Programa de Pós-Graduação em Linguística

PPGL

APLICAÇÃO DE CONHECIMENTO LÉXICO-CONCEITUAL NA SUMARIZAÇÃO AUTOMÁTICA MULTIDOCUMENTO

Rejeane Cassia de Luca

SÃO CARLOS

2019

Universidade Federal de São Carlos

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE EDUCAÇÃO E CIÊNCIAS HUMANAS
PROGRAMA DE PÓS-GRADUAÇÃO EM LINGUÍSTICA

APLICAÇÃO DE CONHECIMENTO LÉXICO-CONCEITUAL NA SUMARIZAÇÃO AUTOMÁTICA MULTIDOCUMENTO

REJEANE CASSIA DE LUCA

Dissertação apresentada ao Programa de Pós-Graduação em Linguística da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Linguística, área de concentração: Descrição, análise e processamento automático de línguas naturais.

Orientadora: Profa. Dra. Ariani Di Felippo

São Carlos - São Paulo - Brasil

2019

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Educação e Ciências Humanas
Programa de Pós-Graduação em Linguística

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Dissertação de Mestrado da candidata Rejeane Cassia de Luca, realizada em 28/02/2019:

Profa. Dra. Ariani Di Felippo
UFSCar

Profa. Dra. Cláudia Dias de Barros
IFSP

Profa. Dra. Débora Domiciano Garcia
Shawee

Certifico que a defesa realizou-se com a participação à distância do(s) membro(s) Débora Domiciano Garcia e, depois das arguições e deliberações realizadas, o(s) participante(s) à distância está(ao) de acordo com o conteúdo do parecer da banca examinadora redigido neste relatório de defesa.

Profa. Dra. Ariani Di Felippo

AGRADECIMENTOS

Ao Criador, primeiramente, por ter me permitido essa experiência humana.

À minha família que tanto me apoiou nessa trajetória e durante toda a minha vida. Em especial à minha irmã Ana Paula pelo amor e confiança de sempre, principalmente nas ocasiões mais difíceis em que o suporte emocional se fez necessário.

Aos meus amigos Theciana, Jackson e Roger pela inspiração, carinho, compreensão, colaboração, palavras de incentivo e ainda descontração em todos os momentos dessa jornada.

À minha orientadora Ariani Di Felippo pela paciência, apoio e conhecimento imensuráveis.

Aos pesquisadores do NILC Paula C.F. Cardoso e Marcio S. Dias pela colaboração e em especial a Fernando A.S. Nóbrega e Marco A. Sobrevilla Cabezudo que tanto se empenharam para a solução dos problemas apresentados pelas ferramentas computacionais.

À UFSCar e PPGL pela infraestrutura disponibilizada.

RESUMO

Na Sumarização Automática Multidocumento (SAM), produz-se automaticamente um único sumário (resumo) a partir de uma coleção de textos de diferentes fontes que versam sobre um mesmo tópico, com o objetivo de facilitar o acesso à informação. Tais sumários são comumente extratos informativos (isto é, sumários compostos por trechos inalterados dos textos-fonte que veiculam a ideia principal da coleção), o que requer a seleção das sentenças mais importantes. Para tanto, pode-se empregar conhecimento linguístico superficial (ou estatística), conhecimento profundo ou híbrido. Os métodos profundos, apesar de mais caros e menos robustos, produzem extratos mais informativos e com mais qualidade linguística. Tendo em vista os resultados promissores do uso de conhecimento profundo do tipo léxico-conceitual em pesquisas incipientes sobre a SAM ou na SAM multilíngue, investigaram-se 4 métodos distintos na SAM monolíngue para o português, os quais se baseiam primordialmente na frequência de ocorrência de conceitos lexicais na coleção para a seleção de conteúdo. Para tanto, selecionou-se o CSTNews, *corpus* multidocumento de referência para o português que possui todos os verbos e 10% dos nomes mais frequentes anotados com os conceitos (*synsets*) da WordNet de Princeton. Para a aplicação dos métodos léxico-conceituais, selecionaram-se 5 coleções do total de 50 do CTSNews, cuja anotação dos nomes foi completada. A partir das 5 coleções com nomes e verbos anotados em nível conceitual, testaram-se os 4 métodos: (i) LCFSummN, baseado na frequência de ocorrência dos nomes na coleção, (ii) LCFSummN-V, baseado na combinação da frequência dos nomes e verbos, (iii) LCFSummN-pond, baseado na média ponderada da frequência dos nomes e (iv) LCFSummN-V-pond, baseado na média ponderada da frequência dos nomes e verbos. Os extratos gerados foram avaliados intrinsecamente quanto à qualidade linguística e informatividade. Quando comparados a um método profundo do estado-da-arte em SAM monolíngue para o português, os resultados do trabalho evidenciam que os métodos léxico-conceituais apresentam bom desempenho.

ABSTRACT

Automatic Multi-document Summarization (MDS) aims at creating automatically a single summary from a collection of texts on the same topic in order to provide an alternative way to deal with the massive amount of information on the web. Since such summary is often an extract (i.e., a summary composed of unchanged excerpts extracted from the source texts that convey the main idea of the collection), it is required the selection of the most important sentences of the collection. For sentence selection, there are superficial (linguistic or statistical), deep linguistic, and hybrid methods. Despite being less robust and more expensive, the deep methods produce extracts that are not only more informative but also have higher linguistic quality. Considering the promising results of lexical-conceptual methods in incipient MDS or in multilingual MDS surveys, we investigated 4 methods in monolingual MDS for Portuguese, which is based on the frequency the lexical concepts in the cluster for content selection. We selected CSTNews, a reference multi-document corpus in Portuguese, whose verbs and 10% of the most frequent nouns are annotated with their correspondent synsets from Princeton WordNet. Specifically, we selected 5 clusters from the 50 in CSTNews, and extended the conceptual annotation to all nouns. Then, we applied 4 methods to the 5 clusters (i) LCFSummN, based on simple frequency of nominal concepts in the cluster, (ii) based on simple frequency of nominal and verbal concepts in the cluster, (iii) based on weighted-average for nominal concepts, and (iv) based on weighted-average frequency for nominal and verbal concepts. We intrinsically evaluated the extracts generated by each method regarding linguistic quality and informativeness. When compared to a deep state-of-art MDS method for Portuguese, the results of our investigation show the good performances of the lexical-conceptual methods.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

ÍNDICE

1. INTRODUÇÃO

1.1. Contextualização

De acordo com um relatório[1] produzido pelas agências de mídia social *We Are Social*[2] e *Hoostsuite*[3], a quantidade atual de usuários de internet no mundo é de aproximadamente 4.1 bilhões de pessoas. Isso equivale a 54% da população mundial, que se aproxima dos 7.6 bilhões de indivíduos. No Brasil, cuja população gira em torno de 210 milhões de pessoas, há aproximadamente 140 milhões de usuários de internet. Em se tratando de webpages, por exemplo, estima-se que há cerca de 60 bilhões de páginas indexadas pelos buscadores Google e Bing segundo o site *World Wide Web Size*[4].

Com tamanho volume de usuários, a quantidade de dados gerados e disponibilizados na web tem crescido exponencialmente. Para 2020, por exemplo, a previsão é de que haverá 40 *zettabytes* (ou 40 trilhões de *gigabytes*) de informação digital em circulação no mundo[5].

Nesse cenário, a busca por um assunto pode retornar inúmeras notícias de fontes jornalísticas distintas, o que dificulta o acesso à informação de interesse pelo usuário. Assim, o massivo volume de dados e a limitação cognitiva e de tempo dos usuários de internet tem impulsionado o desenvolvimento das tecnologias da informação e, por conseguinte, motivado o desenvolvimento de sistemas computacionais em algumas subáreas do Processamento Automático de Línguas Naturais (PLN).

Uma dessas subáreas é a Sumarização Automática Multidocumento (SAM), pois os sistemas de SAM, uma vez que produzem um único sumário (resumo) a partir de uma coleção de textos sobre mesmo assunto, figuram como ferramentas que permitem acesso mais ágil e eficaz à informação de interesse dos usuários (MANI, 2001).

No Quadro 2, apresenta-se um sumário[6] automático multidocumento gerado a partir das 2 notícias que compõem o Quadro 1, as quais versam sobre um mesmo assunto (isto é, a escolha da representante do Brasil no revezamento da tocha olímpica), sendo que cada uma dessas notícias é proveniente de uma fonte jornalística distinta[7].

---

[1] https://digitalreport.wearesocial.com/
[2] https://wearesocial.com/
[3] https://hootsuite.com/
[4] https://www.worldwidewebsize.com/
[5] https://br.okfn.org/2017/09/29/o-que-faremos-com-os-40-trilhoes-de-gigabytes-de-dados-disponiveis-em-2020/
[6] Sumário produzido pelo sumarizador CSTSumm (CASTRO JORGE, PARDO, 2010) (cf. Seção 2.3.2).
[7] As notícias e o sumário foram extraídos do CSTNews (CARDOSO *et al.*, 2011), *corpus* multidocumento de referência do português. Esse *corpus* será descrito em detalhes na seção 3.1 (pág. 38).

Quadro 1 – Notícias sobre um assunto veiculadas por fontes distintas.

Texto 1 (*Folha de S. Paulo*)

[1] A ginasta Jade Barbosa, que obteve três medalhas nos Jogos Pan-Americanos do Rio, em julho, venceu votação na internet e será a representante brasileira no revezamento da tocha olímpica para Pequim-2008.
[2] A tocha passará por vinte países, mas o Brasil não estará no percurso olímpico. [3] Por isso, Jade participará do evento em Buenos Aires, na Argentina, única cidade da América do Sul a receber o símbolo dos Jogos.
[4] O revezamento terminará em 8 de agosto, primeiro dia das Olimpíadas de Pequim.

Texto 2 (*O Estadão*)

[1] Um dos destaques desta temporada do esporte brasileiro, a ginasta Jade Barbosa foi escolhida, na noite desta terça-feira, para ser a representante do Brasil no revezamento da tocha dos Jogos Olímpicos de Pequim.
[2] Em votação pela internet, a ginasta recebeu mais de 100 mil votos e superou o nadador Thiago Pereira, que ganhou seis ouros nos Jogos Pan-Americanos.
[3] O Brasil não faz parte do trajeto da tocha olímpica. Na América do Sul, a chama passará por Buenos Aires, onde Jade participará do revezamento, no dia 11 de abril.
[4] Aos 16 anos, Jade conquistou três medalhas no Pan: ouro na disputa dos saltos, prata na apresentação por equipes e bronze no solo.
[5] Ao todo, a chama olímpica percorrerá 20 países antes de chegar a Pequim para a abertura da competição, no dia 8 de agosto.

Quadro 2 – Sumário multidocumento das notícias do Quadro 1.

Sumário automático multidocumento

A ginasta Jade Barbosa, que obteve três medalhas nos Jogos Pan-Americanos do Rio, em julho, venceu votação na internet e será a representante brasileira no revezamento da tocha olímpica para Pequim-2008.
A tocha passará por vinte países, mas o Brasil não estará no percurso olímpico.

Ao ser suficientemente informativo, um sumário como o do Quadro 2 pode substituir a leitura dos textos-fonte, o que facilitaria o acesso à informação veiculada pelas notícias do Quadro 2. Diante dessa potencialidade dos sumários multidocumento, tem-se focado no desenvolvimento de sumarizadores extrativos, os quais geram sumários como o do Quadro 1, isto é, sumários compostos pela justaposição das sentenças que veiculam a

informação mais relevante da coleção de notícias-fonte, extraídas na íntegra das mesmas. A produção dos extratos informativos tem sido o foco das pesquisas porque os sumários abstrativos, uma vez compostos por material linguístico não presente nos textos-fonte, só podem ser produzidos por técnicas avanças e, por isso, complexas e caras, de interpretação e geração de língua natural (NENKOVA, MCKEOWN 2011).

No cenário da SAM, processam-se tradicionalmente coleções de notícias, como ilustrado pelo sumário do Quadro 2. Mais recentemente, a sumarização de textos opinativos (ou simplesmente sumarização de opiniões) também tem sido foco de várias pesquisas devido à importância de se extrair conteúdo subjetivo para vários domínios da sociedade (p.ex.: LÓPEZ CONDORI, 2015; VARGAS, 2017). Embora a sumarização de opiniões possa ser vista como uma forma de SAM, essa tarefa possui peculiaridades que a diferenciam da aplicação tradicional.

A SAM tradicional envolve inúmeros desafios, os quais são causados pela origem variada das notícias-fonte. Além da seleção das informações mais relevantes, certos fenômenos multidocumento precisam ser tratados, como (i) informações redundantes, complementares e contraditórias, (ii) estilos de escrita diferentes, (iii) ordenação temporal dos eventos e (iv) diferentes perspectivas e focos. Além disso, a SAM também precisa lidar com as questões de coesão e coerência, típicas da clássica sumarização monodocumento (MANI, 2001; NENKOVA, MCKEOWN 2011).

Embora esses aspectos sejam relevantes para a produção automática de um sumário multidocumento, a seleção das sentenças mais relevantes da coleção é central à tarefa de SAM extrativa (NENKOVA; MCKEOWN 2011). O sumário do Quadro 2, por exemplo, é composto exclusivamente por sentenças do Texto 1 (Sentença [1] e [2]), pois foram essas as selecionadas como mais relevantes pelo método subjacente ao sumarizador CSTSumm. De modo geral, a seleção das sentenças para o sumário requer a pontuação e o ranqueamento das sentenças-fonte de acordo com um ou mais critérios de relevância para que, na sequência, as sentenças do topo do ranque sejam selecionadas para compor o sumário, desde que não haja redundância entre elas.

O principal critério de relevância aplicado na SAM é a redundância, posto que esse é o critério mais frequentemente empregado pelos humanos na mesma tarefa (CAMARGO, DI-FELIPPO, PARDO, 2015). Segundo esse critério, as sentenças que veiculam as informações que mais se repetem nos textos-fonte devem compor o extrato informativo, garantindo-se que não haja redundância entre elas.

Para a pontuação e ranqueamento das sentenças, há várias estratégias que buscam capturar a redundância, as quais caracterizam o sistema/método em si. De forma geral, há métodos de SAM para o português que são (i) superficiais, ou seja, que usam estatística ou pouco conhecimento linguístico para selecionar as sentenças, (ii) profundos, isto é, métodos que fazem uso massivo de conhecimento linguístico ou (iii) híbridos, ou seja, métodos que unem conhecimento linguístico e estatístico.

Apesar dos métodos profundos serem os mais caros e de aplicação mais restrita que os demais, pois dependem de recursos (p.ex.: gramáticas, léxicos e modelos de discurso) e ferramentas linguístico-computacionais auxiliares (p.ex.: *parser* discursivo, etiquetadores morfossintáticos), destaca-se que eles geram sumários mais coerentes, coesos e informativos.

Para o português, os métodos profundos pautam-se majoritariamente no uso de conhecimento semântico-discursivo, codificado nas relações da teoria/modelo CST (*Cross-document Structure Theory*) (RADEV, 2000), as quais permitem capturar os fenômenos multidocumento, sobretudo a redundância. Em Castro Jorge e Pardo (2010), por exemplo, os textos-fonte são modelados em um grafo, em que as sentenças são representadas por nós e as relações CST (estabelecidas entre sentenças de textos distintos) por arestas. As sentenças são pontuadas e ranqueadas com base no número de conexões no grafo, sendo que as sentenças com mais conexões ocupam o topo do ranque.

Embora tais métodos gerem extratos informativos com boa qualidade linguística e informatividade, os quais atingem e, por vezes, superam o estado-da-arte, há outro tipo de conhecimento profundo ainda pouco explorado na SAM tradicional, que é o conhecimento léxico-conceitual (ZACARIAS, 2016). Tal conhecimento fora aplicado de forma mais ampla na Sumarização Automática Multidocumento Multilíngue (SAMM) (TOSTA, 2014; DI-FELIPPO *et al.*, 2016).

De forma geral, tais trabalhos obtiveram resultados promissores. Isso se deve principalmente ao fato de que o critério de seleção de conteúdo, isto é, a frequência de ocorrência dos conceitos nominais na coleção, permite capturar a informação central (isto é, redundante) dos textos-fonte, posto que os conceitos lexicalizados por nomes são os mais frequentes e veiculam boa parte da carga semântica dos textos.

Diante do cenário exposto e visando contribuir para as pesquisas sobre SAM profunda do português, apresentam-se a seguir os objetivos (e hipóteses) desta pesquisa.

1.2. Objetivos e hipóteses

Neste trabalho, investigou-se o uso de conhecimento léxico-conceitual na SAM. Especificamente, testou-se a frequência de conceitos lexicalizados na coleção como critério para a seleção de conteúdo. Para tanto, considerou-se a frequência de conceitos nominais, cuja relevância já fora atestada, e também de conceitos verbais. A investigação da relevância da frequência dos conceitos verbais pautou-se na hipótese de que os verbos, sendo centrais à estrutura sentencial, veiculam informação relevante e, por isso, sua frequência na coleção contribui para a detecção da informação central a compor um extrato informativo.

1.3. Metodologia

O equacionamento metodológico desta pesquisa englobou a realização de 4 etapas:

Etapa 1 – Revisão da literatura: leitura constante de trabalhos sobre Sumarização Automática monolíngue (mono e multidocumento) e sobre SAM (mono e multilíngue) baseada em conhecimento profundo, sobretudo o léxico-conceitual.

Etapa 2 – Seleção e Anotação de *corpus*: (i) seleção do CSTNews (CARDOSO *et al.* 2011), *corpus* multidocumento de referência para o português que possui anotação parcial dos conceitos nominais (NÓBREGA, 2013) e anotação completa dos conceitos verbais (SOBREVILLA CABEZUDO, 2015) que constituem suas coleções, e (ii) extensão da anotação dos conceitos expressos pelos nomes.

Etapa 3 – Proposição e Aplicação de Métodos de SAM: proposição e aplicação manual de métodos extrativos baseados na frequência dos conceitos nominais e verbais. Esses métodos caracterizam-se pela segmentação dos textos-fonte anotados em sentenças e pela pontuação/ranqueamento das sentenças em função somente da frequência dos conceitos nominais na coleção ou em combinação com a frequência dos conceitos verbais.

Etapa 4 – Avaliação dos Métodos de SAM: consistiu na avaliação intrínseca dos métodos investigados. Os extratos foram avaliados quanto à sua qualidade linguística de forma manual, com base nos critérios clássicos da Sumarização Automática (SA) (DANG, 2008), e quanto à sua informatividade, por meio do pacote automático de medidas denominado ROUGE *(Recall-Oriented Understudy of Gisting Evaluation*) (LIN, 2004).

1.4. Estrutura da dissertação

O texto desta dissertação está organizado em 6 seções. Na Seção 2, apresenta-se a revisão da literatura. Na Seção 3, descrevem-se as tarefas de seleção e anotação conceitual do *corpus* CSTNews. Na Seção 4, apresentam-se os métodos de SAM baseados em conhecimento léxico-conceitual investigados neste trabalho. Na Seção 5, apresentam-se os procedimentos de avaliação dos métodos e seus resultados. Por fim, na seção 6, algumas considerações finais são feitas, apresentando as contribuições dadas por este trabalho, suas limitações e trabalhos futuros.

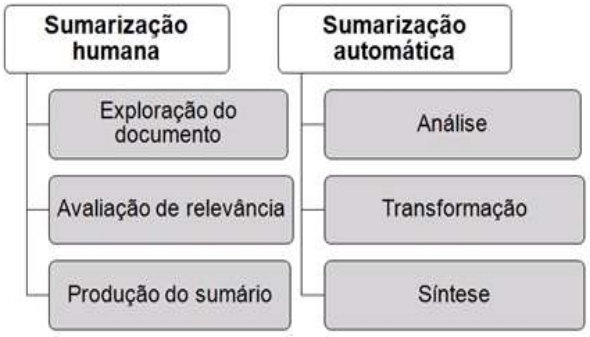
2. REVISÃO DA LITERATURA

Na Seção 2.1, apresentam-se os conceitos gerais sobre a SA, principalmente as etapas que constituem esse processo, os fatores que o influenciam e as estratégias de avaliação. Na Seção 2.2, apresenta-se uma revisão sobre a SA monodocumento, já que a SAM é vista como uma extensão da tarefa monodocumento. Na Seção 2.3, destacam-se os principais aspectos da SAM e os principais trabalhos para o português.

2.1. Noções fundamentais de Sumarização Automática

A SA é uma subárea do PLN na qual visa-se produzir de forma automática um sumário (ou resumo) a partir de um ou mais textos-fonte (MANI, 2001). Os sistemas computacionais que realizam a tarefa de sumarização são os chamados sumarizadores automáticos (SPARCK JONES, 2007). Tendo em vista que a sumarização feita por humanos segue três fases, isto é, (i) exploração do documento, (ii) avaliação de relevância e (iii) produção do sumário, a SA paralelamente envolve três etapas: (i) análise, (ii) transformação e (iii) síntese (SPARCK JONES, 1998; MANI, MAYBURY, 1999; SPARCK JONES, 1999; MANI, 2001). A Figura 1 apresenta a relação entre as etapas manuais e as fases da SA.

Figura 1 - Etapas de sumarização humana e automática



Fonte: Adaptada de Endres-Niggemeyer (1998) e Mani e Maybury (1999)

Na fase de **análise,** os sistemas de sumarização interpretam os textos-fonte e extraem uma representação formal do conteúdo dos mesmos para ser processada automaticamente.

A **transformação** é a etapa principal, pois, a partir da representação gerada na análise, o conteúdo dos textos-fonte é condensado em uma representação interna do sumário. Essa etapa engloba a seleção do conteúdo que irá compor o sumário e, para isso, é necessário ranquear os segmentos dos textos-fonte em função de sua relevância, a qual

pode ser determinada de acordo com diferentes critérios como a função e a audiência do sumário. A partir do ranqueamento, selecionam-se os segmentos de maior pontuação, ou seja, os que contêm as ideias centrais do texto, até que o tamanho desejado do sumário seja atingido (taxa de compressão).

A **síntese** consiste na construção do sumário propriamente dito, em língua natural, a partir da representação interna gerada na transformação. Para isso, métodos de justaposição, ordenação, fusão e correferenciação dos segmentos textuais selecionados podem ser utilizados (SPARCK JONES, 1993).

No geral, a SA sofre influência de determinados fatores, como: (i) taxa de compressão, (ii) audiência, (iii) função, (iv) forma, (v) gênero, (vi) número de textos-fonte, (vii) quantidade de línguas e outros (MCKEOWN, RADEV, 1995; MANI, 2001).

A **taxa de compressão** do sumário é definida por Mani (2001) como a razão entre o tamanho do texto-fonte e o tamanho desejado do sumário. Assim, uma taxa de compressão de 70% determina que o sumário seja composto por 30% do tamanho do texto-fonte. Essa taxa é medida, em geral, pelo número de palavras.

No que diz respeito à **audiência**, o sumário pode ser genérico ou focado nos interesses dos usuários. Os sumários genéricos veiculam as informações mais importantes dos textos-fonte sem que haja um perfil específico de usuário a ser atendido. Os sumários focados nos interesses dos usuários veiculam informações que atendem certo perfil de usuário ou certa consulta (*query,* em inglês) realizada por ele. Diante de um perfil de usuário leigo sobre o assunto coberto por um texto-fonte, por exemplo, a veiculação de informação contextual no sumário pode ser relevante.

