticos se deve ao fato desses possuírem uma linguagem clara e do dia a dia.

Cada grupo de textos contém de 2 a 3 textos sobre um mesmo assunto e seus respectivos sumários humanos e automáticos, além de diversas anotações. Os grupos de textos estão organizados pelos rótulos das seções dos jornais dos quais os textos foram compilados. O critério utilizado para selecionar as notícias foi com base na relevância no momento (dado que deveriam ser noticiadas por diferentes fontes). Isso levou a distribuição de grupos de textos apresentada na Figura 3.

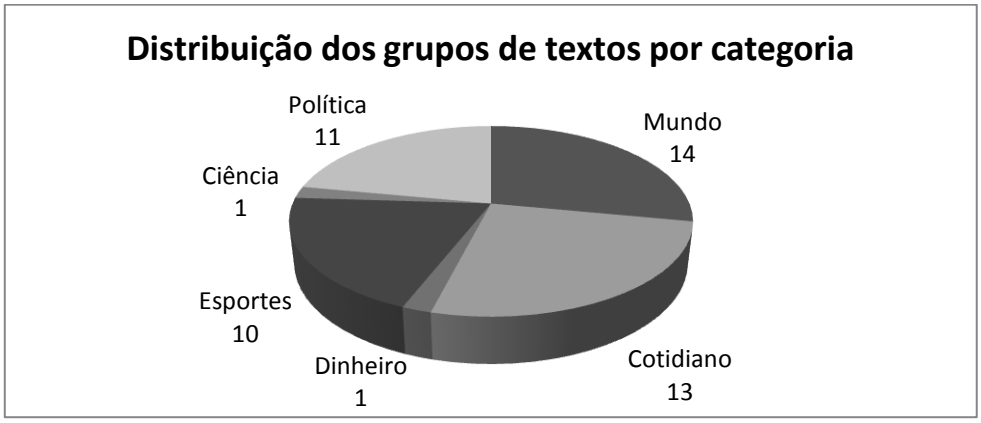

Figura 3 – Distribuição de textos por categoria no CSTNews

O córpus é manualmente anotado de diferentes maneiras. Cada texto-fonte tem a representação discursiva com base na teoria *Rhetorical Structure Theory* (RST) (Mann e Thompson, 1987). O relacionamento entre os textos-fonte é baseado na teoria de relacionamento multidocumento *Cross-document Structure Theory* (CST) (Radev, 2000).

Os sumários manuais são abstrativos e seguem uma taxa de 70%. Consequentemente, os sumários manuais monodocumento contêm no máximo 30% do número de palavras do seu texto-fonte e os sumários manuais multidocumento contêm

30% do número de palavras do maior texto da coleção. No córpus, também há sumários humanos extrativos e automáticos.

Outras anotações também se encontram disponíveis no córpus, como o alinhamento entre os sumários e seus textos-fonte correspondentes (ou seja, a indicação da origem das informações de cada sumário), a delimitação e identificação das expressões temporais nos textos e a indexação dos substantivos dos textos a seus sentidos em uma ontologia, dentre outras.

A seguir, descreve-se o processo de anotação topical do córpus CSTNews.

4. PROCESSO DE ANOTAÇÃO DE SUBTÓPICOS

Para anotação de subtópicos do CSTNews, participaram 14 anotadores com formação em linguística computacional. Durante 3 dias, os anotadores receberam um treinamento de segmentação topical, que foi baseado na metodologia de Hearst (1997). Durante esta etapa, foram apresentados e discutidos os conceitos de tópico e subtópicos, e as regras de segmentação foram definidas.

As regras de segmentação topical adotadas foram as seguintes:

- Acrescentar após cada mudança de subtópico a seguinte marcação: <t rotulo= “breve descrição do subtópico acima”>. A breve descrição deveria ser na forma de palavras-chave.
- A anotação era individual.
- Os títulos dos textos não seriam utilizados. Isso se deve ao fato de que nem todos os textos do córpus tem o título disponível.
- Um mesmo texto deveria ser anotado por no mínimo 5 pessoas. Nesse caso, as marcas indicadas por pelo menos 3 anotadores (a maioria) eram consideradas como a anotação final daquele texto.
- Granularidade não era definida, ou seja, um subtópico poderia ser formado por 1 ou mais sentenças.
- Não era permitido delimitar tópicos dentro de sentenças, ou seja, a anotação deveria ser intersentencial.
- Não era permitido diálogo entre anotadores.