Quanto à **função**, os sumários automáticos podem ser indicativos, informativos ou críticos. Os sumários indicativos, como os índices de livros, são comumente listas de palavras que apontam o conteúdo do texto-fonte, não substituindo, portanto, a leitura deste. Tais podem servir como ponto de partida para a seleção de uma leitura mais aprofundada sobre um tópico de interesse do usuário. Ao contrário dos indicativos, os sumários informativos, como os *abstracts* de artigos científicos, apresentam as informações mais relevantes do texto-fonte a ponto de sua leitura substituir a leitura da fonte. Os sumários críticos, por sua vez, apresentam as informações principais do texto-fonte e também uma avaliação sobre elas. As resenhas de livros e filmes são exemplos de sumários críticos.

Com relação à **forma**, os sumários automáticos podem ser extratos ou *abstracts*. Os extratos são geralmente compostos por sentenças extraídas na íntegra do texto-fonte,

não havendo, portanto, nenhum tipo de reescrita do material-fonte. Os *abstracts* são produzidos por meio da reescrita do conteúdo do texto-fonte, podendo, assim, conter palavras, sintagmas e até mesmo sentenças que não estavam presentes no *input*. A produção de *abstracts*, em especial, requer a aplicação de processos mais complexos de sumarização, como a fusão e a generalização de informação.

O **gênero** dos textos-fonte é outro fator que afeta diretamente a SA, pois é sabido que textos pertencentes a diferentes gêneros, como o científico e o jornalístico, apresentam características específicas que precisam ser consideras na fase de seleção de conteúdo. Assim, a depender do gênero, diferentes estratégias de seleção são necessárias.

Quanto ao **número** de textos-fonte, a SA pode ser monodocumento ou multidocumento. Na SA monodocumento, produz-se um sumário a partir de um único texto-fonte e, na SA multidocumento, o sumário é produzido a partir de uma coleção de textos, de fontes distintas, que abordam o mesmo assunto.

A **quantidade de línguas** é outro fator que influencia a produção automática de sumários. A modalidade de SA denominada monolígue se caracteriza pela geração de um sumário em uma língua *x* a partir de um ou mais textos na mesma língua *x*. A SA que envolve mais de uma língua, por sua vez, pode ser de 3 tipos: (i) *cross-language*, em que, a partir de um ou mais textos em uma língua *x*, produz-se um sumário em uma língua *y*, (ii) multilíngue, em que, a partir de uma coleção de textos em diferentes línguas, produz-se um sumário em uma das línguas dos textos de entrada, e (iii) independente de língua, em que se processam textos em quaisquer línguas para a geração de sumários também expressos em quaisquer línguas.

Outro fator que influencia a SA é a **quantidade e o tipo de conhecimento linguístico** envolvidos no processo. Segundo Mani (2001), quando a SA é realizada com base em pouco ou nenhum conhecimento linguístico, a abordagem é dita superficial. Nesse caso, a sumarização é pautada em informações linguísticas simples ou puramente em estatística. Embora os métodos superficiais comumente gerem sumários com menos coerência, coesão e informatividade, ressalta-se que estes apresentam baixo custo de desenvolvimento e maior aplicabilidade. Na abordagem dita profunda, a SA é pautada no uso massivo de conhecimento linguístico codificado em gramáticas, repositórios de sentidos e modelos de discurso. Nesse caso, os sumários produzidos são mais coerentes, coesos e informativos. Entretanto, os métodos profundos possuem aplicação mais restrita e são mais caros em relação aos métodos superficiais. Nesse contexto, os sumarizadores profundos podem gerar dois tipos de sumários: os extratos e os *abstracts* (que serão

apresentados mais à frente). As abordagens superficial e profunda podem ser unificadas, gerando um processo híbrido.

Para avaliar os métodos/sistemas desenvolvidos nessa tarefa, a comunidade do PLN realiza conferências internacionais periódicas, como a SUMMAC[8] (*Text Summarization Evaluation Conference*) e a TAC (*Text Analysis Conference*)[9], que se dedicam principalmente a estabelecer os parâmetros de avaliação.

De um modo geral, a avaliação de métodos/sistemas de SA pode ser realizada de forma intrínseca ou extrínseca. Na avaliação intrínseca, avalia-se o desempenho dos métodos/sistemas pela análise de seus resultados, ou seja, análise dos sumários. Na avaliação extrínseca, avalia-se a utilidade dos sumários em tarefas específicas, como a recuperação de informação (SPARCK JONES; GALLIERS, 1996). Diante do reconhecimento de que a avaliação extrínseca é demorada, cara e que requer planejamento cuidadoso (VAN-HALTEREN; TEUFEL, 2003), a abordagem mais frequentemente utilizada tem sido a intrínseca. Segundo Mani (2001), a avaliação intrínseca deve focar a qualidade linguística e a informatividade dos sumários.

a) Avaliação intrínseca da qualidade linguística

A qualidade dos sumários automáticos é realizada essencialmente por humanos, uma vez que visa analisar aspectos relacionados à gramaticalidade (ortografia e gramática) e à textualidade (coesão e coerência) (p.ex.: SAGGION; LAPALME, 2000; WHITE *et al.*, 2000), os quais dificilmente podem ser avaliados de forma automática.

Para avaliar a qualidade dos sumários na SAM, a TAC (DANG, 2008) estabeleceu 5 critérios linguísticos: gramaticalidade, não-redundância, clareza referencial, foco, estrutura e coerência.

De acordo com o critério da **gramaticalidade**, o sumário não deve apresentar erros de ortografia, pontuação e sintaxe, como o uso inadequado de letras maiúsculas ou minúsculas, problemas de formatação ou ainda a existência de erros que prejudiquem a legibilidade do texto (por ex.: sentenças agramaticais).

O atributo linguístico da **não-redundância** estabelece que o sumário não deve conter informações repetitivas, por exemplo, a repetição de fatos, substantivos, sintagma nominal ou até sentenças inteiras.

[8] http://www-nlpir.nist.gov/related_projects/tipster_summac/

[9] Em 2008, a *Text Analysis Conference* (TAC) passou a incorporar a DUC (*Document Understanding Conference*). O site http://duc.nist.gov/ engloba as informações referentes à DUC de 2001 a 2007.

Quanto à **clareza referencial**, o sumário deve apresentar de forma clara a identificação de uma pessoa ou entidade sobre a qual os pronomes e sintagmas nominais se referem.

O sumário deve conter um **foco** temático que seja identificável por meio de informações inter-relacionadas, ou seja, as sentenças devem conter informações que se relacionem com as informações do sumário como um todo.

Com relação ao atributo da **estrutura/coerência**, o sumário deve apresentar uma boa estrutura e uma organização adequada de forma que o encadeamento das sentenças construa uma estrutura informativa coerente sobre um mesmo tópico, ou seja, o sumário deve apresentar uma organização textual que preserve o sentido do texto.

b) Avaliação intrínseca da informatividade

A avaliação de informatividade é geralmente realizada de forma automática. A medida automática mais utilizada em conferências internacionais para essa avaliação é o pacote de medidas ROUGE.

As medidas ROUGE mensuram o nível de informatividade de um sumário por meio do cálculo da coocorrência de n-gramas, ou seja, a ROUGE realiza a comparação da quantidade de palavras em comum entre o sumário automático e um ou mais sumários humanos. Na medida ROUGE, o n-grama pode variar de 1 a 4 palavras. Dessa forma, dependendo da extensão do n-grama, a medida ROUGE apresenta especificações, como a ROUGE-1, que calcula a coocorrência de unigramas, e a ROUGE-2, que calcula a coocorrência de bigramas, e assim sucessivamente.

Nesse pacote de medidas, a avaliação é obtida pelos cálculos de precisão (P), cobertura (C) e medida-f (em inglês, *precision*, *recall* e *f-measure*, respectivamente) (LIN, 2004). A precisão corresponde à relação entre o número de n-gramas em comum do(s) sumário(s) de referência com os do sumário automático e o número total de n-gramas do sumário automático. Essa relação é representada pela fórmula:

$$\text{Precisão} = \frac{\text{n-gramas em comum entre sumário automático e humano}}{\text{n-gramas do sumário automático}}$$

A cobertura corresponde à relação entre o número de n-gramas em comum do(s) sumário(s) de referência com os do sumário automático e o número total de n-gramas do sumário de referência. Essa relação é representada pela fórmula:

$$\text{Cobertura} = \frac{\text{n-gramas em comum entre sumário automático e humano}}{\text{n-gramas do sumário humano}}$$

A medida-f, por sua vez, combina as métricas de cobertura e precisão, podendo ser representada pela seguinte fórmula:

$$\text{Medida-f} = \frac{2x \text{ Precisão x Cobertura}}{\text{Precisão + Cobertura}}$$

O resultado da medida-f varia entre 0 e 1. Assim, quanto mais próximo de 1 for o resultado, mais informativo é o sumário; quanto mais próximo de 0, mais baixo será o nível de informatividade.

Em síntese, a ROUGE é um conjunto de métricas automáticas aplicável a qualquer tipo de sumário e os resultados que se alcançam com sua utilização permitem a avaliação intrínseca da informatividade de um sumário. Ao comparar sumários automáticos a sumários de referência obtém-se a quantidade da informação dos textos-fonte presente no sumário automático.

Autores como Sparck Jones (1999), Saggion *et al.* (2002), Nenkova e Passonneau (2004) e Louis e Nenkova (2013) investigaram outras estratégias de avaliação por não haver um consenso a respeito da melhor forma de se avaliar métodos ou sistemas de SA.

Em Saggion *et al.* (2002), por exemplo, há três propostas de avaliação que medem a similaridade entre os sumários: (i) similaridade do cosseno, (ii) sobreposição de unidades lexicais (unigrama ou bigrama) e (iii) sobreposição da maior subsequência de unidades lexicais.

Nenkova e Passonneau (2004) propuseram método conhecido como método de pirâmide, o qual utiliza pouco ou nenhum envolvimento humano. Este modelo de avaliação considera um conjunto de sumários de referência a partir dos quais são extraídas "unidades de conteúdo do sumário" (*summarization contente units*- SCU). As SCU são organizadas em uma pirâmide cujo topo representa as que aparecem na maioria dos sumários de referência. As SCU são pontuadas de acordo com a posição na pirâmide, sendo que a SCU que aparecer em todos os sumários de referência sob avaliação recebe o maior peso e, portanto, estão localizadas mais próximas ao topo do que as demais. Os

sumários automáticos mais informativos são os que possuem maior número de SCU próximas ao topo da pirâmide, uma vez que estas são as unidades mais importantes.

Louis e Nenkova (2013) apresentam três métodos de avaliação de sumários, visando reduzir a influência da subjetividade humana na tarefa. O primeiro método mede a similaridade entre textos-fonte e sumários automáticos, ou seja, considera que, quanto maior a similaridade entre o sumário e seus textos-fonte, melhor é o seu conteúdo. No segundo, pseudomodelos (ou seja, sumários automáticos escolhidos por humanos) são adicionados a um conjunto de sumários de referência (humanos). Dessa forma, a avaliação final se dá pela comparação entre os sumários automáticos e o conjunto de referência expandido. No terceiro método, utilizam-se apenas sumários automáticos como modelo ou referência, os quais são determinados de forma semelhante ao método da pirâmide. Por meio de um cálculo probabilístico das palavras do conjunto de sumários automáticos (referência), obtém-se a distribuição global das palavras nesse conjunto, sendo que tal distribuição indica as informações mais importantes. Dessa forma, a avaliação final de um sumário automático é feita pela comparação de seu conjunto de palavras à distribuição global das palavras do conjunto de referência. Portanto, segundo os autores, bons sumários automáticos tendem a ter propriedades semelhantes à distribuição global.

Dando continuidade à revisão dos conceitos fundamentais da SA, discorre-se brevemente a seguir sobre a SA monodocumento, posto que a SAM é dita como uma extensão dessa tarefa.

2.2 A Sumarização Automática Monodocumento

A SA monodocumento monolíngue é uma modalidade tradicional e, por isso, já gerou muitos trabalhos desde seu surgimento em 1950 (GUPTA, LEHAL, 2010). O objetivo principal da SA monodocumento é a produção de extratos informativos e genéricos e, para isso, os sumarizadores automáticos comumente englobam as 3 etapas da Figura 1: análise, transformação e síntese.

Nos métodos superficiais, aplicam-se estratégias linguísticas simples, as quais guiam a seleção das sentenças de um texto-fonte para a geração de seu respectivo sumário extrativo (genérico e informativo).

Em um trabalho clássico da SA monodocumento, Baxendale (1958) propôs um método em que um sumário científico é produzido pela seleção das sentenças localizadas no início e final dos parágrafos do seu respectivo texto-fonte. Assim, o autor mostrou que

a seleção das sentenças mais relevantes para compor um sumário depende da sua posição no texto-fonte.

Em outro trabalho clássico, Luhn (1958) desenvolve um método em que as sentenças são pontuadas e ranqueadas com base nas palavras mais frequentes no texto-fonte, pois se pressupõe que tais palavras expressam o conteúdo central de um texto.

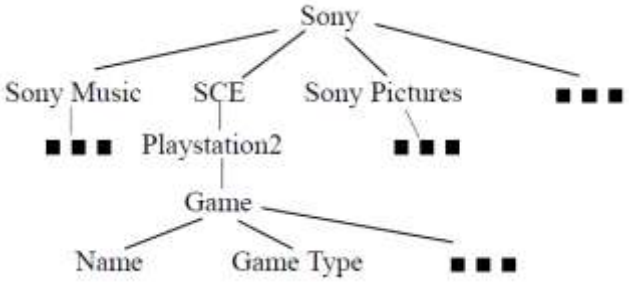
Outra estratégia é a seleção de informações que se relacionam às palavras contidas no título/subtítulo dos textos-fonte (EDMUNDSON, 1969). Essa estratégia consiste em identificar as palavras que compõem o título, subtítulo e tópicos para selecionar sentenças que contenham as ideias principais dos textos.

Expressões-chave ou indicativas de conteúdo que compõem a estrutura discursiva de certos gêneros também são empregadas na seleção de sentenças (PAICE, 1981). Por exemplo, as palavras "resumo", "introdução", "materiais", "métodos", "resultados", "discussão", "conclusão" e "referências" introduzem conteúdo específico em um texto científico e podem ser utilizadas como pistas para a seleção de sentenças, assim como expressões do tipo "este trabalho tem como objetivo".

Quanto aos métodos profundos, destacam-se os de Wu e Liu (2003) e Hennig *et al.* (2008), que se baseiam em conhecimento léxico-conceitual. No método de Wu e Liu (2003), faz-se a identificação dos principais tópicos e subtópicos pela comparação entre os termos que ocorrem no texto-fonte e os termos de uma ontologia[10]. Para tanto, os autores construíram um *corpus* e uma ontologia de domínio. O *corpus* é composto por 51 artigos, os quais foram compilados por meio da *query* (ou seja, termo de busca) SONY. Os conceitos que compõem a ontologia, construída de forma manual, foram organizados hierarquicamente, na forma de uma árvore. Trata-se de uma ontologia de domínio que armazena, por exemplo, (i) conceitos (p.ex.: *Sony*, *Sony Music* e *Sony Pictures*), e (ii) relações de subsunção (p.ex.: *Sony* subsume *Sony Music e Sony Pictures*). Por se tratar de uma árvore conceitual, os conceitos são os nós ou folhas e as relações são os galhos. A Figura 2 ilustra os conceitos mais genéricos da ontologia.

[10] No PLN, *ontologia* pode ser definida como um recurso ou base de conhecimento que fornece um inventário de conceitos, propriedades e relações entre conceitos que representam "uma interpretação da realidade" (isto é, o conhecimento de mundo compartilhado pelos membros de uma comunidade linguística) (GRUBER, 1995).

Figura 2 – *Top-level ontology* do domínio *Sony Corporation*



Fonte: Wu e Liu (2003).

Nesse método, os parágrafos são pontuados em função da informação topical que expressam. Para tanto, os termos neles contidos são comparados aos termos da ontologia. Caso o termo do parágrafo faça parte da ontologia, este é indexado à ontologia e o conceito é pontuado. Quando um conceito da hierarquia é pontuado, seus conceitos superiores também são pontuados. Com base nessa indexação e pontuação, o conceito mais genérico da ontologia (*Sony*) tem sempre a pontuação mais elevada, enquanto os conceitos do segundo nível têm pontuações diferentes. Com isso, apenas os conceitos mais bem pontuados do segundo nível representam os subtópicos e cada parágrafo é pontuado em função desses subtópicos. Os parágrafos são então selecionados até que a taxa de compressão fosse alcançada. Pode-se dizer que o método de Wu e Liu (2003) é uma versão mais refinada do método das palavras-chave, pois identifica o conteúdo central do texto pela frequência dos conceitos em uma ontologia.

No método de Hennig *et al.* (2008) para o inglês, a partir da indexação das palavras de conteúdo das sentenças de uma coleção de textos-fonte a uma hierarquia conceitual (codificada em grafo), utilizam-se informações estruturais da hierarquia para delimitar os conceitos mais relevantes e, por conseguinte, as sentenças que os veiculam.

Outra estratégia utilizada em métodos profundos consiste na modelagem discursiva do texto-fonte (p.ex.: O'DONNELL, 1997; MARCU, 2000). Tais trabalhos pautam-se na teoria/modelo *Rhetorical Structure Theory* (RST) (MANN; THOMPSON, 1987), que permite modelar um texto em uma árvore retórica, na qual as unidades de conteúdo (p.ex.: sentenças) são representadas por nós e as relações semântico-discursivas (p.ex.: *Circumstance, Background, Concession,* etc.) entre as unidades são representadas por arestas. No caso de um texto do gênero jornalístico, a primeira sentença tende a ser a mais nuclear na árvore RST que representa esse texto, sendo, portanto, selecionada para compor o sumário. Uma vez que a RST permite a codificação de conhecimento no nível

semântico-discursivo, a localização de uma unidade de conteúdo no topo de uma árvore RST é considerada um atributo sentencial profundo.

Para o português, há vários métodos/sistemas de SA monodocumento com base nas principais abordagens superficial e profunda da literatura. A maioria deles pode ser encontrada no site do NILC (Núcleo Interinstitucional de Linguística Computacional)[11]. Na sequência, discorre-se sobre a SAM, focalizando especialmente os métodos profundos desenvolvidos para o português.

2.3. A Sumarização Automática Multidocumento

De forma mais específica, define-se a SAM como uma modalidade da SA cujo objetivo é extrair o conteúdo mais importante de uma coleção de textos relacionados, removendo a redundância e levando em conta as semelhanças e diferenças no conteúdo da informação, e apresentá-lo ao usuário de forma condensada (sumário) (MANI, 2001).

Por tratar de múltiplos textos-fonte, a SAM, além de lidar com questões clássicas da SA monodocumento, como a coesão e a coerência, apresenta alguns fenômenos típicos como a ocorrência de informações complementares, contraditórias e redundantes. A redundância é o fenômeno mais comum na SAM, visto que a multiplicidade de textos-fonte, todos eles sobre um mesmo assunto, apresenta repetição de grande quantidade de informações.

Os estudos sobre SAM tiveram início a partir de 1990 com os trabalhos de McKeown e Radev (1995), Radev e McKeown (1998) e Carbonell e Godlstein (1998) e foram ganhando maior relevância a partir dos anos 2000. Nos últimos anos, a SAM tem desempenhado um papel de destaque no âmbito do PLN devido à crescente demanda por tecnologias/sistemas capazes de gerenciar o grande volume de informação disponível na internet, já que o processamento humano dessas informações em sua totalidade é muito custoso.

Na Figura 3, tem-se a arquitetura típica de um sumarizador multidocumento.

[11] http://conteudo.icmc.usp.br/pessoas/taspardo/sucinto/publications.html

Figura 3 – Arquitetura genérica de um sumarizador multidocumento monolíngue.

Fonte: Sparck Jones (1993).

2.3.1. A SAM superficial

Os métodos superficiais de SAM podem ser organizados em 3 grupos de acordo com o tipo de conhecimento linguístico que utilizam para a seleção de conteúdo (GUPTA, LEHAL, 2010; KUMAR, SALIM, 2012).

O primeiro grupo engloba aqueles que se baseiam em atributos linguísticos (do inglês, *feature-based methods*), os quais podem variar em número e combinação (p. ex.: LIN, HOVY, 2002; SCHILDER, KONDADADI, 2008) e apresentar pesos diferentes em função do tipo/gênero dos textos-fonte (cf. BOSSARD, RODRIGUES, 2011; SUANMALI *et al.*, 2011). Um atributo bastante comum é a frequência de ocorrência de palavras de classe aberta. Nesse caso, a etapa de análise é relativamente simples, consistindo na segmentação das sentenças e no cálculo da frequência de ocorrência de cada palavra nos textos-fonte da coleção. A etapa de transformação consiste em (i) pontuar e ranquear as sentenças em função da soma da frequência das palavras que as compõem e (ii) selecionar as mais bem pontuadas até que se atinja a taxa de compressão desejada, garantindo que não haja redundância entre elas. Na síntese, as sentenças selecionadas são justapostas na ordem em que aparecem nos textos-fonte.

Segundo essa abordagem, Pardo (2005) desenvolveu o GistSumm para o português, que pontua e ranqueia as sentenças dos textos da coleção com base no método *keywords* (palavras-chave) (ou as mais frequentes da coleção) ou TF-ISF[12] (*Term Frequency – Inverse Sentence Frequency*). Em ambos os métodos, a sentença de maior pontuação é considerada a *gist sentence*, ou seja, a sentença que expressa o conteúdo

[12] A medida TF-ISF é uma variação do método TF-IDF (*Text Frequency - Inverse Document Frequency*) (SALTON, 1988). A diferença é que a noção de documento é substituída pela noção de sentença. Assim, a importância de uma palavra w em uma sentença s, denotada por TF-ISF(w,s), é calculada pela fórmula TF-ISF(w,s) = TF(w,s)*ISF(w), em que TF(w,s) é o número de vezes que a palavra w ocorre na sentença s, e a frequência inversa da sentença é obtida pela fórmula ISF(w) = log(|S|/SF(w)), em que a frequência sentencial SF(w) é o número de sentenças nas quais a palavra w ocorre.

principal da coleção, a qual é automaticamente selecionada para iniciar o sumário. A seleção das demais sentenças para o sumário pauta-se em dois critérios: (i) a sentença deve conter ao menos um radical em comum com a *gist sentence* e (ii) sua pontuação deve ser maior do que a média da pontuação das demais sentenças. Quanto ao desempenho desse método, o GistSumm foi avaliado de forma intrínseca e extrínseca (PARDO *et al*., 2003). Na primeira, o sistema atingiu média de 51% quanto à informatividade e, na segunda, o sistema recebeu média 3.12 (em uma escala de 0 a 4), o que indica que os extratos têm boa utilidade.

O segundo grupo engloba os trabalhos baseados em *cluster* (grupo) e centroide (*cluster-based methods*) (p.ex.: RADEV *et al.*, 2004). Nesses métodos, a fase de análise consiste em agrupar as sentenças da coleção em *clusters* com base na similaridade lexical. Assim, os *clusters* são formados por sentenças semelhantes entre si, que veiculam os tópicos da coleção. Cada *cluster* é representado por um centroide, ou seja, um conjunto de palavras estatisticamente relevantes. De cada *cluster*, seleciona-se a sentença que contém o maior número de palavras em comum com o centroide.

O terceiro grupo de métodos superficiais engloba aqueles cuja etapa de análise consiste em modelar os textos-fonte como grafos (*graph-based methods*) (p.ex.: SALTON *et al.*, 1997; MIHALCEA, TARAU, 2005; WAN, 2008, NAYEEM, CHALI, 2017, YASUNAGA et al, 2017). Na transformação, as sentenças são representadas por nós e a similaridade entre elas é representada pelas arestas que conectam esses nós. Na fase de síntese, as sentenças mais fortemente conectadas a outras são extraídas para a construção do sumário.

Para o português, o método de Akabane *et al.* (2011), implementado no sistema RCSumm, caracteriza-se como "baseado em grafo". Nele, os textos-fonte são modelados em grafos e medidas de redes complexas são aplicadas para encontrar as sentenças mais relevantes da coleção de textos. Os sumários foram avaliados por meio da medida ROUGE e, como resultado, verificou-se que o grau[13] é a medida de redes complexas que gera os extratos com mais informatividade.

Em Nayeeme Chali (2017), por exemplo, a similaridade entre as sentenças é capturada pela medida do cosseno. Em Yasunaga *et al.* (2017), por sua vez, o grafo pode

[13] O grau de um nó *i* em uma rede complexa (k_i, onde, $0 \leq k_i \leq$ N-1 e N é o número total de nós da rede) representa a quantidade de arestas que ele possui, ou seja, a quantidade de nós diferentes conectados a *i*. Essa medida indica o quão um nó é conectado com seus vizinhos. Portanto, quanto maior o valor do grau, maior será a sua informatividade.

ser construído com base em diferentes tipos de relacionamentos entre as sentenças e não somente com base na similaridade lexical. Entre as informações utilizadas para relacionar as sentenças no grafo, estão, por exemplo, as (i) discursivas, codificadas pela ocorrência de marcadores discursivos e elementos correferenciais, e as (ii) estruturais, que se resumem à posição das sentenças nos seus respectivos textos-fonte e ao tamanho das sentenças.

2.3.2. A SAM profunda

Os métodos profundos também podem ser organizados em 3 grupos, de acordo com o tipo de conhecimento linguístico empregado (MANI, 2001).

No primeiro grupo, reúnem-se os métodos baseados em conhecimento sintático. Em Barzilay *et al.* (1999), por exemplo, a etapa de análise consiste em segmentar as sentenças e analisar sua estrutura sintática por meio de um *parser* (analisador sintático). Após a análise sintática, realiza-se o agrupamento das estruturas predicativas (predicado-argumento) similares, as quais teoricamente expressam o mesmo tópico, e selecionam-se as mais frequentes. Na etapa de síntese, as estruturas predicativas são reordenadas, gerando um sumário abstrativo (*abstract*). No método baseado em sintaxe de Schluter e Sogaard (2015), os textos-fonte são submetidos a um *parser* de dependência e as sentenças que expressam os conceitos mais relevantes da coleção são identificadas com base na frequência de dependências sintáticas não-rotuladas e rotuladas[14].

O segundo grupo inclui os métodos baseados em conhecimento semântico ou conceitual, como os de Mani e Bloedorn (1997), Li *et al.* (2010) e Schluter e Sogaard (2015). Mani e Bloedorn (1997), por exemplo, apresentam um método em que, na fase de análise, cada texto-fonte é modelado em um grafo. Nesse grafo, as palavras são representadas por nós e a similaridade distribucional entre elas por arestas. Dois nós com arestas similares representam palavras sinônimas e, portanto, expressam um conceito. As sentenças dos textos-fonte contendo palavras que representam os conceitos mais importantes da coleção são selecionadas para compor o sumário. Li *et al.* (2010), apresentam um método que consiste em indexar as sentenças de uma coleção aos conceitos de uma ontologia. Para compor o sumário, dada a *query* (consulta) de um usuário (também mapeada à ontologia), o sistema seleciona apenas as sentenças dos

[14]Por exemplo, dada uma sentença como *John walks*, um *parser* de dependência pode identificar duas estruturas, uma não-rotulada (p.ex.: DEPENDENCY (*walk, John*)) e outra rotulada, que, no caso, indica a relação de sujeito (p.ex.: SUBJECT (*walk, John*)).

textos-fonte indexadas aos mesmos conceitos a que os itens lexicais da *query* foram mapeados e/ou aos conceitos mais específicos, gerando sumários focados no interesse do usuário. Um dos métodos de SAM propostos por Schluter e Sogaard (2015) utiliza *frames* semânticos[15] (FILLMORE,1982; FILLMORE *et al.*, 2003) para identificar as sentenças mais relevantes dos textos-fonte. Os autores realizaram várias avaliações por meio da métrica ROUGE e concluíram que os *frames* representam de forma mais adequada os conceitos centrais da coleção de textos-fonte que os simples bigramas (método *baseline*).

Para o português, destacam-se os trabalhos de Tosta (2014) e Zacarias (2016).

Os dois métodos de SAMM descritos em Tosta (2014) e publicados mais recentemente em Di-Felippo *et al.* (2016) foram desenvolvidos especificamente para a SAMM. Neles, a sumarização parte de coleções bilíngues (1 notícia em português e 1 em inglês), cujos nomes foram indexados aos conceitos da WordNet de Princeton (WN.Pr), base léxico-conceitual desenvolvida para o inglês norte-americano (FELLBAUM, 1998)[16]. Uma vez que a indexação tenha sito feita, as sentenças são pontuadas e ranqueadas com base na frequência de ocorrência de seus conceitos constitutivos na coleção. A partir do ranque, o método CFUL (*concept frequency + user language*) seleciona apenas as sentenças em português com pontuação mais alta para compor o sumário até que a taxa de compressão desejada seja atingida. O método CF (*concept frequency*) seleciona as sentenças mais bem pontuadas independentemente da língua e, caso sentenças em inglês sejam selecionadas, faz a tradução destas para o português por meio de um tradutor automático. Segundo os autores, o método CFUL produz extratos mais informativos e com qualidade linguística superior.

Zacarias (2016) investigou métodos de SAM que partem da representação dos conceitos lexicais dos textos-fonte em uma hierarquia (ou árvore) para exploração de certas propriedades de grafo capazes de distinguir os conceitos mais relevantes. Para tanto, selecionaram-se 3 coleções do CSTNews e os nomes que ocorrem nos textos-fonte de cada coleção foram manualmente indexados aos conceitos da WN.Pr, gerando, ao final, uma hierarquia com os conceitos constitutivos da coleção e demais conceitos herdados da WN.Pr para a construção da hierarquia. Os conceitos da hierarquia foram

[15] Os *frames* semânticos são estruturas conceituais esquematizadas de forma a representar uma situação, objeto ou evento a partir de um *background* (pano de fundo) ou "cena" no qual está inserido. Assim, o sentido de uma palavra deve ser descrito em relação aos *frames* semânticos a ele relacionados. Os *frames* semânticos associados ao *frame* "comprar", por exemplo, envolvem necessariamente um "comprador" e um "produto/coisa" e opcionalmente um "vendedor" e um "valor".

[16] Mais detalhes sobre a WordNet de Princeton encontram-se na Seção 3.2.

caracterizados em função de 4 métricas de grafo pertinentes para a identificação dos conceitos relevantes a compor um sumário: *Simple Frequency*, *Cumulative Frequency*, *Closeness e Level.* A partir dessa caracterização, os autores propuseram 2 métodos de seleção de sentenças: (i) o CFSumm, cuja seleção de conteúdo se baseia exclusivamente na métrica *Simple Frequency*, e (ii) o LCHSumm, cuja seleção se baseia nas 4 medidas de relevância. Tais métodos foram avaliados intrinsecamente quanto à informatividade, por meio do pacote ROUGE, e à qualidade linguística, com base nos critérios da TAC. O desempenho dos métodos foi comparado ao do sumarizador GistSumm, tido como *baseline*. No geral, os métodos CFSumm e LCHSumm superam o *baseline*, sendo o CFSumm relativamente superior ao LCHSumm.

No terceiro grupo, apresentam-se os métodos baseados em conhecimento semântico-discursivo, os quais realizam uma modelagem dos textos-fonte de acordo com a teoria CST (do inglês, *Cross-document Structure Theory*) (RADEV, 2000). Tal modelagem é feita por meio de *parsers* ou analisadores discursivos, como o CSTParser (MAZIERO, 2011). A CST, baseada na RST (do inglês, *Rhetorical Structure Theory*) (MANN, THOMPSON, 1987), visa conectar sentenças (ou outras unidades textuais, como palavras, sintagmas, etc.), provenientes de documentos distintos, para estruturar o conteúdo dos mesmos (RADEV, 2000). Tais conexões são rotuladas por um conjunto de relações. Originalmente, essas relações foram agrupadas em um conjunto de 24 rótulos discursivos, conforme listadas no Quadro 3.

Quadro 3 – Conjunto original de relações da CST.

Identity	*Modality*	*Judgment*
Equivalence	*Attribution*	*Fulfillment*
Translation	*Summary*	*Description*
Subsumption	*Follow-up*	*Reader profile*
Contradiction	*Elaboration*	*Contrast*
Historical background	*Indirect speech*	*Parallel*
Cross-reference	*Refinement*	*Generalization*
Citation	*Agreement*	*Change of perspective*

Fonte: Radev (2000).

As conexões intertextuais permitem identificar fenômenos multidocumento, como similaridades e divergências de conteúdo, e até mesmo questões de estilo de escrita. Por

exemplo, uma relação de equivalência (*equivalence*, que corresponde à paráfrase) entre duas sentenças de textos distintos indica que as unidades conectadas contêm informação redundante, sendo, portanto, relevantes para o sumário. A relevância de uma sentença também pode ser determinada pelo número de relações CST que possui com as demais da coleção (MANI, 2001).

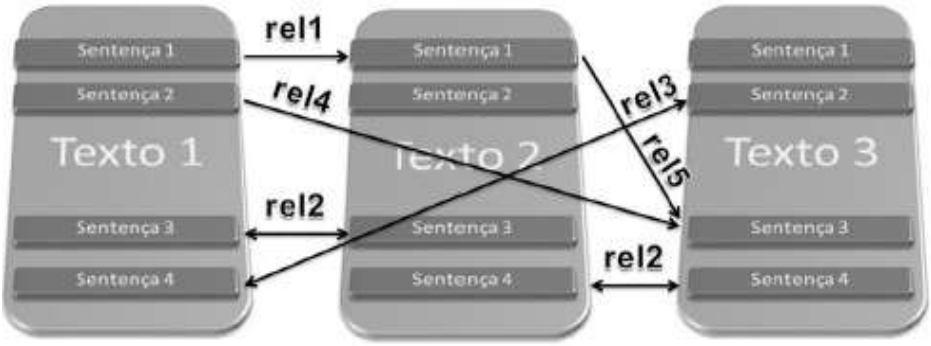
Para o português, destacam-se os trabalhos de Castro Jorge e Pardo (2010) e Cardoso (2014), nos quais o conhecimento semântico-discursivo é codificado pelas relações CST. Nesses trabalhos, tem-se utilizado o conjunto de 14 relações proposto por Aleixo e Pardo (2008) a partir da anotação manual do *corpus* CSTNews (CARSODO *et al.*, 2011a) (Quadro 4).

Quadro 4 – Conjunto de relações CST de Aleixo e Pardo (2008).

Identity	*Elaboration*
Equivalence	*Contradiction*
Summary	*Citation*
Subsumption	*Attribution*
Overlap	*Modality*
Historical background	*Indirect speech*
Follow-up	*Translation*

No CSTSumm de Castro Jorge e Pardo (2010), por exemplo, os textos-fonte são modelados em grafos durante a análise. Nesses grafos, os nós representam as sentenças e as arestas representam as relações CST (Figura 4).

Figura 4 – Esquema genérico de análise multidocumento.



Fonte: Maziero (2011).

Na transformação, as sentenças são pontuadas e ranqueadas com base no número de conexões no grafo, sendo que as sentenças com maior número de conexões recebem a pontuação mais alta. Sobre o ranque, são aplicados operadores de seleção de conteúdo que buscam refletir as preferências do usuário (p.ex.: exibir ou não informação contextual), promovendo uma reordenação das sentenças no ranque que privilegia a informação mais importante para o usuário. Com base no ranque final, selecionam-se as sentenças para composição do sumário. Conforme demonstrado pelos autores, o CSTSumm apresentou bons resultados nos critérios de avaliação, pois foi capaz de gerar sumários com boa informatividade e qualidade linguística.

Cardoso (2014), por sua vez, desenvolveu o método RC-4, publicado mais recentemente em Cardoso e Pardo (2015, 2016). O RC-4 baseia-se na combinação de conhecimento semântico-discursivo das teorias RST e CST. O RC-4 pontua e ranqueia as sentenças de cada coleção multidocumento com base em dois critérios de relevância: (i) a saliência da sentença em seu respectivo texto-fonte, codificada por meio das relações RST, e (ii) a correlação com os fenômenos multidocumento, indicada pelo modelo CST. Em outras palavras, esse método identifica as sentenças mais relevantes por meio da RST e, em seguida, as sentenças são selecionadas de acordo com o número de relações CST. Quanto ao desempenho, o método RC-4 foi avaliado por meio da medida ROUGE e comparado a outros sumarizadores multidocumento para o português. No geral, o RC-4 produz sumários melhores que os do CSTSumm, considerado o estado-da-arte da abordagem profunda, e similares aos obtidos pelo RSumm (RIBALDO, 2013; RIBALDO *et al.*, 2016), considerado o estado-da-arte da abordagem híbrida.

Quanto à abordagem híbrida, destaca-se o método de Schiffman *et al.* (2002), que se caracteriza por utilizar informações superficiais (localização da sentença nos textos-fonte e tamanho das sentenças) e conhecimento léxico-conceitual. Além dos atributos superficiais, os autores utilizam conhecimento de nível profundo ao determinar relações de sinonímia e hiponímia entre as palavras dos textos-fonte para delimitar os conceitos mais relevantes da coleção. As relações são identificadas pela indexação das palavras à WN.Pr.

Para o português, há vários métodos desenvolvidos segundo a abordagem híbrida, a saber: Castro Jorge e Pardo (2011), Castro Jorge *et al.* (2011), Ribaldo *et al.* (2012), Ribaldo *et al.* (2016), Camargo (2013) e Castro Jorge (2015).

O método/sistema de Castro Jorge e Pardo (2011) e Castro Jorge *et al.* (2011) combina atributos linguísticos superficiais (p.ex.: localização nos textos-fonte) e profundos (p.ex.: quantidade de relações CST).

Em Ribaldo *et al.* (2012), exploram-se medidas estatísticas aplicadas a grafos e redes, em combinação com relações CST. Especificamente, nesse trabalho, os textos-fonte são modelados em um grafo em que as sentenças são representadas como nós e as relações CST entre as sentenças como arestas. As arestas codificam o número de relações CST de uma sentença. Os nós mais altamente conectados são selecionados para o sumário, uma vez que estes representam as informações mais relevantes da coleção. Conforme descrito pelos autores, esse método apresentou bons resultados, aproximando-se dos melhores métodos desenvolvidos para o português.

Ribaldo (2013) e Ribaldo *et al.* (2016), por sua vez, exploraram medidas estatísticas aplicadas a grafos e redes em combinação com subtópicos (doravante, RSumm).

Já Camargo (2013) propôs um método de SAM híbrido cujos atributos linguísticos refletem estratégias de sumarização humana multidocumento (SHM). Tal método foi mais recentemente publicado em Camargo *et al.* (2015). Segundo os autores, os humanos selecionam as sentenças que satisfazem as seguintes características: (i) localização no início dos textos-fonte, (ii) redundância (alta) de seu conteúdo, (iii) ocorrência das palavras mais frequentes da coleção e (iv) tamanho médio ou pequeno (em número de palavras). Na avaliação, os extratos gerados pelas estratégias de SHM apresentaram qualidade superior aos gerados pelo melhor dos métodos RSumm que utilizam CST (RIBALDO *et al.*,2012).

Por fim, Castro Jorge (2015) desenvolveu o método MTRST-MCAD, modelado segundo o esquema *Noisy Channel* (em português, *Canal Ruidoso*) (SHANNON, 1948). Para o desenvolvimento desse método, assumiu-se que a fonte do canal (isto é, processo de sumarização) produz um sumário multidocumento que passa pelo canal ruidoso no qual algum tipo de ruído é introduzido, produzindo um conjunto maior de textos. No geral, a SAM modelada via *Noisy Channel* englobou três componentes probabilísticos: (i) P(S) é a probabilidade do sumário e descreve o modelo que captura padrões da boa construção de um sumário multidocumento em termos de coerência; (ii) P(C|S) é o canal ruidoso (ou a etapa de transformação na Figura 3) e modela a seleção de conteúdo via diversos atributos que representam os fatores que influenciam a sumarização, e (iii) P(S|C) é a etapa de decodificação, sendo responsável pela busca do melhor sumário de

acordo com os modelos P(C|S) e P(S), os quais são inferidos a partir de um *corpus* de textos-fonte e seus correspondentes extratos humanos multidocumento. No MTRST-MCAD, o modelo de transformação (MT) engloba atributos baseados na representação dos textos-fonte via RST. O modelo de coerência, por sua vez, foi desenvolvido com base no modelo de entidades e com informações discursivas via CST. Aliás, "MCAD" indica que o modelo de coerência foi aplicado após o processo de decodificação. Quanto ao desempenho, o MTRST-MCAD foi comparado aos principais métodos/sistemas multidocumento para o português via ROUGE e os resultados evidenciam que seus extratos possuem informatividade compatível e, às vezes superior, ao estado da arte.

No Quadro 5, apresenta-se uma síntese dos trabalhos sobre SAM descritos nesta Seção, destacando a autoria, o ano de publicação, a abordagem de SA, a língua alvo e o tipo de conhecimento linguístico predominante.

Quadro 5 – Síntese dos trabalhos sobre SAM da literatura

Autor	**Ano**	**Abordagem**	**Língua**	**Método/Informação**
Mani e Bloedorn	1997	Profunda	Inglês	Conceitual: Similaridade de conceitos
Barzilay *et al.*	1999	Profunda	Inglês	Sintaxe: estruturas predicativas (predicado-argumento)
Radev	2000	Profunda	Inglês	Semântico-discursivo: CST
Pardo *et al.*	2003	Superficial	Português	Frequência de ocorrência de palavras
Radev *et al.*	2004	Superficial	Inglês	*Cluster* e centroide
Pardo	2005	Superficial	Português	Frequência de ocorrência de palavras
Li *et al.*	2010	Profunda	Inglês	Conceitual: indexação de itens lexicais à conceitos de uma ontologia
Maziero *et al.*	2010	Profunda	Português	Semântico-discursivo: CST
Castro Jorge e Pardo	2010	Profunda	Português	Grafo, CST
Akabane *et al.*	2011	Superficial	Português	Grafo: redes complexas
Castro Jorge e Pardo, e, Castro Jorge *et al.*	2011	Híbrida	Português	Localização de sentenças e relações CST
Ribaldo *et al.*	2012	Híbrida	Português	Grafo, CST
Camargo	2013	Híbrida	Português	Combinação de atributos linguísticos e estratégias de Sumarização humana (localização de sentenças, redundância, frequência de palavras e tamanho de sentença).
Cardoso	2014	Profunda	Português	CST e RST
Schluter e Sogaard	2015	Profunda	Inglês	Sintaxe: dependências sintáticas não-rotuladas e rotuladas
Schluter e Sogaard	2015	Profunda	Inglês	Semântico: *frames* semânticos
Castro Jorge	2015	Híbrida	Português	*Noisy Channel* e CST
Ribaldo *et al.*	2016	Híbrida	Português	Medidas de grafos e redes em combinação com subtópicos

Quadro 5 – Síntese dos trabalhos sobre SAM da literatura – "continuação"

Di-Felippo *et al.*	2016	Profunda	Português/Inglês	Conceitual: frequência de conceitos
Zacarias	2016	Profunda	Português	Conceitual: hierarquia conceitual
Nayeeme Chali	2017	Superficial	Inglês	Modelagem em grafo, similaridade do cosseno
Yasunaga *et al.*	2017	Superficial	Inglês	Modelagem em grafo, marcadores discursivos (elementos correferenciais) e estruturais (posição das sentenças)

Diante dos trabalhos sobre SAM descritos nesta Seção, constata-se que, os métodos profundos que utilizam conhecimento léxico-conceitual para o Português ainda são incipientes e necessitam de aplicação mais ampla. Nesse cenário, este trabalho propõe métodos baseados na frequência de conceitos visando contribuir para os avanços na área de SAM monolíngue. Para tanto, apresentam-se nas próximas seções a seleção do *corpus* e proposição dos métodos.

3. SELEÇÃO E ANOTAÇÃO DE *CORPUS*

Por definição, um *corpus* é um conjunto de dados linguísticos sistematizados de acordo com determinados critérios, de forma que possa ser processado por um computador com a finalidade de proporcionar resultados variados e que sejam úteis para a descrição e análise (SINCLAIR, 2005).

Dessa forma, com base no objetivo desta pesquisa, fez-se necessária a seleção de um *corpus* com características específicas. Idealmente, o *corpus* a ser escolhido deveria ser: (i) monolíngue, especificamente do português, já que este trabalho focaliza a investigação de métodos de SAM no cenário específico do processamento automático do português; (ii) multidocumento, pois um *corpus* desse tipo fornece *clusters* de textos provenientes de diferentes fontes que versam sobre um mesmo assunto, (iii) jornalístico, em função da tradição dos trabalhos em SAM, e (iv) anotado, no caso, com informações de nível léxico-conceitual.

A partir desses requisitos, identificou-se o CSTNews (CARDOSO *et al.* 2011a), *corpus* de referência para o português, cujas características estão descritas a seguir.

3.1. O *corpus* CSTNews

O CSTNews possui 50 *clusters* (ou coleções) compostos por 2 ou 3 textos em português provenientes de diferentes fontes jornalísticas. As coleções abordam domínios diversos e os textos no interior de cada uma delas versam necessariamente sobre o mesmo tópico.

Os textos foram compilados dos principais jornais *online* do Brasil, a saber: *Folha de São Paulo, Estadão, O Globo, Jornal do Brasil* e *Gazeta do Povo*. Essas fontes foram escolhidas devido à popularidade e circulação na web, garantindo a coleta de uma mesma notícia veiculada por fontes distintas. A coleta foi realizada de forma manual durante aproximadamente 60 dias, no período de agosto a setembro de 2007.

No total, o CSTNews possui 140 textos, que somam 2.088 sentenças e 47.240 palavras. Cada coleção do *corpus* possui ainda sumários humanos e automáticos, monodocumento e multidocumento, e uma série de anotações que explicitam conhecimento linguístico de diferentes níveis. Os *clusters* estão organizados em categorias, cujos rótulos indicam a seção do jornal da qual os textos que os constituem foram coletados. Assim, o *corpus* é composto por coleções de várias categorias distribuídas conforme Quadro 6.

Quadro 6 – Relação de *clusters* por categoria (CSTNews)

Categoria	**Nº de *clusters***
Mundo	14
Política	10
Cotidiano	14
Ciência	1
Dinheiro	1
Esporte	10

Especificamente, cada *cluster* do CSTNews é composto por:

a) 2 ou 3 textos-fonte,
b) sumário manual de cada texto-fonte;
c) 6 *abstracts* e 6 extratos multidocumento manuais;
d) sumários automáticos multidocumento;
e) interconexão entre os textos-fonte via CST[17];
f) anotação de expressões temporais nos textos-fonte;
g) etiquetação morfossintática e sintática dos textos-fonte;
h) anotação dos sentidos dos nomes e verbos por meio dos conceitos da WN.Pr;
i) anotação de aspectos informativos (p.ex.: o quê, onde, etc.) de um dos sumários multidocumento de referência (manuais);
j) anotação discursiva de cada texto-fonte via RST e
k) anotação de subtópicos informativos dos textos-fonte.

Todos os sumários multidocumento do CSTNews (manuais e automáticos) possuem taxa de compressão de 70%, ou seja, apresentam 30% do número de palavras do maior texto-fonte de sua coleção correspondente.