Após o treinamento, 7 dias foram necessários para a anotação propriamente dita do córpus. Reuniões diárias foram organizadas, com duração de 1 hora, aproximadamente.

Os anotadores eram organizados em grupos de 5 e 7 pessoas cada. Cada grupo recebia um pacote com 10 textos por dia. Os grupos eram formados aleatoriamente, para evitar possíveis tendências nas anotações. Os anotadores utilizaram o editor de textos de sua preferência para fazer a anotação.

A seguir, apresentam-se os resultados da anotação.

5. RESULTADOS

A fim de certificar que as instruções foram suficientemente claras e os anotadores de fato anotaram o mesmo fenômeno, foi medida a concordância entre anotadores por meio da medida Kappa (Carletta, 1996). A kappa é uma medida clássica de concordância em PLN, indicando a concordância entre anotadores ao mesmo tempo em que desconta a concordância ao acaso, por sorte. Apesar de não existir um valor específico (referenciado por k) a partir do qual se deva considerar o valor da kappa como adequado, encontram-se na literatura algumas sugestões que orientam esta decisão: um valor menor do que 0.4 pode indicar uma anotação na qual não se pode confiar; se estiver entre 0.4 e 0.75, a anotação é satisfatória; e, se for maior do que 0.75, é muito boa.

A Tabela 1 mostra a quantidade de grupos, anotadores, textos e sentenças por dia, e seus respectivos valores kappa. Pode ser observado que, no primeiro dia de anotação, houve a melhor concordância entre anotadores, sendo k=0.656 para o grupo 1 e k=0.566 para o grupo 2. Por outro lado, a concordância mais baixa aconteceu no segundo dia, sendo k=0.458 para o grupo 1 e k=0.447 para o grupo 2. A média geral foi k=0.560; esse valor pode ser considerado satisfatório dada a subjetividade da tarefa. Por outro lado, esse valor médio de 0.560 pode ser dito baixo quando comparado com o experimento realizado por Hearst (1997), no qual se obteve k=0.611. Vale ressaltar que Hearst contou com 7 anotadores para segmentarem 12 textos expositivos (*magazine articles*) (escritos em inglês), que são textos de natureza diferente dos anotados aqui. A autora também orientou os anotadores colocarem as quebras entre parágrafos quando houvesse mudança de subtópico.

Tabela 1 – Concordância por grupo de anotadores

Dia	Grupos	Número de anotadores	Número de textos	Número de sentenças	Valor Kappa
1	Grupo 1	6	10	144	0.656
	Grupo 2	7		118	0.566
2	Grupo 1	6	10	135	0.458
	Grupo 2	6		123	0.447
3	Grupo 1	7	10	142	0.515
	Grupo 2	5		158	0.638
4	Grupo 1	5	10	203	0.544
	Grupo 2	7		171	0.562
5	Grupo 1	5	10	141	0.643
	Grupo 2	5		227	0.528
6	Grupo 1	5	12	178	0.570
	Grupo 2	5	13	231	0.549
7	Grupo 1	5	15	158	0.611
				Média	**0.560**

A partir dos textos anotados, foi criada uma segmentação de referência para cada texto. Hearst (1997) considerou a opinião de pelo menos três dos sete anotadores disponíveis. Para o córpus CSTNews, contabilizou-se a opinião da metade mais 1 de anotadores concordantes nas quebras. Para exemplificar esse cenário, a Figura 4 apresenta um texto com 7 sentenças (numeradas e referenciadas por S1 a S7) e a Figura 5 representa a anotação de 5 juízes para o mesmo texto. Na Figura 5, as linhas numeradas de 1 a 5 representam as delimitações de subtópicos feitas por cada um dos 5 anotadores, sendo que cada quadrinho representa uma das 7 sentenças e a segmentação é indicada por barras verticais. A última linha, identificada por "Final", representa os locais de quebra de maior concordância, nesse caso, por pelo menos 3 anotadores. Portanto, nesse texto, as quebras que representam a opinião da maioria são após as sentenças 5, 6 e 7.