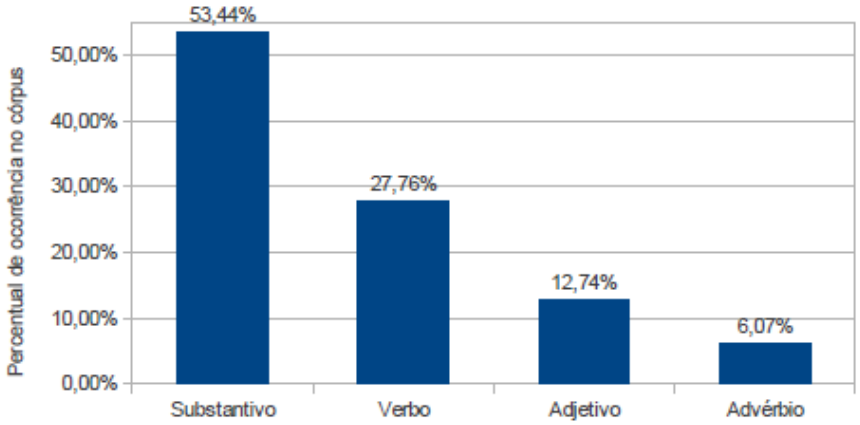
3.2. A anotação conceitual prévia do CSTNews

A anotação léxico-conceitual do CSTNews, em especial, foi inicialmente motivada pelo desenvolvimento de métodos de desambiguação lexical de sentidos para os nomes (NÓBREGA, 2013; NÓBREGA, PARDO, 2014) e posteriormente para os verbos (SOBREVILLA CABEZUDO, 2015; SOBREVILLA CABEZUDO *et al.*, 2017).

[17] O *corpus* CSTNews é assim denominado porque os textos das coleções estão alinhados por meio das relações estabelecidas pela teoria/modelo linguístico-computacional CST.

Para tanto, Nóbrega (2013) realizou um pré-processamento do CSTNews, que consistiu na etiquetação morfossintática[18] realizada por meio do *tagger* MXPOST (RATNAPARKHI, 1996), cujo objetivo era o de identificar o total de palavras pertencentes às classes abertas que compõem o *corpus* (Figura 5).

Figura 5 – Frequência de ocorrência por categoria gramatical no *corpus* CSTNews.



Fonte: Nóbrega (2013).

Tendo em vista a alta quantidade de unidades lexicais da classe dos nomes, a anotação destes restringiu-se aos mais frequentes de cada coleção, limitando-se a 10% do total de nomes presentes em todos os textos de cada coleção. Para a investigação da desambiguação de sentido dos verbos, Sobrevilla Cabezudo (2015) realizou a anotação total das unidades lexicais dessa classe no CSTNews.

A anotação conceitual dos nomes e verbos do CSTNews caracteriza-se pela utilização da WN.Pr (FELLBAUM, 1998) como repositório de conceitos, o que, aliás, determinou certas etapas da anotação (cf. Figura 6).

Os repositórios são recursos computacionalmente tratáveis que contêm pares de palavras e seus respectivos conceitos ou sentidos, como dicionários, ontologias[19], *thesaurus*, etc. Após investigar os repositórios do português, como o TEP 2.0 (MAZIERO *et al.*), a Onto.PT (OLIVEIRA, GOMES, 2012) e a WordNet.BR (DIAS-DA-SILVA, 2005), Nóbrega (2013) optou pela WN.Pr e essa escolha também foi seguida posteriormente por Sobrevilla Cabezudo (2015).

[18] A etiquetação morfossintática automática ou *tagging* é realizada por uma ferramenta comumente denominada *tagger* e consiste na atribuição da categoria gramatical (nome, verbo, adjetivo, etc.) a cada palavra de um texto.

[19] Por "ontologia", entende-se um inventário de conceitos, propriedades e relações entre conceitos que representam "uma interpretação da realidade", ou seja, o conhecimento de mundo compartilhado pelos membros de uma comunidade linguística (GRUBER, 1995).

A escolha pela WN.Pr se deveu a: (i) adequação linguística, já que segue princípios claros sobre a organização do conhecimento lexical advindos da Psicolinguística e Ciência Cognitiva, (ii) pertinência computacional e acessibilidade, já que possui um formato tratável por máquina e é de livre acesso na *web*, e (iii) extensão, pois é uma das bases lexicais mais bem estruturadas do inglês, armazenando, em sua versão 3.0, um total de 155.287 unidades lexicais, distribuídas em 117.798 nomes, 11.529 verbos, 21.479 adjetivos e 4.481 advérbios.

Na WN.Pr, as palavras ou expressões do inglês estão organizadas em quatro classes: nome, verbo, adjetivo e advérbio. No interior de cada classe, as palavras e as expressões estão codificadas na forma de *synsets* (do inglês, *synonym sets*), ou seja, conjuntos de formas sinônimas ou quase-sinônimas (p.ex.: {car, auto, automobile, machine, motorcar}). Cada *synset* representa um único conceito lexicalizado[20] por suas formas constituintes[21]. Entre os *synsets*, codificam-se 5 relações: antonímia, hiponímia, meronímia, acarretamento e causa (LYONS, 1979; CRUSE, 1986; FELLBAUM, 1998).

A anotação dos nomes e verbos realizadas por Nóbrega (2013) e Sobrevilla Cabezudo (2015) seguiu uma metodologia bastante similar e consistiu em explicitar o conceito subjacente a nomes e verbos por meio da associação de um *synset* da WN.Pr a cada ocorrência dessas unidades lexicais no *corpus*. Devido à utilização da WN.Pr, aliás, a anotação do CSTNews é tida como "léxico-conceitual".

Quanto à equipe, ambas anotações léxico-conceitual foram realizadas por um grupo misto de anotadores (composto por linguistas e cientistas da computação) que se reuniu em sessões diárias de 1 hora por um período de 6/7 semanas. As coleções de textos foram distribuídas entre grupos de 2 ou 3 linguistas e cientistas da computação e cada coleção foi anotada uma única vez por cada grupo. No total foram 10 anotadores distribuídos em grupos.

Os anotadores utilizaram ferramentas de suporte ou editores que agilizaram o processo de anotação. No caso dos nomes, Nóbrega (2013) desenvolveu a ferramenta denominada NASP (NÓBREGA, 2013), a qual foi estendida por Sobrevilla Cabezudo (2015) para a anotação dos verbos, originando o NASP++. Os editores possuem

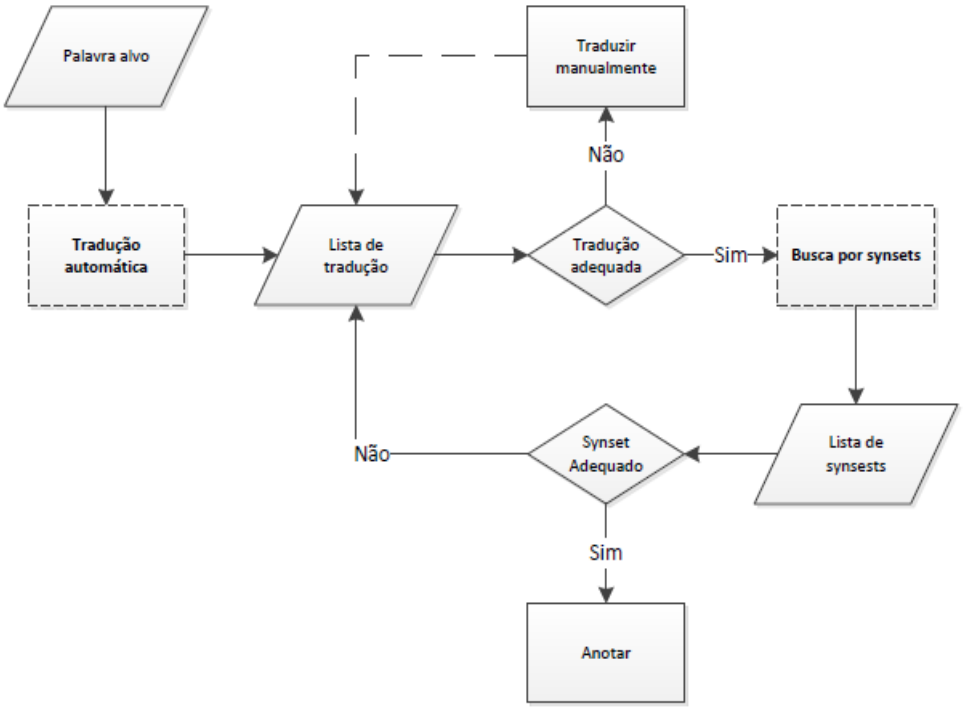
[20] A lexicalização é o processo pelo qual um conteúdo semântico é expresso por uma unidade lexical, seja ela simples, como "casa", composta, como "guarda-roupa", ou mesmo complexa, como "nota fiscal". (TAYLOR, 1985; LAKOFF, 1987).

[21] Os *synsets* também podem codificar conceitos não-lexicalizados (ou seja, conceitos para os quais não há uma expressão lexical, como é o caso dos conceitos expressos pelos *synsets* {natural object} e {external bodyparts}. A principal razão da inclusão desses conceitos é auxiliar a organização da hierarquia conceitual (VOSSEN, 1998).

funcionalidades muito semelhantes, sendo que o NASP++ apresenta duas extensões, que são: (i) a não-restrição da anotação às unidades lexicais mais frequentes, e (ii) a geração automática de uma representação hierárquica dos conceitos anotados na coleção, que é obtida a partir da identificação da relação de hiponímia entre esses conceitos na WN.Pr.

Tais ferramentas foram construídas segundo o princípio heurístico de que uma palavra tende a assumir um mesmo significado ou conceito em um mesmo contexto (MIHALCEA, 2006). Portanto, diante da anotação de uma palavra *x* no texto-fonte inicial com um *synset*, os editores automaticamente pré-anotam as demais ocorrências de *x* com o *synset* nos demais textos-fonte da mesma coleção. Com isso, a anotação se deu com base nas seguintes diretrizes gerais: (i) iniciar a anotação pelo texto de menor tamanho, pois, assim, garantia-se que os demais textos-fonte teriam várias palavras pré-anotadas, agilizando o processo geral; (ii) anotar todas as ocorrências das palavras (nomes ou verbos) na coleção; (iii) anotar outro texto da coleção somente após ter anotado todas as ocorrências das palavras em questão no texto inicial, e assim sucessivamente; (iv) avaliar as pré-anotações para confirmar, alterar ou excluí-las; (v) revisar e salvar a anotação. Além disso, as ferramentas de edição codificam uma sequência específica de passos, a qual está ilustrada na Figura 6.

Figura 6 – Metodologia de anotação léxico-conceitual do *corpus* CSTNews.



Fonte: Nóbrega (2013).

Com base na Figura, vê-se que cada palavra *x* etiquetada como nome, por exemplo, é inicialmente traduzida para o inglês com base no dicionário WordReference®[22]. Essa etapa de tradução fez-se necessária para recuperar os conceitos da WN.Pr, os quais estão representados por conjuntos de formas sinônimas em inglês. Quando o dicionário não fornece uma possibilidade de tradução ou fornece possibilidades equivocadas, o anotador humano pode inserir a opção mais adequada manualmente. Após selecionar a tradução mais adequada *y*, o editor recupera da WN.Pr todos os *synsets* que possuem *y* como um de seus elementos constitutivos e os exibe ao anotador. Após analisar todos os *synsets*, o anotador escolhe o que representa mais adequadamente o conceito ou sentido subjacente à *x* no texto-fonte.

Se para o desenvolvimento de métodos de desambiguação de sentidos, a anotação dos conceitos subjacentes aos nomes que compõem a parcela dos 10% mais frequentes de cada coleção era suficiente, o mesmo não ocorre para a investigação de métodos de SAM. Tendo em vista o objetivo da SAM de identificar as sentenças que veiculam as informações mais importantes da coleção, todos os conceitos nominais que ocorrem na coleção precisam ser considerados no processo de ranqueamento das sentenças. Por conseguinte, fez-se necessária a extensão da anotação léxico-conceitual dos nomes, a qual é detalhada a seguir.

3.3. A extensão da anotação conceitual dos nomes do CSTNews

Essa etapa consistiu em completar a anotação dos conceitos subjacentes aos nomes das coleções do CSTNews. Como o processo de anotação é complexo e demorado, optou-se pelo emprego do NASP++, posto que essa ferramenta não apresenta a restrição inicial do NASP. No entanto, tal editor, desenvolvido por Sobrevilla Cabezudo em 2015, não reconhecia ou não carregava adequadamente (isto é, não exibia corretamente a anotação conceitual prévia) certas coleções da nova versão (6.0) do CSTNews.

Após um longo trabalho colaborativo com o desenvolvedor[23] do editor, em que o arquivo de cada texto-fonte do *corpus* foi analisado de forma individual, a ferramenta passou a reconhecer e carregar todas as coleções do CSTNews. Se, por um lado, a ferramenta era fundamental para agilizar a anotação léxico-conceitual, a qual seria quase

[22] http://www.wordreference.com/

[23] A Marco Antonio Sobrevilla Cabezudo, doutorando do Programa de Pós-graduação em Ciências da Computação da USP (São Carlos) e pesquisador do NILC, sinceros agradecimentos pelo esforço e tempo dispendidos para gerar uma versão estável do editor NASP++ que pudesse ser utilizada neste trabalho.

que inviável se fosse realizada manualmente, o problema inesperado do não-reconhecimento dos arquivos acabou por atrasar o pré-processamento do *corpus.*

A solução demorada do problema do NASP++ afetou o tempo originalmente destinado à anotação do *corpus*. Por conseguinte, restringiram-se a quantidade de coleções e o número de anotadores (humanos). Especificamente, a anotação dos conceitos nominais de apenas 5 coleções do *corpus* foi completada (cf. Quadro 7). Diante de um *corpus* composto por 5 coleções, ressalta-se que os resultados obtidos neste trabalho precisarão ser futuramente validados em um *corpus* maior. Quanto à equipe de anotação, ressalta-se que a tarefa foi realizada apenas pela autora deste trabalho, posto que não houve tempo hábil para treinar outros anotadores.

Quadro 7 – Relação de *clusters* cujos conceitos nominais foram 100% anotados.

Cluster	**Domínio**	**Assunto/Tópico**
C1	Mundo	Queda de um avião na República Democrática do Congo
C2	Política	Reeleição de Lula
C3	Cotidiano	Acidente aéreo com avião da TAM
C4	Cotidiano	Alagamentos na cidade de São Paulo
C5	Cotidiano	Indicação de nome para a presidência da ANAC

3.3.1. O editor NASP++

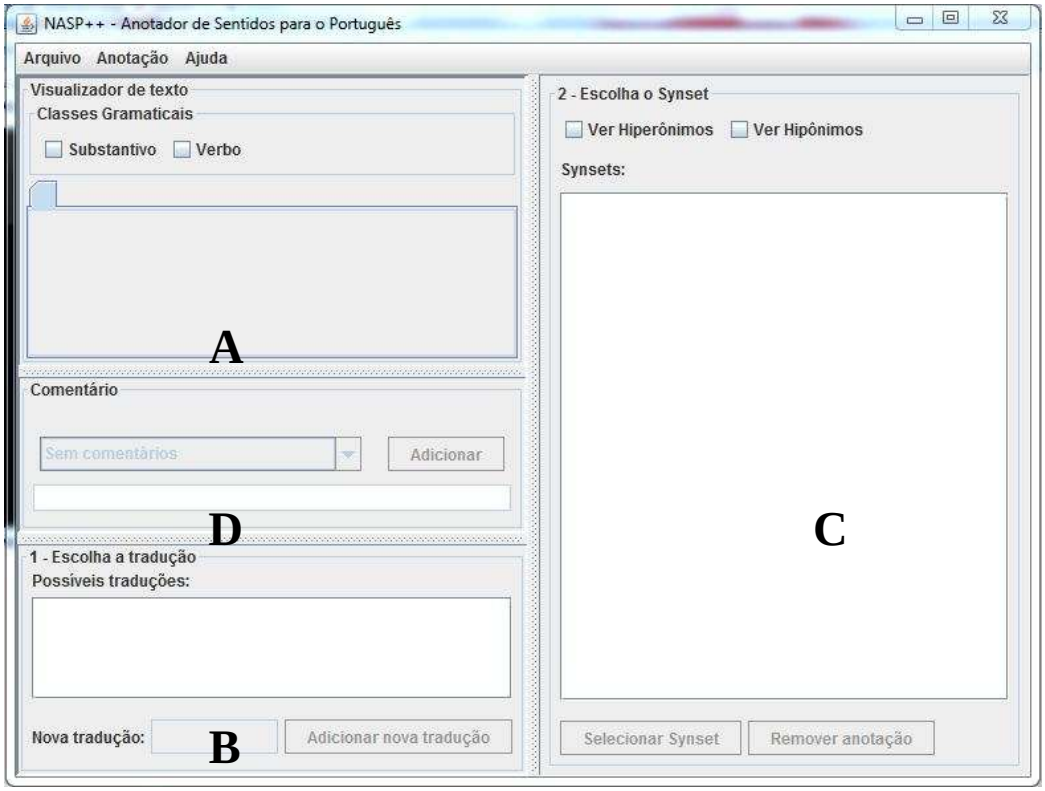
Para a anotação/indexação dos nomes aos conceitos da WN.Pr, o editor realiza o pré-processamento dos textos-fonte da coleção, que engloba os processos de (i) tokenização (isto é, processo de delimitação de *tokens*, que geralmente são sequências de caracteres separados por espaços em branco), (ii) etiquetação morfossintática[24] (isto é, atribuição da categoria gramatical a cada palavra) e (iii) lematização[25] (isto é, processo de redução das palavras à sua forma canônica ou básica). Somente após as tarefas (i)-(iii), os textos-fonte são exibidos para anotação.

A anotação é feita por meio dos seguintes passos: (i) seleção do nome a ser anotado, (ii) a tradução desse nome para o inglês, o que é feito pelo acesso ao tradutor automático WordReference, (iii) recuperação dos *synsets* dos quais esse nome é elemento constitutivo, (iv) seleção do *synset* que mais adequadamente representa o conceito expresso pelo nome e (v) anotação do nome com o *synset* escolhido em (iv).

[24] O NASP++ utiliza o etiquetador MXPost (RATNAPARKHI, 1996).
[25] Disponível em http://www.icmc.usp.br/pessoas/taspardo/LematizadorV2a.rar

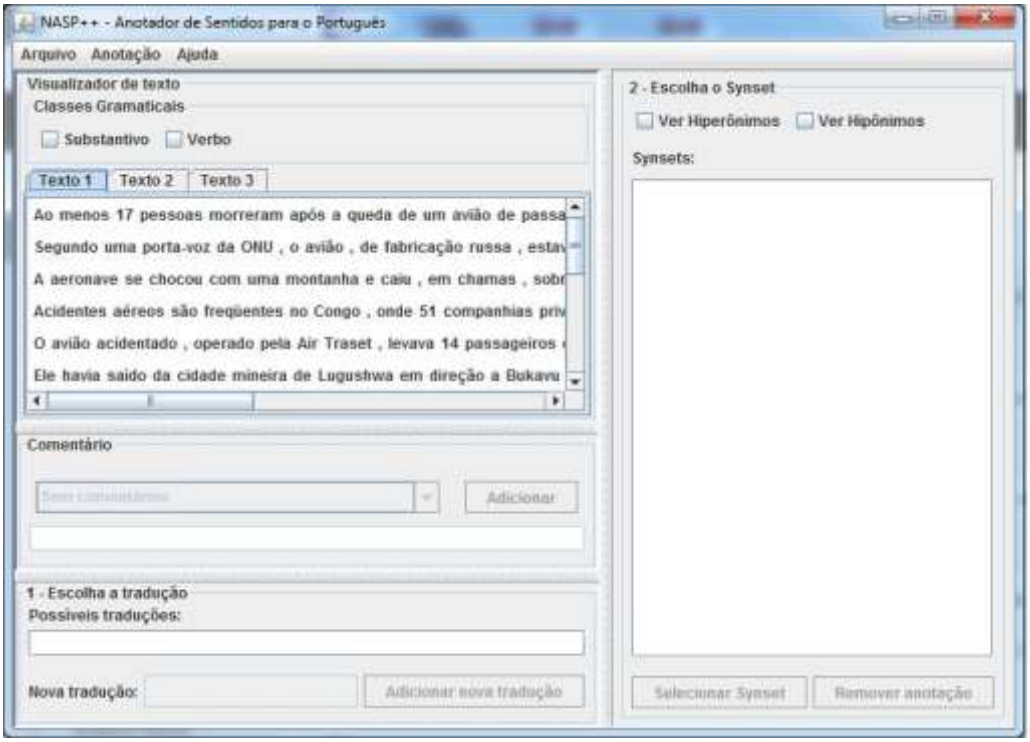
Na Figura 7, apresenta-se a tela principal do NASP++, composta pelos campos: (i) visualizador dos textos-fonte (A); (ii) painel para exibição e seleção das traduções (B); (iii) painel para exibição e seleção dos *synsets* (C), e (iv) painel para anotação de comentários (D).

Figura 7 – Tela principal do NASP++.



O campo A é responsável por exibir os textos-fonte da coleção. No caso da abertura de dois ou três textos-fonte, essa janela será composta por duas ou três abas de exibição, uma para cada texto, como pode ser visto na Figura 8. Nesse mesmo campo (A), o anotador escolhe a classe das palavras que irá anotar (nome ou verbo).

Figura 8 – Exibição dos textos



Na Figura 9, os três textos que compõem uma coleção a ser anotada foram carregados e exibidos para anotação. Nos textos-fonte exibidos, as unidades lexicais automaticamente etiquetadas pelo MXPost como nome estão destacadas em vermelho ou em verde**.** As unidades em verde já foram previamente anotadas; as em vermelho ainda precisavam ser anotadas. No Texto 1 da Figura 9, por exemplo, a anotação tem início com o primeiro nome destacado em vermelho, ‘‘avião’’.

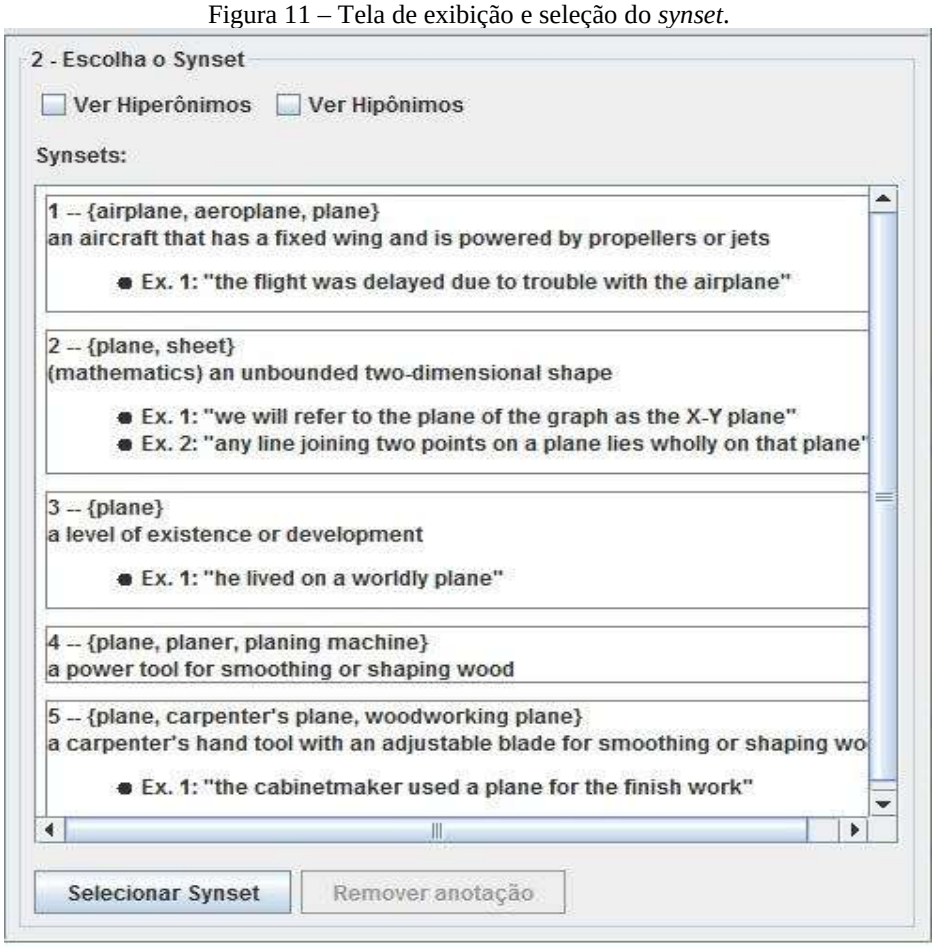
Figura 9 – Tela de visualização dos textos.