As descrições indicadas por cada juiz após as quebras não foram utilizadas para definir a marcação final. Por outro lado, as descrições serviram para verificar se os anotadores tinham compreendido a tarefa, de forma que alguma palavra da descrição deveria estar contida no subtópico correspondente.

[S1] 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.

[S2] As vítimas do acidente foram 14 passageiros e três membros da tripulação. **[S3]** 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.

[S4] O avião explodiu e se incendiou, acrescentou o porta-voz da ONU em Kinshasa, Jean-Tobias Okala. **[S5]** "Não houve sobreviventes", disse Okala.

[S6] 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.

[S7] Segundo fontes aeroportuárias, os membros da tripulação eram de nacionalidade russa.

Figura 4 – Texto-fonte "Acidente aéreo em Bukavu"

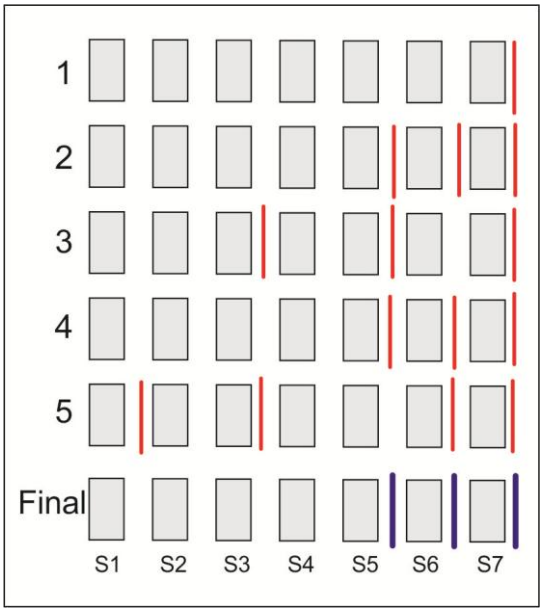

Figura 5 – Representação da segmentação topical de diferentes anotadores

Na Figura 6, mostra-se o número de subtópicos na anotação de referência. Pode ser observado que a menor quantidade de subtópicos encontrada foi de 1 subtópico em 8 textos (corresponde a 6% do córpus), 2 subtópicos em 24 textos (17%), 3 subtópicos em 50 textos (36%), 4 subtópicos em 32 textos (23%), 5 subtópicos em 18 textos (13%), 6 subtópicos em 4 textos (3%), 7 subtópicos em 2 textos (1%) e 8 subtópicos em 2 textos (1%). Portanto, a média de subtópicos por texto no córpus é 3.

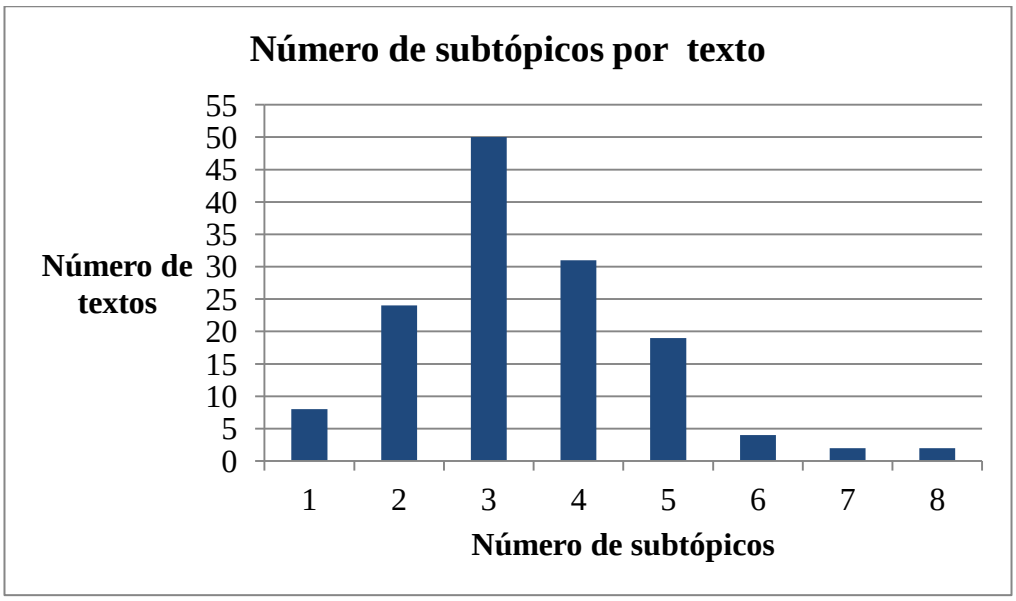

Figura 6 – Quantidade de subtópicos por documento

Para identificar os subtópicos, os anotadores podem utilizar diversos conhecimentos e isso implica em diferentes segmentações. Além disso, alguns textos são mais difíceis do que outros. A Figura 7 mostra um texto com 23 sentenças para o qual houve muitas divergências entre 5 anotadores. A Figura 8 mostra as quebras indicadas pelos anotadores para o texto da Figura 7. As quebras mais proeminentes estão após as sentenças 12, 20 e 23, pois tais quebras foram indicadas por 3 anotadores. Entretanto, verificam-se outras 5 quebras (após as sentenças 1, 7, 13, 15 e 16) indicadas por uma minoria.

[S1] Termina hoje, às 20 horas, o prazo para que os deputados acusados de participar do esquema dos sanguessugas renunciem para escapar da abertura de processo por quebra de decoro parlamentar. **[S2]** A expectativa de lideranças da Câmara e do Conselho de Ética é que pouco mais de 10% dos 69 deputados denunciados no relatório parcial da CPI dos Sanguessugas abrirão mão de seus mandatos.

[S3] Integrante da cúpula da Câmara que, nos últimos dias, conversou com ao menos 30 parlamentares acusados no caso calcula que sete podem renunciar - Nilton Capixaba (PTB-RO), Marcelino Fraga (PMDB-ES), César Bandeira (PFL-MA), Benedito Dias (PP-AP), Carlos Nader (PL-RJ), João Caldas (PL-AL) e Reginaldo Germano (PP-BA). **[S4]** Ex-líder do PP, Pedro Henry (MT) cogitou sair da função, mas teria desistido da ideia.

[S5]Até ontem, só Coriolano Sales (PFL-BA) havia apresentado renúncia. **[S6]** Ele não quis arriscar a chance de assumir a prefeitura de Vitória da Conquista. **[S7]** Segundo colocado em 2004, Sales processou seu adversário por abuso do poder econômico e aguarda resultado.

[S8] "Não dá para avaliar quantos vão renunciar", disse ontem o presidente do Conselho de Ética, Ricardo Izar (PTB-SP). **[S9]** Ele vai instaurar o processo contra os deputados envolvidos com a máfia dos sanguessugas amanhã, às 10h30. **[S10]** Formalizada antes da abertura, a renúncia cessa o procedimento. **[S11]** Izar pretende que os casos dos 15 parlamentares que receberam depósito na própria conta bancária ou na de parentes sejam os primeiros julgados pelo Conselho. **[S12]**"Vou instaurar todos os processos juntos, mas a idéia é que os 15 casos mais graves, que têm provas contundentes, sejam julgados na frente", afirmou Izar.

[S13] O horário-limite para que o parlamentar renuncie - 20 horas - foi estabelecido pela direção da Câmara a fim de que o ato seja oficializado com a sua publicação já no Diário Oficial do Congresso de amanhã. **[S14]** A maioria dos 69 deputados acusados de envolvimento com a máfia dos sanguessugas é candidato à reeleição e, com a renúncia, tentará escapar do risco de cassação e da perda dos direitos políticos por oito anos. **[S15]** Outros parlamentares resistem à renúncia por considerarem ter chance de não sofrer punição. **[S16]** Segundo investigações iniciadas pela Polícia Federal, o esquema consistia no desvio de recursos públicos por meio da apresentação de emendas parlamentares para a compra superfaturada de ambulâncias.