Ao clicar em “avião”, ainda no campo A, o editor recupera automaticamente todas as possíveis traduções em inglês para a palavra em questão com base no dicionário WordReference®. As possíveis traduções são exibidas no painel B. Para “avião”, o editor recuperou dois equivalentes de tradução, “*plane*” e “*airplane*” (Figura 10).

Figura 10 – Tela com a lista de traduções possíveis

Ao escolher a tradução “*plane*”, o editor recupera automaticamente todos os *synsets* da WN.Pr que possuem esse nome como elemento constitutivo, como pode ser visto na Figura 11. A lista de *synsets* é exibida na janela (C).

Figura 11 – Tela de exibição e seleção do *synset*.



Caso o NASP++ não tenha sugerido traduções adequadas para as palavras em português, o editor fornece a possibilidade de inserção de um equivalente em inglês por meio do campo "Nova tradução", localizado na parte inferior do campo B. Essa inserção deve ser realizada manualmente. Somente por meio da escolha de um equivalente em inglês, o NASP++ é capaz de exibir, para o anotador, os *synsets* da WN.Pr que possuem tal equivalente e que podem representar o conceito da palavra a ser anotada. É possível observar ainda, na Figura 11, que o editor recuperou 5 *synsets* para o nome "*plane*", sendo que cada um deles representa ou codifica um conceito distinto.

O NASP++ também recupera a glosa (ou seja, definição informal do conceito) e os exemplos de uso (ou frases-exemplo) de algumas palavras que constituem o *synset*. O primeiro *synset* recuperado e exibido ao anotador, por exemplo, foi {airplane, aeroplane, plane}, cuja glosa é "*an aircraft that has a fixed wing and is powered by propellers or jets*" (uma aeronave que tem uma asa fixa e é alimentada por hélices ou jatos). Para "*plane*", com o sentido representado pela glosa em questão, tem-se a frase-exemplo "*the flight was delayed due to trouble with the airplane*" (o vôo atrasou devido a problemas com o avião). As glosas e as frases-exemplo auxiliam a identificação do *synset* que representa ou codifica o conceito subjacente ao nome original em português de forma mais adequada. Dentre os *synsets* recuperados, cabe ao anotador selecionar o que mais adequadamente representa o conceito subjacente ao nome "avião" no texto-fonte.

Caso os *synsets* constituídos pelo equivalente de tradução ("*plane*"), as glosas e as frases-exemplo não sejam suficientes para se definir a representação mais adequada do sentido do nome em português, o editor oferece, na parte superior da janela (C) da tela principal, a exibição dos hiperônimos e hipônimos dos *synsets* listados.

Dentre os *synsets* recuperados, para selecionar o mais adequado, ou seja, aquele que será utilizado como rótulo semântico para a anotação da palavra em português, o anotador deve clicar sobre o *synset* em questão, por exemplo, {airplane, aeroplane, plane}. Após selecionado o *synset*, o processo de anotação semântica é finalizado por meio do botão "Selecionar *synset*", localizado na parte inferior da janela (C) do editor. Na sequência, o editor exibe uma janela de confirmação. Diante da certeza sobre a escolha do *synset*, o anotador deve clicar no botão "Sim".

Assim que um *synset* é selecionado, a palavra sob anotação ("avião") é destacada no campo "visualizador de texto" em verde, como ilustrado na Figura 12. O destaque indica que foi associado um rótulo conceitual (*synset*) ao nome.

Uma vez que uma palavra *x* tenha sido anotada com um sentido *y*, todas as demais ocorrências de *x* na coleção também são pré-anotadas pelo editor com *y*. Observa-se na Figura 12, por exemplo, que outras ocorrências de "avião" foram pré-anotadas com o *synset* selecionado para a anotação da primeira ocorrência de "avião". Cabe ressaltar que a pré-anotação semântica é realizada para todas as ocorrências do nome "avião", independentemente de sua forma flexionada. Assim, caso ocorra o nome "aviões", este também será pré-anotado, como pode ser visto na Figura 12. Ao anotador humano, cabe a tarefa de verificar se, de fato, o sentido/*synset* pré-anotado é pertinente para cada ocorrência do nome na coleção.

Figura 12 – Anotação da 1ª ocorrência de "avião" e pré-anotação das demais.

Ao final, o NASP++ gera um arquivo no formato XML (do inglês, *Extensible Markup Language*), um dos mais utilizados na anotação de *corpus*. Na Figura 13, ilustra-se o arquivo gerado pelo NASP++ após a anotação do nome "avião". Nessa Figura, observa-se que a palavra "avião" recebeu a etiqueta "N", sendo categorizado como "substantivo" e lematizado para "avião". Entre as traduções disponibilizadas ("*plane*" e "*airplane*"), "*plane*" foi a selecionada (o que é indicado pelo valor "*true*"). Entre os *synsets* que possuem "*plane*" como elemento constitutivo, o anotador selecionou o *synset* codificado pelo ID 2691156, sendo que essa seleção também é indicada pelo valor "*true*".

Figura 13 – Formato XML da anotação conceitual gerada pelo NASP++.

3.3.2. Os procedimentos de anotação

Em conformidade com Nóbrega (2013), a anotação foi realizada com base em diretrizes gerais e específicas. Quanto às **diretrizes gerais** descritas a seguir, a etapa VI foi a única estabelecida neste trabalho; as demais foram incorporadas de Nóbrega.

I – *Ler integralmente os textos-fonte da coleção*

A leitura integral dos textos-fonte de uma coleção como etapa inicial da anotação visa à familiarização do(s) anotador(es) com o seu conteúdo global, o que auxilia na identificação dos conceitos/*synsets* subjacentes às unidades lexicais.

II - *Iniciar a anotação preferencialmente pelo texto-fonte de menor tamanho*

A anotação deve ter início preferencialmente pelo texto de menor extensão. Dessa forma, os textos maiores apresentarão uma quantidade maior de palavras pré-anotadas, o que torna a tarefa mais dinâmica. Somente após anotar todas as palavras do texto inicial, deve-se anotar outro texto da coleção.

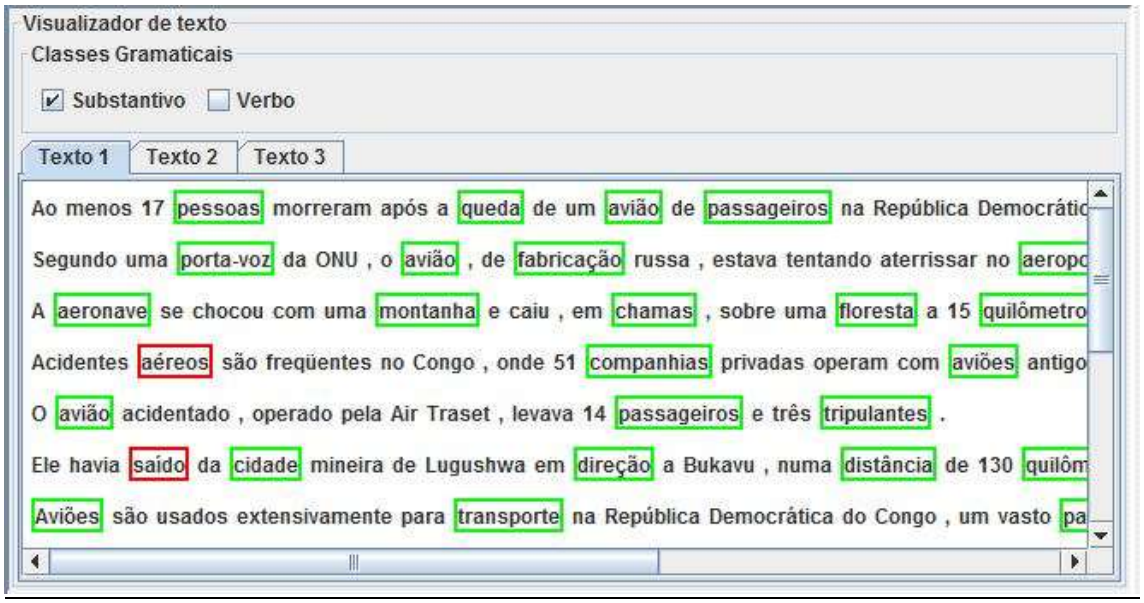
III – *Anotar todos os nomes comuns e siglas*

Todas as palavras ou expressões etiquetadas como "nome" pelo etiquetador morfossintático automático MXPost devem ser anotadas. A inclusão das siglas justifica-se pelo fato de que estas são relevantes para a veiculação do conteúdo no gênero jornalístico.

IV – *Ignorar os casos erroneamente anotados como nome*

Os casos de ruídos (isto é, palavras anotadas equivocadamente como nome) produzidos pelo *tagger* devem ser ignorados, não sendo, portanto, alvo da anotação conceitual. Na Figura 14, por exemplo, o adjetivo "aéreos" e o verbo "saído" como nomes, destacando-os em vermelho. De acordo com essa diretriz, tais casos foram ignorados.

Figura 14 – Ilustração dos casos de erros gerados pelo *tagger*.



V – *Avaliar as pré-anotações*

Todas as palavras ou expressões pré-anotadas (indicadas em amarelo) devem ser analisadas para confirmar ou não a seleção inicial dos *synsets*.

VI - *Revisar a anotação prévia dos nomes mais frequentes da coleção*

A anotação realizada por Nóbrega (2013) deve ser revisada. Isso foi necessário porque se verificou que a anotação de alguns dos nomes mais frequentes não estava sendo recuperada adequadamente. Esse foi o caso, por exemplo, de "assessor", que, embora ocorresse 5 vezes em uma mesma coleção, apenas uma ocorrência anotada era exibida pelo NASP++. Nesses casos, a anotação devia ser corrigida.

Além das diretrizes gerais, estabeleceu-se um conjunto de **diretrizes específicas** para completar a anotação dos nomes nas 5 coleções selecionadas a partir do CSTNews. Tal conjunto combina as diretrizes específicas de Nóbrega (2013) e outras que foram elaboradas devido à peculiaridade dos casos que ainda não tinham sido anotados.

I – *Anotar somente o núcleo de expressões multipalavras com o synset correspondente ao conceito representado pela expressão*

Tendo em vista que o *tagger* constitutivo do NASP++ anota somente unidades simples (ou seja, sequências de caracteres separadas por espaços em branco), estabeleceu-se que somente o núcleo de uma expressão multipalavra deve ser anotado com o *synset* que representa o conceito expresso pela expressão. Esse é o caso, por exemplo, de "**pista de aterrisagem**" que deve ser associado ao *synset* {***runway***}. Assim, anota-se apenas o núcleo "pista" cujo *synset* {*runway*} representa mais adequadamente o conceito ("*a strip of level paved surface where planes can take off or land*") (isto é, "uma faixa nivelada de superfície pavimentada onde aviões podem decolar ou pousar") subjacente à expressão.

II – *Anotar os nomes constitutivos de uma expressão multipalavras quando não há um synset correspondente ao conceito veiculado pela expressão*

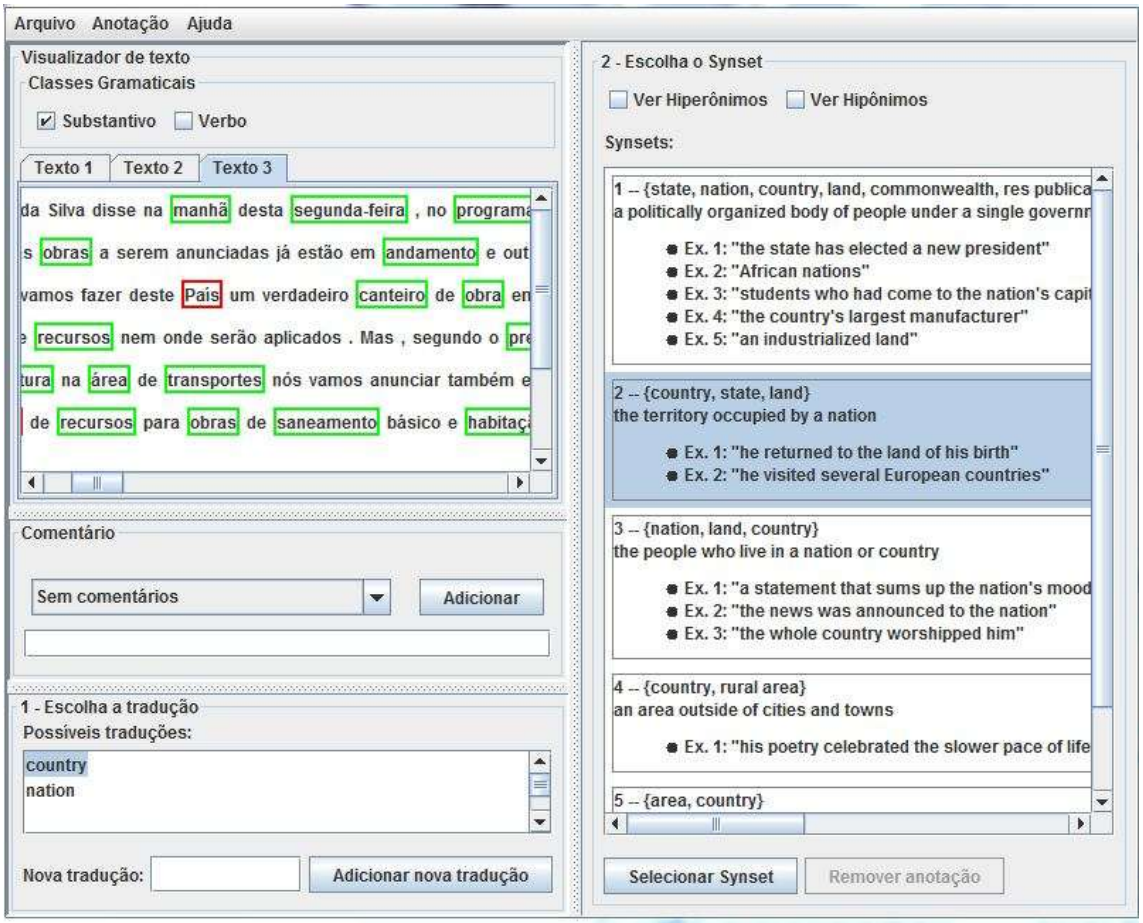
Diante de expressões multipalavras como "canteiro de obras" em "[...] obras de infra-estrutura que transformarão o Brasil em um "verdadeiro **canteiro de obras**" [...]", ambos os nomes "canteiro" e "obras" foram anotados. Isso se justifica porque não há um *synset* que codifica o conceito subjacente a "canteiro de obras" (isto é, "espaço ao redor de uma construção, onde os operários ficam em alojamentos provisórios e que também é usado como depósito de materiais"[26]) na WN.Pr. Assim, na tentativa de se explicitar esse conceito específico, anotam-se os nomes constitutivos na expressão com seus respectivos *synsets*.

III – *Analisar todas as traduções sugeridas pelo NASP++ e selecionar a mais adequada*

Todas as traduções sugeridas pelo dicionário WordReference® devem ser analisadas antes da seleção definitiva do equivalente de tradução. O objetivo dessa regra é o de selecionar a tradução mais adequada em inglês, posto que a recuperação dos conceitos (*synsets*) na WN.Pr é diretamente determinada pelo equivalente de tradução. Na Figura 15, observa-se que, para a palavra "país", por exemplo, o editor sugeriu 2 equivalentes na janela de possíveis traduções, "*country*" e "*nation*". Ao analisar os *synsets* recuperados pelas duas traduções possíveis, verificou-se que somente "*country*" compõe o *synset* 2 ({***country***, *state*, *land*}), o qual representa mais adequadamente o conceito subjacente a "país" ("*the territory occupied by a nation*") (isto é, "o território ocupado por uma nação"). Assim, o equivalente escolhido foi "*country*".

[26]https://michaelis.uol.com.br/moderno-portugues/busca/portugues-brasileiro/canteiro/

Figura 15 – Ilustração das "possíveis traduções" e *synsets* recuperados.



IV – *Testar diferentes equivalentes antes de adicionar uma tradução ao NASP++*

Quando o editor não recupera possíveis equivalentes de tradução do WordReference®, essa regra estabelece que todas as possibilidades de tradução identificadas em fontes externas ao editor sejam testadas antes de se inserir efetivamente um equivalente por meio do botão "Adicionar nova tradução". Essa regra foi estabelecida porque uma possível tradução *y* pode não estar contida na WN.Pr, mas isso não significa necessariamente que o conceito não esteja codificado na base. Esse é o caso, por exemplo, de "pista" (que é o núcleo da expressão multipalavra "pista de aterrisagem") em que nenhuma das 4 possíveis traduções oferecidas pelo WordReference® (*track, hint, clue, dance floor*) contempla o conceito adequado à expressão. Assim, busca-se o seu equivalente de tradução mais adequado em outras fontes (nesse caso, "*runway*") e insere-o no campo "nova tradução". Como fontes externas, destacam-se a versão *online* do dicionário *Michaelis Moderno Dicionário Inglês & Português*[27] e os diferentes dicionários disponíveis no site

[27]http://michaelis.uol.com.br/

Cambridge Dictionaries Online[28], além dos serviços *online* denominados *Google translate*[29] e o Linguee[30].

V – *Analisar os synsets recuperados pelo NASP++ e selecionar o mais adequado*

Deve-se analisar todos os *synsets* recuperados da WN.Pr pelo NASP++ e selecionar o que representa mais adequadamente o conceito subjacente a um nome *x*. Essa regra foi formulada porque a WN.Pr comumente apresenta conceitos muito próximos, cuja distinção nem sempre é simples. Por exemplo, ao selecionar a tradução "*country*", o NASP++ recupera 5 *synsets* no total. Utilizando a barra de rolagem da interface ilustrada na Figura 14, todos os 5 *synsets* podem ser vistos. Dois deles, {*country, state, land*} ("isto é, "o território ocupado por uma nação") e {*area, country*} ("uma região geográfica particular geralmente servindo a algum propósito especial ou distinguido por seu povo ou cultura ou geografia), codificam conceitos que podem ser considerados relativamente semelhantes, o que requer uma análise atenta dos *synsets*.

VI – *Selecionar synsets hiperônimos*

Diante da ausência de um *synset* que representa o conceito subjacente a um nome, essa regra estabelece que o *synset* hiperônimo (ou seja, mais genérico) seja selecionado. Para a anotação, por exemplo, da sigla CPI em "[...] O relatório final da **CPI** do Apagão da Câmara, que começou a ser lido nesta terça-feira [...]", que expressa um conceito específico do cenário político brasileiro ("Comissão Parlamentar de Inquérito"), não há um *synset* diretamente correspondente na WN.Pr. Assim, essa regra estabelece que a sigla seja anotada com o *synset* que representa um conceito mais genérico, como "investigação". A partir da tradução *investigation*, o NASP++ recuperou dois *synsets* (Quadro 8). Dentre eles, selecionou-se o *synset* 1 {probe, investigation}, que codifica o conceito "*an inquiry into unfamiliar or questionable activities*" ("um inquérito sobre atividades desconhecidas ou questionáveis").

[28]http://dictionary.cambridge.org/
[29]http://translate.google.com.br/
[30]http://www.linguee.com.br/

Quadro 8 – Seleção de conceitos/*synsets* hiperônimos

	Synset	**Glosa/Frase-exemplo (Tradução da glosa)**
1	{probe, investigation}	an inquiry into unfamiliar or questionable activities. "there was a congressional probe into the scandal" (“uma investigação sobre atividades desconhecidas ou questionáveis”)
2	{investigation, investigating}	the work of inquiring into something thoroughly and systematically (“o trabalho de se investigar algo de forma completa e sistemática”)

3.3.3. As estatísticas da extensão da anotação

Com base nas diretrizes anteriormente descritas e na ferramenta NASP++, completou-se a anotação conceitual dos nomes dos 5 *clusters* iniciais do CSTNews, totalizando 14 textos. Tal anotação foi realizada em sessões diárias de 3 horas no período de 8 semanas. As horas diárias despendidas não foram necessariamente consecutivas, uma vez que a anotação dos *clusters* mais extensos ou com conceitos mais complexos demandou muito esforço cognitivo, não podendo ser realizada em uma única sessão contínua. Considerando a quantidade de anotadores, as horas diárias, o tempo total de anotação e a quantidade de nomes anotados, exibe-se, no Quadro 9, uma comparação entre os processos realizados por Nóbrega (2013) e o conduzido neste trabalho.

Quadro 9 – Comparação entre a anotação prévia do CSTNews e a deste trabalho.

	Quant. anotadores	Quant. horas diárias	Quant. tempo total	Quant. nomes
Nóbrega (2013)	10	01	06 semanas	423
Anotação atual	01	03	08 semanas	327

Assim, anotou-se o total de 327 nomes distintos para os quais atribuíram-se 453 ocorrências de conceitos nominais, sendo 314 conceitos distintos. Nas Tabelas 1 e 2, exibem-se os dados estatísticos da anotação.

Tabela 1 – Quantidade de nomes anotados por *cluster* neste trabalho

Cluster	**Total de palavras**	**Total de nomes distintos**	**Nomes distintos anotados por Nóbrega**	**Nomes distintos anotados neste trabalho**
C1	432	44	8	36
C2	996	48	14	34
C3	1249	151	19	132
C4	833	81	16	65
C5	572	72	12	60
TOTAL	**4082**	**396**	**69**	**327**

Tabela 2 – Quantidade de conceitos/*synsets* por *cluster*

Cluster	**Total de conceitos distintos**	**Quant. de conceitos distintos atribuídos por Nóbrega**	**Quant. de conceitos distintos atribuídos neste trabalho**	**Total de ocorrências de conceitos/*synsets* anotadas neste trabalho**
C1	42	8	34	59
C2	49	15	34	52
C3	142	17	125	181
C4	81	18	63	91
C5	70	12	58	70
TOTAL	**384**	**70**	**314**	**453**

No Quadro 10, ilustram-se os 151 nomes distintos do *cluster* C3 anotados e seus conceitos/*synsets* correspondentes, sendo que 19 destes já haviam sido anotados por Nóbrega (2013), pois pertenciam ao grupo dos 10% mais frequentes da coleção.

Quadro 10 – Conjunto de nomes e seus respectivos *synsets* (C3).