[S17] Dois dos 69 deputados acusados são da Mesa Diretora, mas se afastaram das funções. **[S18]** Contra João Caldas, por exemplo, pesam 12 pagamentos no total de R$ 136 mil, alguns dos quais em sua própria conta. **[S19]** Ele, porém, resiste à ideia de abrir mão do mandato. **[S20]**Já Capixaba, acusado de ter recebido R$ 646 mil, considera seriamente a hipótese de renunciar.

[S21] No caso dos integrantes da Igreja Universal, a possibilidade de saída do cargo está afastada, pois os envolvidos, entre eles Edna Macedo (PTB-SP), foram proibidos pela direção da instituição de concorrer à reeleição. **[S22]** Eleitos na esteira do deputado Enéas Carneiro (Prona-SP), ex-integrantes do partido suspeitos, como Irapuan Teixeira (PTB-PR) e Ildeu Araújo (PP-SP), tiveram votação insignificante em 2002 e não têm chance de reeleição. **[S23]** Devem preferir manter o resto do mandato.

Figura 7 – Texto-fonte "Sanguessugas"

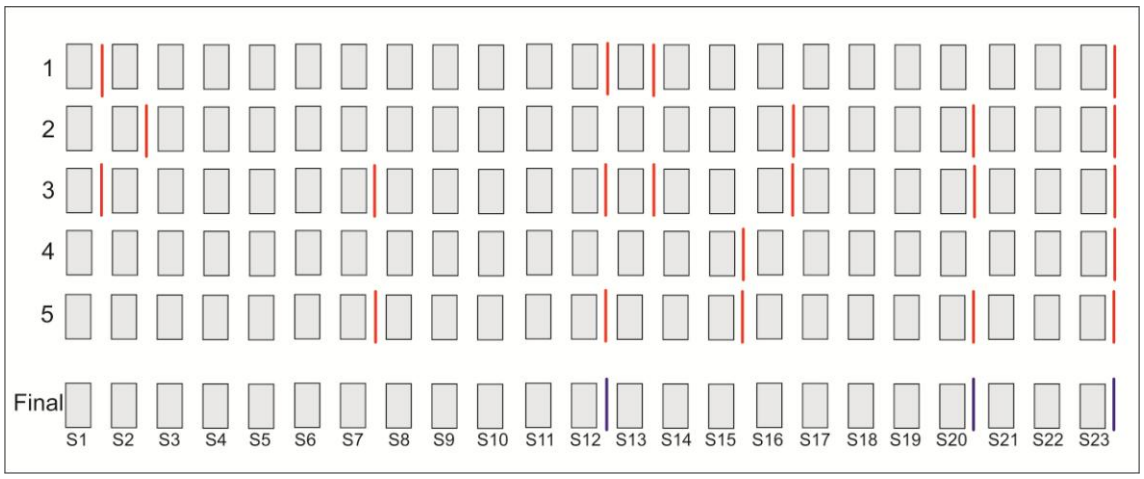

Figura 8– Várias segmentações para o texto “Sanguessugas”

A Figura 9 exibe as descrições dos anotadores para alguns dos subtópicos da Figura 8. Observa-se que os anotadores 1 e 3 colocaram descrições similares para o subtópico descrito na sentença 1. Já os anotadores 1, 3 e 4 concordaram com a quebra após a sentença 12, mas colocaram rótulos diferentes para o subtópico. Essa variedade na descrição do subtópico também reflete o fato de que as pessoas utilizam diferentes critérios/conhecimentos para delimitar segmentos topicais.

Sentença 1	Anotador 1	<t rotulo="fim do prazo para renúncia">
	Anotador 3	<t rotulo="o prazo para a renúncia">
Sentença 12	Anotador 1	<t rotulo="quem vai renunciar">
	Anotador 3	<t rotulo="fala do presidente do conselho">
	Anotador 4	<t rotulo="processo contra acusados">

Figura 9 – Descrições para os subtópicos por diferentes anotadores

A maioria das quebras topicais é entre parágrafos. Somente em 4% do córpus (6 textos) aconteceram quebras dentro de parágrafos. Acredita-se que isso está relacionado a forma como as pessoas foram ensinadas a escrever e estão habituadas a estruturar seus textos, utilizando parágrafos como um recurso para organização do texto. O parágrafo é organizado em torno de uma ideia-núcleo, que é desenvolvida por ideias secundárias. Quando se pretende mudar de assunto, deve-se iniciar um novo parágrafo. O parágrafo é sabidamente uma unidade discursiva bem delimitada.