Nome	**Conceito/Synset**
acidente	accident
aeronave	aircraft
aeroporto	airport, airdrome, aerodrome, drome
água	water, H2O
airbus	airbus
assessor	adviser, advisor, consultant
assinante	subscriber, reader
aviação	aviation
avião	airplane, aeroplane, plane
bombeiro	fireman, firefighter, fire fighter, fire-eater
braço	arm
brigadeiro	brigadier, brigadier general
câmera	television camera, tv camera, camera
carga	freight, freightage
caso	event, case
causa	cause, reason, grounds
cena	scene
checagem	confirmation, verification, check, substantiation
chuva	rain, rainfall
coluna	column, editorial, newspaper column
colunista	columnist, editorialist
comemoração	celebration, festivity
companhia	company
comportamento	behavior, behaviour, conduct, doings
conclusão	conclusion
condição	condition, status

Quadro 10 – Conjunto de nomes e seus respectivos *synsets* (C3) (continuação)

conhecimento	understanding, apprehension, discernment, savvy
controle	control, controller
corpo	body, organic structure, physical structure
culpa	fault
decolagem	takeoff
defeito	defect, fault, flaw
departamento	department, section
derrapagem	brake shoe, shoe, skid
desaceleração	deceleration
dia	day, twenty-four hours, twenty-four hour period, 24-hour interval, solar day, mean solar day
	day
edição	issue, number
efusão	effusion, gush, outburst, blowup, ebullition
elemento	component, constituent, element, factor, ingredient
emissora	television, television system
empregado	employee
empresa	company
equipamento	equipment
escoamento	drain, drainage
espaço	space
especialista	specialist, specializer, specialiser
esquerda	left
exibição	exhibition
fabricante	manufacturer, maker, manufacturing business
falha	defect, fault, flaw
família	family, household, house, home, menage
fato	fact
fator	factor
filme	video, picture
final	end
força	power, force
frente	front
fumaça	smoke, smoking
futebol	soccer, association football
gesto	gesticulation
gestual	gesticulation
governo	government, authorities, regime
h	hour, time of the day
hipótese	hypothesis, possibility, theory
hora	moment, minute, second, instant
imagem	picture, image, icon, ikon
imprensa	press, public press

Quadro 10 – Conjunto de nomes e seus respectivos *synsets* (C3 (continuação)

indignação	indignation, outrage
influência	influence
informação	information, info
início	beginning
instrumento	instrument
investigação	investigation, investigating
jornal	newspaper, paper
jornalista	journalist
julho	July
lado	side
limite	terminus ad quem, terminal point, limit
mão	hand, manus, mitt, paw
matéria	topic, subject, issue, matter
mecanismo	mechanism
membro	member, fellow member
momento	moment, minute, second, instant
morte	death
motor	engine
noite	night
normal	regular, habitue, fixture
nota	note
notícia	information, info
noticiário	news
número	number
obra	work, piece of work
obstáculo	obstacle, obstruction
ocasião	occasion
ocupante	traveler, traveller
ontem	yesterday
país	state, nation, country, land, commonwealth, res publica, body politic
parte	part, portion
peça	piece
perito	expert
pessoa	person, individual, someone, somebody, mortal, soul
piloto	pilot, airplane pilot
pista	site, land site
possibilidade	hypothesis, possibility, theory
pouso	landing
prédio	building, edifice
presidência	presidency, presidential term, administration
presidente	president
problema	trouble, problem
procedimento	operation, procedure

Quadro 10 – Conjunto de nomes e seus respectivos *synsets* (C3) (continuação)

propulsão	propulsion
quarta-feira	Wednesday, Midweek, Wed
quinta-feira	Thursday, Th
raiva	anger, choler, ire
ranhura	groove, channel
reação	reaction, response
realização	accomplishment, achievement
recomendação	recommendation
registro	record
reportagem	coverage, reporting, reportage
reto	straightway, straight
reverso	reverse, reverse gear
reversor	reverse, reverse gear
revisão	inspection, review
risco	risk, peril, danger
segunda-feira	Monday, Mon
segundo	second, sec, s
segurança	security, security department
sentido	direction, way
sistema	system
solo	land, dry land, earth, ground, solid ground, terra firma
sul	south
suspeita	intuition, hunch, suspicion
terça	Tuesday, Tue
terça-feira	Tuesday, Tue
terra	land, dry land, earth, ground, solid ground, terra firma
tipo	type
torcida	cheering, shouting
torre	tower
tragédia	calamity, catastrophe, disaster, tragedy, cataclysm
trajeto	path, route, itinerary
transmissão	transmission
transporte	transportation, shipping, transport
turbina	turbine
velocidade	speed, velocity
versão	version
vice-presidente	vice president, V.P.
vídeo	video, picture
vilão	villain, scoundrel
voo	flight
zona	zone, geographical zone

Dentre todos os nomes anotados nos 5 *clusters*, incluindo-se a parcela dos 10% mais frequentes, constataram-se 18 casos de sinonímia (cf. Quadro 11) e 6 casos de polissemia (cf. Quadro 12). Aqui, emprega-se a noção de sinonímia contextual, a mesma utilizada para a construção dos *synsets* da WN.Pr, ou seja, "duas unidades lexicais são sinônimas em um contexto C, se a substituição de uma pela outra em C não altera o valor de verdade denotado por C" (MILLER, FELLBAUM, 1991). Já a polissemia ocorre quando um item lexical possui significados diferentes em um mesmo contexto (FIORIN, 2003).

Quadro 11 – Lista dos 18 casos de sinonímia nos 5 *clusters* anotados.

Cluster	***Nome e Contexto***	***Synset***
C1	As vítimas do acidente foram 14 passageiros e três membros da **tripulação**.	{crew}
	O avião acidentado, operado pela Air Traset, levava 14 passageiros e três **tripulantes**.	
	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.	{kilometer, kilometre, km, klick}
	Ele havia saído da cidade mineira de Lugushwa em direção a Bukavu, numa distância de 130 **quilômetros**.	
C2	A pesquisa de hoje apresentou uma variação na lista espontânea (quando os entrevistados dizem em quem pretendem votar sem um **cartão** com nomes).	{list, listing}
	(...) é a primeira vez que o Ibope utiliza a **lista** oficial de candidatos a presidente da República.	
C3	De acordo com a **companhia** aérea, a recomendação da Airbus -- fabricante do avião-- é que a revisão no reversor seja feita até dez dias depois de o defeito ser detectado.	{company}
	Na quarta-feira, o presidente da **empresa**, Marco Antonio Bologna, havia declarado que o avião passara por checagem justamente no dia 13 e estava em perfeito estado.	
	O Airbus-A320 da TAM acidentado em São Paulo na última terça-feira (17) tinha um **defeito** no reverso da turbina direita desde o último dia 13.	{defect, fault, flaw}
	A **falha** no reversor --mecanismo que ajuda o avião a frear-- foi detectada pelo sistema eletrônico de checagem da própria aeronave, que continuou voando nos dias seguintes, com o reversor direito desligado.	
	Reportagem da colunista Eliane Cantanhêde (íntegra disponível só para assinantes do jornal ou do UOL), publicada na edição da Folha desta quinta, mostra que nova versão da Infraero inclui falha mecânica entre as **hipóteses** para o acidente --o maior da aviação brasileira.	{hypothesis, possibility, theory}
	Caso a **possibilidade** de que uma falha na aeronave tenha provocado a tragédia com o Airbus, as suspeitas de que o acidente esteja relacionado a falhas na recém-reformada pista do aeroporto perdem força.	
	(...) o que explicaria, por exemplo, por que o piloto não conseguiu parar o avião, que continuou em velocidade bem alta depois de tocar o **solo** e girou (...)	{land, dry land, earth, ground, solid ground, terra firma}
	Além dos ocupantes, o acidente vitimou pessoas em **terra**, entre elas funcionários da empresa de transporte de cargas.	

Quadro 11 – Lista dos 18 casos de sinonímia nos 5 *clusters* anotados (continuação).

C3	O reverso é um instrumento auxiliar na **hora** de frear o avião (...)	{moment, minute, second, instant}
	Procurado pelo JB, Marco Aurélio se disse incapaz de relembrar todo o seu gestual em um dia inteiro. Também disse não saber em que **momento** foi filmado.	
	O Airbus-A320 da TAM acidentado em São Paulo na última terça-feira (17) tinha um defeito no **reverso** da turbina direita desde o último dia 13.	{reverse, reverse gear}
	A falha no **reversor** --mecanismo que ajuda o avião a frear-- foi detectada pelo sistema eletrônico de checagem da própria aeronave (...)	
	Às 18h50 da última **terça**, sem controle, a aeronave --que havia decolado de Porto Alegre-- passou pela pista de Congonhas (...)	{Tuesday, Tue}
	O Airbus-A320 da TAM acidentado em São Paulo na última **terça-feira** (17) tinha um defeito no reverso da turbina direita desde o último dia 13	
	O presidente da Infraero (estatal que administra os aeroportos do país), brigadeiro José Carlos Pereira, disse ontem que os peritos detectaram "fumaça forte" no motor esquerdo do Airbus-A320 da TAM no **filme**, de acordo com a reportagem publicada pela Folha.	{video, picture}
	Os **vídeos** comparam pousos de outras aeronaves com o do Airbus A-320 acidentado.	
	Seu assessor puxava os braços ao lado do corpo simulando o **gesto** de copular.	{gesticulation}
	Procurado pelo JB, Marco Aurélio se disse incapaz de relembrar todo o seu **gestual** em um dia inteiro.	
	Esta notícia acabou criando uma polêmica em torno da reação de membros do governo ao tomarem conhecimento da **informação** através da imprensa.	{information, info}
	Esta **notícia** acabou criando uma polêmica em torno da reação de membros do governo ao tomarem conhecimento da informação através da imprensa.	
C4	O congestionamento esteve ainda maior às 9**h**, quando chegou a 113 km de extensão para uma média de 32 km.	{hour, time of day}
	Às 9 **horas**, a cidade tinha 113 km de lentidão, sendo que a média para o horário é de 82 km, segundo a Companhia de Engenharia de Tráfego (CET).	
	Naquele horário, segundo a CET (Companhia de Engenharia de Tráfego), havia 110 **km** de congestionamento em toda a cidade enquanto a média para o horário era de 76 **km**.	{kilometer, kilometre, km, klick}
	De acordo com a CET, o índice de congestionamento era de 54 **quilômetros** às 8h, bem acima da média.	
	O Aeroporto de Congonhas funciona normalmente nesta **segunda**.	{Monday, Mon}
	Nesta **segunda-feira**, os termômetros não devem passar de 18 graus.	
C5	Mas, diante da dificuldade para encontrar pessoas que aceitassem assumir uma das diretorias da **agência** reguladora, após a renúncia de três diretores, Jobim decidiu indicar a economista para o cargo.	{agency, federal agency, government agency, bureau, office, authority}
	Inicialmente, Solange Vieira, que é assessora especial de Jobim havia sido escolhida para comandar a **Secretaria** Nacional de Aviação Civil, a ser criada na estrutura do ministério, segundo a assessoria de imprensa do ministério.	
	Mas também conquistou muito inimigos, principalmente entre os **dirigentes** de fundos de pensão (...)	{director}
	Também renunciaram ao cargo de **diretor** da Anac Denise Abreu e Leur Lomanto.	

Quadro 12 – Lista dos 6 casos de polissemia nos 5 *clusters* anotados.

Cluster	**Nome**	***Synset* / glosa / exemplo**
C2, C3, C4	dia	**{day}** (*a day assigned to a particular purpose or observance*) Uma possível disputa entre Lula e Heloísa num segundo turno não foi cogitada pelo Ibope nesse levantamento do **dia** 25. (C2)
		{day, twenty-four hours, twenty-four hour period, 24-hour interval, solar day, mean solar day} (*time for Earth to make a complete rotation on its axis*) Na comparação com a pesquisa Ibope divulgada há nove **dias**, Lula permaneceu estável (...)(C2)
C2	número	**{number, figure}** (*the property possessed by a sum or total or indefinite quantity of units or individuals*) Os **números** da última pesquisa Datafolha são semelhantes aos divulgados nesta sexta-feira.
		{number} (*a numeral or string of numerals that is used for identification and may be attached to accounts, memberships, etc.*) A pesquisa foi realizada entre os dias 29 e 31 de julho e foi registrada no TSE com o **número** 12.197/2006.
C4	estado	**{state}**(*the way something is with respect to its main atributes*) O **estado** de atenção na cidade foi suspenso às 9h25.
		{state, province} (*the territory occupied by one of the constituent administrative districts of a nation*) A temperatura deve permanecer baixa, por conta da massa de ar polar que acompanha a frente fria que passa pelo **estado**.
	centro	**{center, centre}** (*a building dedicated to a particular activity*) A forte chuva em São Paulo complicava o trânsito na manhã desta segunda-feira, 16, e fez com que o **Centro** de Gerenciamento de Emergência (CGE) da Prefeitura colocasse a cidade em estado de atenção.
		{center, centre, middle, heart, eye} (*an area that is approximately central within some larger region*) O motorista que percorria a Radial Leste enfrentava 4,3 de trânsito ruim, a partir do Viaduto Pires do Rio até a Rua Wandenkolk, no sentido **centro.**

Tendo em vista o número reduzido de palavras polissêmicas encontradas nas 5 coleções anotadas, comprova-se o princípio heurístico de que, em dada coleção sobre mesmo assunto, as palavras tendem a assumir o mesmo significado (MIHALCEA, 2006). Sobre a sinonímia, ressalta-se que a quantidade mais elevada de casos por coleção corrobora a aplicação de métodos léxico-conceituais na sumarização em detrimento de métodos puramente pautados na ocorrência de palavras, os quais não capturam o fenômeno da sinonímia.

Após a extensão da anotação dos nomes, deu-se início à próxima etapa do trabalho, que consistiu na proposição de método(s) extrativo(s) de SAM cuja seleção de conteúdo fosse pautada nos conceitos lexicalizados na coleção.

4. PROPOSIÇÃO DE MÉTODOS LÉXICO-CONCEITUAIS DE SAM

Investigaram-se 4 métodos de SAM baseados na frequência de conceitos lexicalizados na coleção. Tais métodos seguem um mesmo algoritmo (isto é, sequência de passos), o qual foi adaptado de Tosta (2014) (Quadro 13). A principal diferença entre os 4 métodos reside na estratégia de pontuação das sentenças, que é codificado pelo item 4 do Quadro 13.

Quadro 13 – Algoritmo genérico de SAM baseado na frequência léxico-conceitual.

Método de Frequência dos Conceitos	
ANÁLISE	1. Anotar cada um dos textos da coleção em nível léxico-conceitual, rotulando os nomes comuns e/ou verbos com os *synsets* da WordNet de Princeton. 2. Calcular a frequência de cada *synset* na coleção.
TRANSFORMAÇÃO	3. Calcular uma taxa de compressão *x* (em função do número de palavras do maior texto da coleção) 4. Pontuar as sentenças em função da frequência de ocorrência dos *synsets*/conceitos na coleção 5. Ranquear as sentenças em função da pontuação dos conceitos 6. Selecionar a 1ª sentença do ranque 7. Caso a taxa de compressão não tenha sido atingida: a) Selecionar a próxima sentença do ranque b) Verificar a redundância da sentença em questão com a já selecionada c) Selecionar a sentença somente se não for redundante 8. Repetir o passo 6 até que a taxa de compressão *x* seja atingida
SÍNTESE	9. Justapor as sentenças na ordem em que foram selecionadas 10. Ordenar as sentenças pelo Método Baseado na Posição Textual (OPT).

No algoritmo genérico do Quadro 13, doravante LCFSumm (do inglês, *Lexical Concept-Frequency Summarization*), a análise consiste na segmentação das sentenças dos textos-fonte e no cálculo da frequência de ocorrência dos conceitos na coleção.

O primeiro passo da fase de transformação é o cálculo da taxa de compressão. No cenário da SAM, a taxa de compressão é comumente especificada com base na extensão do maior texto da coleção (medido em número de palavras). Assim, diante de uma taxa

de 70%, por exemplo, o extrato terá tamanho equivalente a 30% do maior texto-fonte da coleção. Se o maior texto de uma coleção contiver 100, por exemplo, a taxa estipulada em 70% gera um extrato com 30 palavras.

Na sequência, realizam-se a pontuação e o ranqueamento das sentenças dos textos-fonte da coleção em função da frequência de seus conceitos na coleção. Neste trabalho, 4 estratégias distintas foram empregadas na pontuação, as quais variam quanto ao tipo de conceito lexicalizado e quanto ao tipo de frequência, originando 4 métodos distintos:

- **LCFSummN**: método de SAM em que a pontuação das sentenças é feita com base na frequência simples dos conceitos lexicalizados por nomes na coleção;
- **LCFSummN-V**: método de SAM em que a pontuação das sentenças é feita com base na combinação da frequência simples dos conceitos nominais e verbais na coleção;
- **LCFSummN-pond**: método de SAM em que a pontuação das sentenças é feita com base na média ponderada da frequência dos conceitos nominais na coleção;
- **LCFSummN-V-pond**: método de SAM em que a pontuação das sentenças é feita pela combinação da média ponderada da frequência dos conceitos nominais e verbais.

No caso, a frequência simples foi utilizada porque sua relevância já fora comprovada em outros trabalhos da literatura. A média ponderada, no caso, foi calculada com base na fórmula *FS*peso/Quant. C em S*, em que FS equivale à "soma da frequência simples dos conceitos constitutivos de S" e o peso é um valor relativo à posição de S no texto-fonte. Vale ressaltar que o denominador da fórmula é a quantidade de conceitos distintos em S. Essa média ponderada buscou restringir o privilégio das sentenças mais longas e destacar a posição das mesmas no texto-fonte, posto que a localização já se mostrou uma estratégia relevante para a seleção de conteúdo na SAM.

A partir do ranque, seleciona-se a sentença de maior pontuação para iniciar o extrato. Caso a taxa de compressão não seja atingida, seleciona-se a segunda sentença do ranque, desde que esta não tenha conteúdo similar à sentença previamente selecionada. Para garantir que as sentenças selecionadas não sejam similares entre si, deve-se calcular um fator de redundância, comumente capturado pela *word overlap* (JURAFSKY, MARTIN, 2001). Neste trabalho, optou-se pela medida aqui denominada *concept overlap* (isto é, sobreposição de conceitos), por considerá-la mais "rica" que a pura sobreposição lexical. Para calcular a *concept overlap* entre um par de sentenças (S1 e S2), divide-se o número total de conceitos idênticos entre as sentenças (*Conceitos comuns*) pela soma do número total de conceitos de cada sentença (*Conceitos(S1) + Conceitos(S2)*), conforme fórmula abaixo. O resultado obtido será entre 0 e 1, sendo que, quanto mais próximo a 1 for

o resultado da fórmula, mais redundante será o par entre si, e, quanto mais próximo a 0, menos redundante.

$$Col\ (S1, S2) = \frac{\#Conceitos\ comuns}{\#Conceitos(S1)\ +\ \#Conceitos(S2)}$$

Caso a medida calculada entre uma sentença já selecionada e uma sentença do ranque candidata a compor o sumário seja superior a um limiar determinado empiricamente (do inglês, *threshold*), a sentença candidata não é selecionada.

O processo de seleção se repete até que o extrato obtido tenha uma extensão que mais se aproxime da taxa de compressão. Considerando que um extrato deva conter 30 palavras, por exemplo, verifica-se se a inclusão de uma sentença do ranque ultrapassa ou não esse limite. Caso a inclusão ultrapasse a extensão desejada em 15 palavras (totalizando 45) e a não-inclusão gere um extrato com 5 palavras a menos que o tamanho esperado (isto é, 25), opta-se pela não inserção da sentença.

Por fim, na síntese, constroi-se efetivamente os extratos. Para tanto, as sentenças selecionadas são comumente justapostas na ordem em que ocorreram nos textos-fonte, sob a hipótese de que a ordenação contribui para a coesão/estrutura textual.

4.1. Aplicação dos métodos LCFSummN e LCFSummN-pond

Tendo em vista o número reduzido de coleções anotadas, os 4 métodos foram aplicados de forma manual. Nesta seção, ilustra-se a aplicação de dois deles, o LCFSummN e o LCFSummN-pond. Para tanto, calculou-se uma taxa de compressão de 70%, que é a mesma empregada para a geração dos demais extratos que atualmente compõem o CSTNews. Assim, o tamanho dos extratos gerados equivale a 30% do maior texto-fonte da coleção. Para a ilustração do funcionamento dos métodos LCFSummN e LCFSummN-pond, utilizou-se a coleção C3, cujo maior texto-fonte possui 607 palavras, o que determinou que a taxa de 70% fosse de 182 palavras. O *cluster* C3, aliás, é composto por 3 textos compilados dos jornais *on-line Folha de São Paulo, Estadão* e *Jornal do Brasil* e possui um total de 50 sentenças, 151 nomes e 84 verbos. Os textos são do domínio "mundo" e relatam um "acidente aéreo com um avião da TAM".

4.1.1 A pontuação e ranqueamento das sentenças

O processo de pontuação de ambos os métodos se iniciou com o cálculo da frequência simples de cada *synset/*conceito nominal da coleção C3. No Quadro 14, tem-se o ranque dos conceitos em função da frequência simples na coleção.

Quadro 14 – Ranque dos conceitos nominais de C3 em função da frequência simples.

Posição	**Synset/conceito**	**Freq.**
1º	reverse, reverse gear	13
2º	airplane, aeroplane, plane	12
3º	defect, fault, flaw	10
3º	accident	10
4º	aircraft	9
4º	site, land site	9
5º	company	8
5º	information, info	8
6º	trouble, problem	6
7º	gesticulation	5
7º	day, twenty-four hours, twenty-four hour period, 24-hour interval, solar day, mean solar day	5
7º	adviser, advisor, consultant	5
7º	landing	5
8º	hypothesis, possibility, theory	4
8º	moment, minute, second, instant	4
8º	day	4
8º	airport, airdrome, aerodrome, drome	4
8º	flight	4
8º	person, individual, someone, somebody, mortal, soul	4
9º	Tuesday, Tue	3
9º	video, picture	3
9º	airbus	3
9º	calamity, catastrophe, disaster, tragedy, cataclysm	3
9º	cause, reason, grounds	3
9º	confirmation, verification, check, substantiation	3
9º	coverage, reporting, reportage	3
9º	issue, number	3
9º	picture, image, icon, ikon	3
9º	pilot, airplane pilot	3
9º	second, sec, s	3
10º	land, dry land, earth, ground, solid ground, terra firma	2
10º	beginning	2
10º	brigadier, brigadier general	2
10º	building, edifice	2
10º	condition, status	2
10º	control, controller	2

Quadro 14 – Ranque dos conceitos nominais de C3em função da frequência simples (continuação).

10º	death	2
10º	end	2
10º	engine	2
10º	expert	2
10º	fact	2
10º	government, authorities, regime	2
10º	intuition, hunch, suspicion	2
10º	night	2
10º	note	2
10º	president	2
10º	press, public press	2
10º	side	2
10º	smoke, smoking	2
10º	south	2
10º	specialist, specializer, specialiser	2
10º	speed, velocity	2
10º	system	2
10º	television camera, tv camera, camera	2
10º	turbine	2
10º	Wednesday, Midweek, Wed	2
10º	zone, geographical zone	2
11º	accomplishment, achievement	1
11º	anger, choler, ire	1
11º	arm	1
11º	aviation	1
11º	behavior, behaviour, conduct, doings	1
11º	body, organic structure, physical structure	1
11º	brake shoe, shoe, skid	1
11º	celebration, festivity	1
11º	cheering, shouting	1
11º	column, editorial, newspaper column	1
11º	columnist, editorialist	1
11º	component, constituent, element, factor, ingredient	1
11º	conclusion	1
11º	deceleration	1
11º	department, section	1
11º	direction, way	1
11º	drain, drainage	1
11º	effusion, gush, outburst, blowup, ebullition	1
11º	employee	1
11º	equipment	1
11º	event, case	1
11º	exhibition	1
11º	factor	1
11º	family, household, house, home, menage	1

Quadro 14 – Ranque dos conceitos nominais de C3em função da frequência simples (continuação).

11º	fault	1
11º	fireman, firefighter, fire fighter, fire-eater	1
11º	freight, freightage	1
11º	front	1
11º	groove, channel	1
11º	hand, manus, mitt, paw	1
11º	hour, time of the day	1
11º	indignation, outrage	1
11º	influence	1
11º	inspection, review	1
11º	instrument	1
11º	investigation, investigating	1
11º	journalist	1
11º	July	1
11º	left	1
11º	manufacturer, maker, manufacturing business	1
11º	mechanism	1
11º	member, fellow member	1
11º	Monday, Mon	1
11º	news	1
11º	newspaper, paper	1
11º	number	1
11º	obstacle, obstruction	1
11º	occasion	1
11º	operation, procedure	1
11º	part, portion	1
11º	path, route, itinerary	1
11º	piece	1
11º	power, force	1
11º	presidency, presidential term, administration	1
11º	propulsion	1
11º	rain, rainfall	1
11º	reaction, response	1
11º	recommendation	1
11º	record	1
11º	regular, habitue, fixture	1
11º	risk, peril, danger	1
11º	scene	1
11º	security, security department	1
11º	soccer, association football	1
11º	space	1
11º	state, nation, country, land, commonwealth, res publica, body politic	1
11º	straightway, straight	1
11º	subscriber, reader	1
11º	takeoff	1

Quadro 14 – Ranque dos conceitos nominais de C3em função da frequência simples (continuação)

11º	television, television system	1
11º	terminus ad quem, terminal point, limit	1
11º	Thursday, Th	1
11º	topic, subject, issue, matter	1
11º	tower	1
11º	transmission	1
11º	transportation, shipping, transport	1
11º	traveler, traveller	1
11º	type	1
11º	understanding, apprehension, discernment, savvy	1
11º	version	1
11º	vice president, V.P.	1
11º	villain, scoundrel	1
11º	water, H2O	1
11º	work, piece of work	1
11º	yesterday	1

De acordo com o LCFSummN, as sentenças foram pontuadas em função da soma da frequência simples de seus conceitos constitutivos. No Quadro 15, por exemplo, tem-se uma sentença ("*As imagens, divulgadas pela Infraero, mostram que o avião da TAM levou três segundos para fazer o trajeto na pista que, em condições, normais, levariam 11 segundos*") que possui 7 *synsets* ou conceitos nominais, sendo 6 distintos. No Quadro, os conceitos subjacentes aos nomes estão representados pelos números de identificação (ID) fornecido pela WN.Pr, dispostos entre os símbolos "< >". A frequência simples de ocorrência de cada um deles na coleção inteira está descrita entre parênteses. Ao se somar a frequência de todos os conceitos nominais (3+12+3+1+9+2+3=33), a sentença do Quadro 15 obteve a pontuação 33.

Quadro 15 – Exemplo da frequência simples dos conceitos nominais.

As **imagens<3931044>(3)**, divulgadas pela Infraero, mostram que o **avião<2691156>(12)** da TAM levou três **segundos<15235126>(3)** para fazer o **trajeto<8616311>(1)** na **pista<8651247>(9)** que, em **condições<13920835>(2)**, normais, levariam 11 **segundos<15235126>(3)**. **(=33)**

Para pontuar as sentenças segundo o método LCFSummN-pond, atribuíram-se pesos diferentes às sentenças conforme sua localização no texto-fonte. A primeira sentença de um texto recebeu peso 3, as intermediárias receberam peso 2 e a última recebeu peso 1. Assim, multiplicou-se a soma da frequência de ocorrência dos conceitos nominais de uma

sentença S pelo peso atribuído a S (3, 2 ou 1) e dividiu-se o resultado pelo número de conceitos distintos que compõem S. No caso da sentença do Quadro 15, multiplicou-se o valor 33 pelo peso 2, posto que se trata de uma sentença intermediária em seu texto-fonte, o que resultou em 66. Esse resultado foi dividido pelo número de conceitos distintos que compõem a sentença, ou seja, 6 (<3931044>, <2691156>, <15235126>, <8616311>, <8651247> e <13920835>). Dessa forma, a média ponderada da sentença é 11.

Com base na pontuação obtida pelas sentenças em função de cada método, construiu-se um ranque de sentenças, cujo topo é ocupado pela sentença de maior pontuação, as quais tendem a veicular o conteúdo central da coleção. No Quadro 16, apresenta-se o ranque das sentenças da coleção C3 construído com base na frequência simples dos conceitos nominais, estratégia que constitui o método LCFSummN. Nesse Quadro, cada sentença está especificada pelo documento de origem na coleção e pela posição ocupada por ela no mesmo (p.ex.: D3_S4).

Quadro 16 – Ranque sentencial de C3 pela frequência dos conceitos nominais

Ranque	**Sentenças de C3**	**Pontuação**
1º	Uma câmera da Rede Globo posicionada estrategicamente em frente ao Palácio do Planalto, pouco antes do início do Jornal Nacional, captou imagens do assessor especial da Presidência para Assuntos Internacionais, Marco Aurélio Garcia, e seu assessor de imprensa, Bruno Gaspar, comemorando com gestos obscenos a notícia de que o reverso direito do avião da TAM - que explodiu na última terça-feira contra um prédio da companhia aérea provocando a morte de pelo menos 198 pessoas - estava com defeito. (D3_S4)	86
2º	A falha no reversor --mecanismo que ajuda o avião a frear-- foi detectada pelo sistema eletrônico de checagem da própria aeronave, que continuou voando nos dias seguintes, com o reversor direito desligado. (D1_S5)	68
3º	Caso a possibilidade de que uma falha na aeronave tenha provocado a tragédia com o Airbus, as suspeitas de que o acidente esteja relacionado a falhas na recém-reformada pista do aeroporto perdem força. (D1_S15)	62
4º	A informação de que o Airbus A320 da TAM prefixo PR-MBK apresentava um defeito no reverso - peça acionada durante o pouso para inverter o sentido da propulsão das turbinas - direito foi antecipada pela jornalista Hildegard Angel, em sua coluna da edição de ontem do JB e confirmada pelo vice-presidente técnico da companhia, Ruy Amparo. (D3_S15)	56
5º	De acordo com a empresa aérea, o reversor estava travado, mas argumentou que a aeronave tinha condições de pouso normais, mesmo sem ele. (D3_S2)	55

Quadro 16 – Ranque sentencial de C3 pela frequência dos conceitos nominais (continuação).

6º	O reverso é um instrumento auxiliar na hora de frear o avião, mas não é considerado fator importante para o pouso, portanto, mesmo não funcionando, ele poderia ter brecado a aeronave sem problemas, diz. (D2_S7)	51
6º	De acordo com a companhia aérea, a recomendação da Airbus --fabricante do avião-- é que a revisão no reversor seja feita até dez dias depois de o defeito ser detectado. (D1_S11)	51
7º	Um dia antes do acidente, na segunda-feira, 16, o avião também teria apresentado problemas ao aterrissar em Congonhas, durante o vôo 3215, procedente de Belo Horizonte (Confins), só conseguindo parar muito próximo do final da pista. (D2_S8)	49
8º	O problema teria sido detectado pelo sistema eletrônico de checagem do próprio avião, e ainda assim a aeronave da TAM, um Airbus A320, continuou voando, com o reverso direito desligado. (D3_S5)	45
9º	A notícia, que retirou dos ombros do governo parte da suspeita de culpa sobre o acidente, foi comemorada no Palácio com efusão singular por Marco Aurélio e seu assessor, que fizeram gestos que seriam comuns a uma torcida de futebol, mas que soam inadequados ao momento. (D3_S6)	41
9º	Às 18h50 da última terça, sem controle, a aeronave --que havia decolado de Porto Alegre-- passou pela pista de Congonhas (zona sul de São Paulo) com velocidade acima do normal, atravessou a movimentada avenida Washington Luís e atingiu um prédio da própria empresa --da TAM Express. (D1_S7)	41
10º	Em 1996, uma falha no reverso foi a causa do acidente com o Fokker-100 da TAM, ocorrido segundos depois da decolagem, também em Congonhas. (D1_S13)	40
10º	A fumaça, de acordo com a reportagem, pode indicar que os motores estavam funcionando em sentidos opostos, um impulsionando para frente e outro freando, o que explicaria, por exemplo, por que o piloto não conseguiu parar o avião, que continuou em velocidade bem alta depois de tocar o solo e girou para a esquerda ao final da pista, em vez de seguir reto. (D1_S25)*	40
11º	Ainda segundo o "Jornal Nacional", o mesmo avião teve problemas para pousar, um dia antes do acidente, também em Congonhas. (D1_S17)	33
11º	As imagens, divulgadas pela Infraero, mostram que o avião da TAM levou três segundos para fazer o trajeto na pista que, em condições normais, levariam 11 segundos. (D1_S22)	33
12º	O Airbus-A320 da TAM acidentado em São Paulo na última terça-feira (17) tinha um defeito no reverso da turbina direita desde o último dia 13. (D1_S1)	32

Quadro 16 – Ranque sentencial de C3 pela frequência dos conceitos nominais (continuação).

12º	Segundo o perito aposentado do Departamento de Segurança de Vôo, Roberto Peterka, mesmo com problemas no reverso, a aeronave teria pousado tranqüilamente. (D2_S2)	32
13º	Na quarta-feira, o presidente da empresa, Marco Antonio Bologna, havia declarado que o avião passara por checagem justamente no dia 13 e estava em perfeito estado. (D3_S16)*	31
14º	Em nota enviada após a exibição da reportagem, a TAM afirma "que não teve registro de qualquer problema mecânico neste avião no dia 16 de julho". (D1_S19)	30
15º	O especialista também afirma que o reverso, apontado como possível causa, não pode ser considerado o vilão do acidente. (D2_S6)	29
16º	Além dos ocupantes, o acidente vitimou pessoas em terra, entre elas funcionários da empresa de transporte de cargas. (D1_S8)	28
17º	Imagens gravadas no aeroporto de Congonhas (zona sul de São Paulo) e divulgadas na quarta-feira (18) mostram o momento do pouso do vôo 3054 da TAM. (D1_S20)	26
18º	De acordo com Peterka, o fato de Congonhas ter uma pista menor não significa uma causa direta para o acidente. (D2_S4)	24
19º	Segundo o brigadeiro, a fumaça introduz um elemento novo no início das investigações: a possibilidade de falha mecânica no equipamento. (D1_S24)	23
19º	Esta notícia acabou criando uma polêmica em torno da reação de membros do governo ao tomarem conhecimento da informação através da imprensa. (D3_S3)	23
19º	Foi um gesto de raiva e indignação com o comportamento de certo tipo de noticiário da imprensa diante da tragédia de 200 famílias , que acabou tirando conclusões precipitadas sobre o fato - tentou explicar o assessor especial da Presidência da República. (D3_S14)	23
20º	Reportagem da colunista Eliane Cantanhê de (íntegra disponível só para assinantes do jornal ou do UOL), publicada na edição da Folha desta quinta, mostra que nova versão da Infraero inclui falha mecânica entre as hipóteses para o acidente --o maior da aviação brasileira. (D1_S3)	22
20º	A TAM confirmou , na noite desta quinta-feira , que ao airbus da TAM estava com o reverso do lado direito desligado, desde o último dia 13. (D3_S1)	22
21º	A informação, revelada na edição desta quinta do "Jornal Nacional", da TV Globo, foi confirmada , em nota , pela companhia aérea. (D1_S2)	21
22º	As obras foram entregues incompletas, sem o chamado "grooving" (ranhuras que ajudam no escoamento da água) --que reduz o risco de derrapagens de aviões em casos de chuva . (D1_S16)	20
23º	O presidente da Infraero (estatal que administra os aeroportos do país), brigadeiro José Carlos Pereira, disse ontem que os peritos detectaram "fumaça forte" no motor esquerdo do Airbus-A320 da TAM no filme , de acordo com a reportagem publicada pela Folha. (D1_S23)	19
24º	A TAM afirma que "o procedimento não configura qualquer obstáculo ao pouso da aeronave ". (D1_S6)	16

Quadro 16 – Ranque sentencial de C3 pela frequência dos conceitos nominais (continuação).

25º	Para a TAM, a falha não impediria a realização dos vôos. (D1_S12)	15
26º	"Isso não quer dizer que o aeroporto de Congonhas teve influência decisiva no acidente ", conclui. (D2_S5)	15
26º	O piloto teria relatado à torre de controle que a pista estava muito escorregadia. (D2_S9)	15
26º	A câmera da emissora teria registrado a cena logo depois da transmissão da notícia , o que sugere uma comemoração festiva da matéria veiculada na TV. (D3_S9)	15
27º	O vôo, que havia saído de Minas conseguiu parar apenas no limite da pista. (D1_S18)	14
27º	"Por ser a pista maior, o piloto não precisaria ter que arremeter, se realmente foi isso que ele fez , e teria espaço maior para a desaceleração ", explica. (D2_S3)	14
27º	Seu assessor puxava os braços ao lado do corpo simulando o gesto de copular. (D3_S8)	14
28º	Os vídeos comparam pousos de outras aeronaves com o do Airbus A-320 acidentado. (D1_S21)	12
29º	Procurado pelo JB, Marco Aurélio se disse incapaz de relembrar todo o seu gestual em um dia inteiro. (D3_S10)	10
30º	Até a noite desta quinta, os bombeiros confirmavam a morte de 188 pessoas. (D1_S9)	9
31º	Na ocasião, 99 pessoas morreram. (D1_S14)	5
31º	Para especialistas, a tragédia do Vôo 3054 não teria acontecido caso o Airbus da TAM tivesse pousando em Cumbica. (D2_S1)	5
32º	A TAM negou a hipótese. (D1_S4)	4
32º	Também disse não saber em que momento foi filmado. (D3_S11)	4
33º	O número pode chegar a 200. (D1_S10)	1
33º	Garcia bateu com uma das mãos espalmada na outra cerrada, como se dissesse "se ferrou ". (D3_S7)	1
34º	Mas contemporizou. (D3_S12)	0
34º	- Certamente não foi regozijo. (D3_S13)	0

No Quadro 17, tem-se o ranque das sentenças de C3 produzido com base na média ponderada dos conceitos nominais, que caracteriza o método LCFSummN-pond.

Quadro 17 – Ranque sentencial de C3 pela média ponderada dos conceitos nominais.

Ranque	**Sentenças de C3**	**Pontuação**
1º	De acordo com a empresa aérea, o reversor estava travado, mas argumentou que a aeronave tinha condições de pouso normais, mesmo sem ele. (D3_S2)	22
2º	O Airbus-A320 da TAM acidentado em São Paulo na última terça-feira (17) tinha um defeito no reverso da turbina direita desde o último dia 13. (D1_S1)	19,2

Quadro 17 – Ranque sentencial de C3 pela média ponderada dos conceitos nominais (continuação).

3º	A falha no reversor --mecanismo que ajuda o avião a frear-- foi detectada pelo sistema eletrônico de checagem da própria aeronave, que continuou voando nos dias seguintes, com o reversor direito desligado. (D1_S5)	17
4º	Ainda segundo o "Jornal Nacional", o mesmo avião teve problemas para pousar, um dia antes do acidente, também em Congonhas. (D1_S17)	16,5
5º	Caso a possibilidade de que uma falha na aeronave tenha provocado a tragédia com o Airbus, as suspeitas de que o acidente esteja relacionado a falhas na recém-reformada pista do aeroporto perdem força. (D1_S15)	15,5
6º	O problema teria sido detectado pelo sistema eletrônico de checagem do próprio avião, e ainda assim a aeronave da TAM, um Airbus A320, continuou voando, com o reverso direito desligado. (D3_S5)	15
7º	Em 1996, uma falha no reverso foi a causa do acidente com o Fokker-100 da TAM, ocorrido segundos depois da decolagem, também em Congonhas. (D1_S13)	13,33333333
8º	A TAM confirmou, na noite desta quinta-feira, que ao airbus da TAM estava com o reverso do lado direito desligado, desde o último dia 13. (D3_S1)	13,2
9º	O reverso é um instrumento auxiliar na hora de frear o avião, mas não é considerado fator importante para o pouso, portanto, mesmo não funcionando, ele poderia ter brecado a aeronave sem problemas, diz. (D2_S7)	12,75
9º	De acordo com a companhia aérea, a recomendação da Airbus --fabricante do avião-- é que a revisão no reversor seja feita até dez dias depois de o defeito ser detectado. (D1_S11)	12,75
10º	Um dia antes do acidente, na segunda-feira, 16, o avião também teria apresentado problemas ao aterrissar em Congonhas, durante o vôo 3215, procedente de Belo Horizonte (Confins), só conseguindo parar muito próximo do final da pista. (D2_S8)	12,25
11º	De acordo com Peterka, o fato de Congonhas ter uma pista menor não significa uma causa direta para o acidente. (D2_S4)	12
11º	Os vídeos comparam pousos de outras aeronaves com o do Airbus A-320 acidentado. (D1_S21)	12
12º	O especialista também afirma que o reverso, apontado como possível causa, não pode ser considerado o vilão do acidente. (D2_S6)	11,6
13º	Uma câmera da Rede Globo posicionada estrategicamente em frente ao Palácio do Planalto, pouco antes do início do Jornal Nacional, captou imagens do assessor especial da Presidência para Assuntos Internacionais, Marco Aurélio Garcia, e seu assessor de imprensa, Bruno Gaspar, comemorando com gestos obscenos a notícia de que o reverso direito do avião da TAM - que explodiu na última terça-feira contra um prédio da companhia aérea provocando a morte de pelo menos 198 pessoas - estava com defeito. (D3_S4)	11,46666667
14º	As imagens, divulgadas pela Infraero, mostram que o avião da TAM levou três segundos para fazer o trajeto na pista que, em condições normais, levariam 11 segundos. (D1_S22)	11

Quadro 17 – Ranque sentencial de C3 pela média ponderada dos conceitos nominais (continuação).

15º	Segundo o perito aposentado do Departamento de Segurança de Vôo, Roberto Peterka, mesmo com problemas no reverso, a aeronave teria pousado tranqüilamente. (D2_S2)	10,66666667
16º	A informação, revelada na edição desta quinta do "Jornal Nacional", da TV Globo, foi confirmada, em nota, pela companhia aérea. (D1_S2)	10,5
17º	Isso não quer dizer que o aeroporto de Congonhas teve influência decisiva no acidente, conclui. (D2_S5)	10
17º	Procurado pelo JB, Marco Aurélio se disse incapaz de relembrar todo o seu gestual em um dia inteiro. (D3_S10)	10
17º	Para a TAM, a falha não impediria a realização dos vôos. (D1_S12)	10
18º	O vôo, que havia saído de Minas conseguiu parar apenas no limite da pista. (D1_S18)	9,333333333
19º	A TAM negou a hipótese. (D1_S4)	8
19º	A TAM afirma que "o procedimento não configura qualquer obstáculo ao pouso da aeronave". (D1_S6)	8
19º	Também disse não saber em que momento foi filmado. (D3_S11)	8
19º	A informação de que o Airbus A320 da TAM prefixo PR-MBK apresentava um defeito no reverso - peça acionada durante o pouso para inverter o sentido da propulsão das turbinas - direito foi antecipada pela jornalista Hildegard Angel, em sua coluna da edição de ontem do JB e confirmada pelo vice-presidente técnico da companhia, Ruy Amparo. (D3_S15)	8
20º	Esta notícia acabou criando uma polêmica em torno da reação de membros do governo ao tomarem conhecimento da informação através da imprensa. (D3_S3)	7,666666667
21º	Em nota enviada após a exibição da reportagem, a TAM afirma "que não teve registro de qualquer problema mecânico neste avião no dia 16 de julho". (D1_S19)	7,5
21º	Para especialistas, a tragédia do Vôo 3054 não teria acontecido caso o Airbus da TAM tivesse pousando em Cumbica. (D2_S1)	7,5
22º	Às 18h50 da última terça, sem controle, a aeronave --que havia decolado de Porto Alegre-- passou pela pista de Congonhas (zona sul de São Paulo) com velocidade acima do normal, atravessou a movimentada avenida Washington Luís e atingiu um prédio da própria empresa --da TAM Express. (D1_S7)	7,454545455
23º	Por ser a pista maior, o piloto não precisaria ter que arremeter, se realmente foi isso que ele fez, e teria espaço maior para a desaceleração, explica. (D2_S3)	7
23º	Além dos ocupantes, o acidente vitimou pessoas em terra, entre elas funcionários da empresa de transporte de cargas. (D1_S8)	7
24º	A notícia, que retirou dos ombros do governo parte da suspeita de culpa sobre o acidente, foi comemorada no Palácio com efusão singular por Marco Aurélio e seu assessor, que fizeram gestos que seriam comuns a uma torcida de futebol, mas que soam inadequados ao momento. (D3_S6)	6,833333333

Quadro 17 – Ranque sentencial de C3 pela média ponderada dos conceitos nominais (continuação).

25º	Foi um gesto de raiva e indignação com o comportamento de certo tipo de noticiário da imprensa diante da tragédia de 200 famílias, que acabou tirando conclusões precipitadas sobre o fato - tentou explicar o assessor especial da Presidência da República. (D3_S14)	6,571428571
26º	Imagens gravadas no aeroporto de Congonhas (zona sul de São Paulo) e divulgadas na quarta-feira (18) mostram o momento do pouso do vôo 3054 da TAM. (D1_S20)	6,5
27º	Reportagem da colunista Eliane Cantanhêde (íntegra disponível só para assinantes do jornal ou do UOL), publicada na edição da Folha desta quinta, mostra que nova versão da Infraero inclui falha mecânica entre as hipóteses para o acidente --o maior da aviação brasileira. (D1_S3)	6,285714286
28º	Segundo o brigadeiro, a fumaça introduz um elemento novo no início das investigações: a possibilidade de falha mecânica no equipamento. (D1_S24)	5,75
29º	Seu assessor puxava os braços ao lado do corpo simulando o gesto de copular. (D3_S8)	5,6
30º	Na quarta-feira, o presidente da empresa, Marco Antonio Bologna, havia declarado que o avião passara por checagem justamente no dia 13 e estava em perfeito estado. (D3_S16)*	5,166666667
31º	Na ocasião, 99 pessoas morreram. (D1_S14)	5
32º	O presidente da Infraero (estatal que administra os aeroportos do país), brigadeiro José Carlos Pereira, disse ontem que os peritos detectaram "fumaça forte" no motor esquerdo do Airbus-A320 da TAM no filme, de acordo com a reportagem publicada pela Folha. (D1_S23)	4,75
33º	Até a noite desta quinta, os bombeiros confirmavam a morte de 188 pessoas. (D1_S9)	4,5
34º	As obras foram entregues incompletas, sem o chamado "grooving" (ranhuras que ajudam no escoamento da água) --que reduz o risco de derrapagens de aviões em casos de chuva. (D1_S16)	4,444444444
35º	A câmera da emissora teria registrado a cena logo depois da transmissão da notícia, o que sugere uma comemoração festiva da matéria veiculada na TV. (D3_S9)	4,285714286
36º	O piloto teria relatado à torre de controle que a pista estava muito escorregadia. (D2_S9)*	3,75
37º	A fumaça, de acordo com a reportagem, pode indicar que os motores estavam funcionando em sentidos opostos, um impulsionando para frente e outro freando, o que explicaria, por exemplo, por que o piloto não conseguiu parar o avião, que continuou em velocidade bem alta depois de tocar o solo e girou para a esquerda ao final da pista, em vez de seguir reto. (D1_S25)*	3,333333333
38º	Garcia bateu com uma das mãos espalmada na outra cerrada, como se dissesse "se ferrou". (D3_S7)	2
38º	O número pode chegar a 200. (D1_S10)	2
39º	Mas contemporizou. (D3_S12)	0
39º	Certamente não foi regozijo. (D3_S13)	0

4.1.2 A seleção das sentenças para o extrato

Diante do ranque das sentenças, inicia-se a seleção das sentenças para o sumário.