6. CONSIDERAÇÕES FINAIS

Neste relatório técnico, descreveu-se o processo de anotação de subtópicos dos textos-fonte do córpus CSTNews. A anotação foi realizada de forma manual contando com a participação de 13 anotadores. Essa iniciativa permitiu criar uma segmentação topical de referência que poderá ser utilizada em diversas aplicações.

Como trabalho futuro, pretende-se desenvolver algoritmos de segmentação topical baseados em diversas características textuais. O desempenho de tais algoritmos poderá ser avaliado em relação a segmentação topical de referência, que está disponível no CSTNews.

REFERÊNCIAS

Aleixo, P. e Pardo, T.A.S. (2008). *CSTNews: Um Córpus de Textos Jornalísticos Anotados segundo a Teoria Discursiva Multidocumento CST*. Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, N. 326. São Carlos-SP, 15p.

Biryukov, M.; Angheluta, R.; Moens, M-F. (2005). Multidocument question answering text summarization using topic signatures. *Journal on Digital Information Management*. Vol. 3, N. 1, pp. 27-33.

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. (2011). CSTNews - A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese. In: *Proceedings of the 3rd RST Brazilian Meeting*, pp. 88-105. October 26, Cuiabá/MT, Brazil.

Carletta, J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistic*, Vol. 22, N. 2, pp. 249-254.

Carlson, L. and Marcu, D. (2001). *Discourse Tagging Reference Manual*. Technical Report ISI-TR-545, University of Southern, California.

Hearst, M. (1997). TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages. *Computational Linguistics*, Vol. 23, N. 1, pp. 33-64.

Hennig, L. (2009). Topic-based multi-document summarization with probabilistic latent semantic analysis. In: *Recent Advances in Natural Language Processing* (RANLP), pp. 144-149. Borovets, Bulgaria.

Hovy, E. and Lin, C-Y. (1998). Automated Text Summarization and the SUMMARIST system. In: *Proceedings of TIPSTER'98*, pp. 197-214.

Jubran, C.C.A.S; Urbano, H; Fávero, L.L.; Koch, I.G.V. (1992). Organização tópica da conversação. In: ILARI, R. (org.). *Gramática do português falado*, Vol. II. Campinas/SP: UNICAMP, São Paulo: FAPESP, pp. 322-384.

Jubran, C.C.A.S. (2006). Revisitando a noção de tópico discursivo. *Cadernos de Estudos Linguísticos*, Vol. 48, N. 1, pp. 33-41.

Kazantseva, A. and Szpakowicz, S. (2012). Topical Segmentation: a study of human performance and a new measure of quality. In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 211-220.

Koch, I.G.V. (1998). *A coesão textual*. São Paulo: Contexto.

Koch, I.G.V. (2009). *Introdução à linguística textual*. São Paulo: Contexto.

Lin, C-Y. (1995). Knowledge-based automatic topic identification. In: *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics* (ACL-95), pp. 308-310. Cambridge, Massachusetts.

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

Oh, H-J.; Myaeng, S.H.; Jang, M-G. (2007). Semantic passage segmentation based on sentence topics for question answering. *Journal Information of Sciences*, Vol. 177, N.18, pp. 3696-3717.

Pinheiro, C.L. (2008). Organização tópica e sumarização do texto: estratégia para ensino de leitura. *Revista Horizontes de Linguística Aplicada*, Vol. 7, N. 1.

Prince, V. and Labadié, A. (2007). Text segmentation based on document understanding for information retrieval. In: *Proceedings of NLDB'07*, pp. 295-304.

Radev, D.R. (2000). A common theory of information fusion from multiple text sources, step one: Cross-document structure. In: *Proceedings of the 1st ACL SIGDIAL Workshop on Discourse and Dialogue*. Hong Kong, China.