Utilizando como exemplo o *cluster* C3 e o ranque do Quadro 16, selecionou-se inicialmente a 1ª sentença do ranque para compor o sumario, isto é, "*Uma câmera da Rede Globo posicionada estrategicamente em frente ao Palácio do Planalto, pouco antes do início do Jornal Nacional, captou imagens do assessor especial da Presidência para Assuntos Internacionais, Marco Aurélio Garcia, e seu assessor de imprensa, Bruno Gaspar, comemorando com gestos obscenos a notícia de que o reverso direito do avião da TAM - que explodiu na última terça-feira contra um prédio da companhia aérea provocando a morte de pelo menos 198 pessoas - estava com defeito*". Essa sentença possui 79 palavras.

Como a taxa de compressão de 70% não foi atingida (isto é, 182 palavras), selecionou-se a 2ª sentença do ranque ("*A falha no reversor --mecanismo que ajuda o avião a frear-- foi detectada pelo sistema eletrônico de checagem da própria aeronave, que continuou voando nos dias seguintes, com o reversor direito desligado*"), que possui 32 palavras. Seguindo o algoritmo (Quadro 13), calculou-se, então, a redundância entre a 2ª sentença e a 1ª já selecionada por meio da medida *concept overlap* (cf. Seção 4). Para tanto, calculou-se o número de conceitos idênticos entre a sentença já selecionada para o sumário e a segunda sentença, que foi 5. Esse número foi dividido pela soma do número total de conceitos das duas sentenças, que foi 34. O resultado obtido foi 0.147059. Já que a medida *concept overlap* determina que quanto mais próximo de 1 for o resultado da fórmula, mais redundante será o par entre si, e, quanto mais próximo de 0, menos redundante, o limiar adotado empiricamente neste trabalho para se considerar que as sentenças não sejam redundantes entre si, foi o de que o resultado da medida *concept overlap* não deve ultrapassar 0.4, pois essa margem garante o menor número de informações idênticas entre as sentenças. Assim, verificou-se que não há redundância entre as sentenças em questão, sendo a 2ª sentença selecionada a compor o extrato.

Como a taxa de compressão ainda não tinha sido atingida (79+32=111), selecionou-se a 3ª sentença do ranque ("*Caso a possibilidade de que uma falha na aeronave tenha provocado a tragédia com o Airbus, as suspeitas de que o acidente esteja relacionado a falhas na recém-reformada pista do aeroporto perdem força*"), que possui 33 palavras e não apresenta redundância frente às duas sentenças anteriores já selecionadas. Como o tamanho do sumário (182) ainda não tinha sido atingido (79+32+33=144), selecionou-se também a 4ª sentença do ranque ("*A informação de que*

o Airbus A320 da TAM prefixo PR-MBK apresentava um defeito no reverso - peça acionada durante o pouso para inverter o sentido da propulsão das turbinas - direito foi antecipada pela jornalista Hildegard Angel, em sua coluna da edição de ontem do JB e confirmada pelo vice-presidente técnico da companhia, Ruy Amparo"), que possui 56 palavras e não apresenta redundância com as demais selecionadas. A inclusão dessa sentença geraria um extrato com 200 palavras, ultrapassando, assim, o tamanho desejado (182) em 18 palavras e a não-inclusão geraria um extrato com 38 palavras a menos. Como a não-inclusão se distanciaria mais do tamanho desejado do que a sua inclusão, optou-se pela inclusão da sentença. Uma vez que a taxa de compressão tenha sido atingida, procedeu-se à ordenação das sentenças para a geração dos extratos propriamente ditos.

4.1.3 A ordenação das sentenças e a geração dos extratos

Objetivando produzir sumários coerentes e coesos, as sentenças selecionadas foram justapostas com base no método de Ordenação da Posição Textual denominado OPT (LIMA, PARDO, 2012).

Segundo tal método, as sentenças selecionadas para compor o sumário são justapostas na ordem em que aparecem nos textos-fonte, isto é, se a primeira sentença selecionada para compor o sumário ocupa posição 7 em seu texto-fonte e a segunda sentença selecionada ocupa posição 5, a sentença de posição 5 em seu respectivo texto-fonte deverá ser a primeira sentença a compor o sumário, sendo seguida pela sentença de posição 7. Caso haja empate, ou seja, caso mais de uma sentença ocupe a mesma posição em seus textos-fonte, o critério de desempate é o seu tamanho em número de palavras, sendo que as sentenças menores devem ocupar posição anterior à sentença maior.

Na coleção C3, por exemplo, a primeira sentença selecionada para compor o sumário ocupa a posição 4 em seu respectivo texto-fonte (D3-S4). Já a segunda ocupa posição 5 em seu texto-fonte (D1-S5). Assim, com base no método OPT, a sentença D1_S5 foi ordenada na segunda posição no extrato. Como as duas últimas sentenças selecionadas ocupam a mesma posição em seus respectivos textos-fonte (D1_S15 e D3_S15), aplicou-se o critério de desempate e a sentença a ocupar a terceira posição no extrato foi a D1_S15, pois esta apresenta o menor número de palavras em comparação à D3_S15.

Nos Quadros 18 e 19, tem-se a ordenação final dos extratos da coleção C3 produzidos com base nos métodos LCSummN e LCSummN-pond, respectivamente.

Quadro 18 – Extrato de C3 produzido pelo método LCFSummN

Uma câmera da Rede Globo posicionada estrategicamente em frente ao Palácio do Planalto, pouco antes do início do Jornal Nacional, captou imagens do assessor especial da Presidência para Assuntos Internacionais, Marco Aurélio Garcia, e seu assessor de imprensa, Bruno Gaspar, comemorando com gestos obscenos a notícia de que o reverso direito do avião da TAM - que explodiu na última terça-feira contra um prédio da companhia aérea provocando a morte de pelo menos 198 pessoas - estava com defeito. (D3_S4)
A falha no reversor --mecanismo que ajuda o avião a frear-- foi detectada pelo sistema eletrônico de checagem da própria aeronave, que continuou voando nos dias seguintes, com o reversor direito desligado. (D1_S5)
Caso a possibilidade de que uma falha na aeronave tenha provocado a tragédia com o Airbus, as suspeitas de que o acidente esteja relacionado a falhas na recém-reformada pista do aeroporto perdem força. (D1_S15)
A informação de que o Airbus A320 da TAM prefixo PR-MBK apresentava um defeito no reverso - peça acionada durante o pouso para inverter o sentido da propulsão das turbinas - direito foi antecipada pela jornalista Hildegard Angel, em sua coluna da edição de ontem do JB e confirmada pelo vice-presidente técnico da companhia, Ruy Amparo. (D3_S15)

Quadro 19 – Extrato de C3 produzido pelo método LCFSummN-pond

O Airbus-A320 da TAM acidentado em São Paulo na última terça-feira (17) tinha um defeito no reverso da turbina direita desde o último dia 13. (D1_S1)
A TAM confirmou, na noite desta quinta-feira, que ao airbus da TAM estava com o reverso do lado direito desligado, desde o último dia 13. (D3_S1)
De acordo com a empresa aérea, o reversor estava travado, mas argumentou que a aeronave tinha condições de pouso normais, mesmo sem ele. (D3_S2)
A falha no reversor --mecanismo que ajuda o avião a frear-- foi detectada pelo sistema eletrônico de checagem da própria aeronave, que continuou voando nos dias seguintes, com o reversor direito desligado. (D1_S5)
Em 1996, uma falha no reverso foi a causa do acidente com o Fokker-100 da TAM, ocorrido segundos depois da decolagem, também em Congonhas. (D1_S13)
Caso a possibilidade de que uma falha na aeronave tenha provocado a tragédia com o Airbus, as suspeitas de que o acidente esteja relacionado a falhas na recém-reformada pista do aeroporto perdem força. (D1_S15)
Ainda segundo o "Jornal Nacional", o mesmo avião teve problemas para pousar, um dia antes do acidente, também em Congonhas. (D1_S17)

A título de comparação, nos Quadros 20 e 21, exibem-se também a ordenação final dos extratos da coleção C3 produzidos com base nos métodos LCSummN-V e LCSummN-V-pond, respectivamente.

Quadro 20 – Extrato de C3 produzido pelo método LCFSummN-V

De acordo com a empresa aérea, o reversor estava travado, mas argumentou que a aeronave tinha condições de pouso normais, mesmo sem ele. (D3_S2)
Uma câmera da Rede Globo posicionada estrategicamente em frente ao Palácio do Planalto, pouco antes do início do Jornal Nacional, captou imagens do assessor especial da Presidência para Assuntos Internacionais, Marco Aurélio Garcia, e seu assessor de imprensa, Bruno Gaspar, comemorando com gestos obscenos a notícia de que o reverso direito do avião da TAM - que explodiu na última terça-feira contra um prédio da companhia aérea provocando a morte de pelo menos 198 pessoas - estava com defeito. (D3_S4)
A falha no reversor --mecanismo que ajuda o avião a frear-- foi detectada pelo sistema eletrônico de checagem da própria aeronave, que continuou voando nos dias seguintes, com o reversor direito desligado. (D1_S5)
O reverso é um instrumento auxiliar na hora de frear o avião, mas não é considerado fator importante para o pouso, portanto, mesmo não funcionando, ele poderia ter brecado a aeronave sem problemas, diz. (D2_S7)
Caso a possibilidade de que uma falha na aeronave tenha provocado a tragédia com o Airbus, as suspeitas de que o acidente esteja relacionado a falhas na recém-reformada pista do aeroporto perdem força. (D1_S15)

Quadro 21 - Extrato de C3 produzido pelo método LCFSummN-V-pond

O Airbus-A320 da TAM acidentado em São Paulo na última terça-feira (17) tinha um defeito no reverso da turbina direita desde o último dia 13. (D1_S1)
A TAM confirmou, na noite desta quinta-feira, que ao airbus da TAM estava com o reverso do lado direito desligado, desde o último dia 13. (D3_S1)
De acordo com a empresa aérea, o reversor estava travado, mas argumentou que a aeronave tinha condições de pouso normais, mesmo sem ele. (D3_S2)
A falha no reversor --mecanismo que ajuda o avião a frear-- foi detectada pelo sistema eletrônico de checagem da própria aeronave, que continuou voando nos dias seguintes, com o reversor direito desligado. (D1_S5)
Em 1996, uma falha no reverso foi a causa do acidente com o Fokker-100 da TAM, ocorrido segundos depois da decolagem, também em Congonhas. (D1_S13)
Caso a possibilidade de que uma falha na aeronave tenha provocado a tragédia com o Airbus, as suspeitas de que o acidente esteja relacionado a falhas na recém-reformada pista do aeroporto perdem força. (D1_S15)
Ainda segundo o "Jornal Nacional", o mesmo avião teve problemas para pousar, um dia antes do acidente, também em Congonhas. (D1_S17)

Os extratos gerados de acordo com os 4 métodos para as 5 coleções do CSTNews anotadas estão no Apêndice 1 desta dissertação. Por fim, após a geração dos extratos, estes foram submetidos à avaliação intrínseca conforme descrito na sequência.

5. A AVALIAÇÃO DOS MÉTODOS

A avaliação dos extratos gerados pelos métodos propostos nesta investigação foi a intrínseca, visto que a avaliação extrínseca é demorada, cara e que requer planejamento cuidadoso (VAN-HALTEREN; TEUFEL, 2003). Dessa forma, avaliaram-se os extratos quanto à qualidade linguística e informatividade (MANI, 2001).

5.1 A avaliação intrínseca da qualidade linguística

A avaliação da qualidade linguística seguiu os parâmetros propostos pela TAC (DANG, 2008). Assim, avaliaram-se os extratos de forma manual quanto aos critérios de gramaticalidade, não-redundância, clareza referencial, foco e estrutura/coerência.

Os sumários gerados para as 5 coleções do *corpus* foram avaliados por 6 juízes, sendo 3 linguistas e 3 cientistas da computação; cada juiz avaliou 20 sumários. O procedimento adotado para essa tarefa pautou-se na confecção de formulários *online* por meio dos quais os juízes acessaram os sumários e submeteram o julgamento individualmente. Os juízes atribuíram pontuações em uma escala de 1 a 5 para cada atributo linguístico, conforme quadro abaixo.

Quadro 22 – Pontuações e níveis para a avaliação da qualidade linguística

Pontuação	**Nível**
1	Péssimo
2	Ruim
3	Regular
4	Bom
5	Excelente

Tendo em vista os 4 métodos investigados, geraram-se 4 extratos para cada uma das 5 coleções do *corpus*, os quais foram analisados de acordo com os 5 critérios linguísticos. Os resultados da avaliação estão apresentados nas Tabelas 3, 4, 5, e 6, em termos absolutos e porcentagens.

Na Tabela 3, em que se exibe a avaliação da gramaticalidade, observa-se que o método LCFSummN recebeu (i) 18 vezes a classificação “excelente” (isto é, 60% do total), (ii) 10 vezes “bom” (isto é, 33.3% do total), (iii) 2 vezes “regular” (isto é, 6,7% do total) e (iv) 0 vezes as classificações “ruim” e “péssimo”.

Tabela 3 – Pontuações das versões do método: critério de “gramaticalidade”

Gramaticalidade											
Método	Péssimo (1)		Ruim (2)		Regular (3)		Bom (4)		Excelente (5)		Média
LCFSummN	0	0%	0	0%	2	6.7%	10	33.3%	18	60%	4.5 (excelente)
LCFSummN-V	0	0%	0	0%	1	3.3%	9	30%	20	66.7%	4.6 (excelente)
LCFSummN-pond	0	0%	0	0%	2	6.7%	9	30%	19	63.3%	4.6 (excelente)
LCFSummN-V-pond	0	0%	0	0%	2	6.7%	10	33.3%	18	60%	4.5 (excelente)

Na média, a gramaticalidade dos extratos gerados pelos métodos obteve nível “excelente”, com pontuação 4,6 para LCFSummN-V e LCFSummN-pond, e pontuação 4,5 para LCFSummN e LCFSummN-V-pond. A partir disso, pode-se inferir que os extratos gerados apresentam poucos problemas de ortografia, pontuação e sintaxe.

Tabela 4 – Pontuações das versões do método: critério de “não-redundância”

Não-redundância											
Método	Péssimo (1)		Ruim (2)		Regular (3)		Bom (4)		Excelente (5)		Média
LCFSummN	0	0%	2	6.7%	0	0%	9	30%	19	63.3%	4.5 (excelente)
LCFSummN-V	0	0%	1	3.3%	2	6.7%	10	33.3%	17	56.7%	4.4 (bom)
LCFSummN-pond	1	3.3%	2	6.7%	4	13.3%	6	20%	17	56.7%	4.2 (bom)
LCFSummN-V-pond	1	3.3%	4	13.3%	4	13.3%	6	20%	15	50%	4.0 (bom)

Com base na Tabela 4, observa-se que o atributo não-redundância dos extratos gerados pelos métodos LCFSummN-V, LCFSummN-pond e LCFSummN-V-pond foi classificado na média como “bom”. Ainda quanto aos dados da Tabela 4, destaca-se que o método LCFSummN foi o único a gerar extratos cuja não-redundância foi em média classificada como “excelente”.

Tabela 5 – Pontuações das versões do método: critério de “clareza referencial”

Clareza referencial											
Método	Péssimo (1)		Ruim (2)		Regular (3)		Bom (4)		Excelente (5)		Média
LCFSummN	0	0%	1	3.3%	7	23.3%	6	20%	16	53.3%	4.2 (bom)
LCFSummN-V	0	0%	3	10%	5	16.7%	12	40%	10	33.3%	4.0 (bom)
LCFSummN-pond	0	0%	2	6.7%	5	16.7%	9	30%	14	46.7%	4.2 (bom)
LCFSummN-V-pond	0	0%	3	10%	8	26.7%	11	36.7%	8	26.7%	3.8 (bom)

Quanto à clareza referencial, observa-se, a partir dos dados da Tabela 5, que esse atributo nos extratos gerados pelos 4 métodos propostos neste trabalho foi em média classificado como “bom”. O nível “bom” obtido pelos métodos léxico-conceituais pode indicar que a aplicação, na fase de síntese, do método OPT (LIMA, PARDO, 2012) de ordenação de sentenças contribui para a produção de extratos com boa clareza referencial sem que se faça uso de um método específico de correferenciação.

Tabela 6 – Pontuações das versões do método: critério de “Foco”

Foco											
Método	Péssimo (1)		Ruim (2)		Regular (3)		Bom (4)		Excelente (5)		Média
LCFSummN	0	0%	0	0%	5	16.7%	7	23.3%	18	60%	4.4 (bom)
LCFSummN-V	0	0%	3	10%	3	10%	9	30%	15	50%	4.2 (bom)
LCFSummN-pond	0	0%	2	6.7%	4	13.3%	10	33.3%	14	46.7%	4.2 (bom)
LCFSummN-V-pond	0	0%	4	13.3%	9	30%	8	26.7%	9	30%	3.7 (bom)

O atributo foco foi avaliado em média como “bom” nos extratos gerados pelos 4 métodos. Assim, pode-se dizer que, no geral, os extratos contêm informações inter-relacionadas, isto é, sentenças que possuem certo grau de relação entre si a ponto de garantir um sentido ao todo textual. Embora o foco dos extratos gerados pelo método LCFSummN-V-pond tenha sido classificado como “bom”, esse foi o único método que obteve pontuação média abaixo de 4.0 (3.7).

Tabela 7 – Pontuações das versões do método: critério de “estrutura/coerência”

Estrutura/coerência											
Método	Péssimo (1)		Ruim (2)		Regular (3)		Bom (4)		Excelente (5)		Média
LCFSummN	0	0%	2	6.7%	7	23.3%	9	30%	12	40%	4.0 (bom)
LCFSummN-V	0	0%	3	10%	5	16.7%	11	36.7%	11	36.7%	4.0 (bom)
LCFSummN-pond	1	3.3%	5	16.7%	5	16.7%	11	36.7%	8	26.7%	3.7 (bom)
LCFSummN-V-pond	3	10%	6	20%	7	23.3%	7	23.3%	7	23.3%	3.3 (regular)

Quanto ao atributo estrutura/coerência a pontuação média dos extratos gerados pelos métodos LCFSummN, LCFSummN-V e LCFSummN-pond foi 4.0, 4.0 e 3.7, respectivamente, as quais equivalem ao nível “bom”. O método LCFSummN-V-pond foi o único a gerar extratos com estrutura/coerência considerada “regular”. Com base nos resultados da Tabela 7, pode-se dizer que a maioria dos métodos léxico-conceituais, com exceção do LCFSummN-V-pond, geram extratos multidocumento com razoável nível de coerência, o que é significativo em se tratando de métodos extrativos.

Sobre a comparação dos métodos léxico-conceituais entre si descrita na Tabela 8, ressalta-se os métodos baseados em conhecimento léxico-conceitual geram, no geral, extratos com “boa” qualidade linguística. Além disso, observa-se na Tabela 8 que o LCFSummN obteve as melhores médias na maioria dos critérios da TAC. Isso indica que os conceitos nominais, sendo os mais frequentes nas coleções, veiculam de fato as informações mais relevantes. Frente ao método LCFSummN, a inclusão dos conceitos verbais aos nominais no método LCFSummN-V melhorou apenas a gramaticalidade dos extratos.

Tabela 8 – Comparação entre os métodos LCFSumm

Critérios	LCFSummN	LCFSummN-V	LCFSummN-pond	LCFSummN-V-pond
Gramaticalidade	4.5	**4.6**	**4.6**	4.5
Não-redundância	**4.5**	4.4	4.2	4.0
Clareza referencial	**4.2**	4.0	4.2	3.8
Foco	**4.4**	4.2	4.2	3.7
Estrutura/Coerência	**4.0**	**4.0**	3.7	3.3

5.2 A avaliação intrínseca da informatividade

A informatividade dos extratos gerados pelos métodos léxico-conceituais foi avaliada automaticamente por meio do pacote de medidas ROUGE (LIN, 2004) (cf. Seção 2.1). No caso, utilizou-se a interface *online*[31] NILC-WISE (NILC - *Web Interface for Summary Evaluation*) (NÓBREGA, PARDO, 2016) para avaliação de sumários automáticos[32]. Neste trabalho, utilizaram-se duas medidas ROUGE-1 e ROUGE-2. A ROUGE-1 calcula a informatividade pela sobreposição de unigramas entre o extrato automático e o de referência, enquanto a ROUGE-2 faz o cálculo com base na sobreposição de bigramas. Os resultados são fornecidos em termos de precisão, cobertura e medida-f. Para a aplicação da ROUGE, os extratos gerados por cada um dos métodos léxico-conceituais para cada um dos *clusters* C1, C2, C3, C4 e C5 foram automaticamente comparados aos 6 sumários humanos (*abstracts*) disponíveis no CSTNews para esses *clusters*. Os resultados da ROUGE obtidos por cada um dos métodos léxico-conceituais estão disponíveis nas Tabelas 9, 10, 11 e 12.

Tabela 9 – Avaliação ROUGE: método LCFSummN

Coleção	ROUGE-1			ROUGE-2		
	Cobertura	Precisão	Medida-F	Cobertura	Precisão	Medida-F
C1	0.40851	0.41026	0.40938	0.17467	0.17544	0.17505
C2	0.49153	0.48876	0.49014	0.24952	0.24811	0.24881
C3	0.41586	0.36375	0.38806	0.17952	0.15686	0.16743
C4	0.29333	0.27500	0.28387	0.15541	0.14557	0.15033
C5	0.46618	0.50000	0.48250	0.27357	0.29365	0.28325

Tabela 10 – Avaliação ROUGE: método LCFSummN-pond

Coleção	ROUGE-1			ROUGE-2		
	Cobertura	Precisão	Medida-F	Cobertura	Precisão	Medida-F
C1	0.35319	0.35470	0.35394	0.16157	0.16228	0.16192
C2	0.58004	0.52381	0.55049	0.32000	0.28866	0.30352
C3	0.42698	0.42287	0.42492	0.16410	0.16250	0.16330
C4	0.33333	0.34722	0.34013	0.17342	0.18075	0.17701
C5	0.47532	0.47101	0.47316	0.24584	0.24359	0.24471

[31] http://nilc.icmc.usp.br/nilcwise

[32] Para auxiliar os próximos usuários de tal interface, sobretudo os linguistas, produziu-se um manual de utilização, o qual está descrito no Apêndice 2.

Tabela 11 – Avaliação ROUGE: método LCFSummN-V

Coleção	ROUGE-1			ROUGE-2		
	Cobertura	Precisão	Medida-F	Cobertura	Precisão	Medida-F
C1	0.40851	0.41026	0.40938	0.17467	0.17544	0.17505
C2	0.51601	0.46599	0.48973	0.27619	0.24914	0.26197
C3	0.46732	0.42105	0.44298	0.19215	0.17298	0.18206
C4	0.34222	0.34685	0.34452	0.18468	0.18721	0.18594
C5	0.39305	0.42157	0.40681	0.21811	0.23413	0.22584

Tabela 12 – Avaliação ROUGE: método LCFSummN-V-pond

Coleção	ROUGE-1			ROUGE-2		
	Cobertura	Precisão	Medida-F	Cobertura	Precisão	Medida-F
C1	0.35319	0.34583	0.34947	0.16157	0.15812	0.15983
C2	0.52919	0.45915	0.49169	0.25905	0.22442	0.24049
C3	0.42698	0.42287	0.42492	0.16410	0.16250	0.16330
C4	0.40000	0.37037	0.38462	0.20721	0.19167	0.19914
C5	0.39488	0.40909	0.40186	0.20702	0.21456	0.21072

Na Tabela 13 a seguir, descreve-se a média das pontuações de cada medida ROUGE para cada método, assim como a média total do método para cobertura, precisão e medida-f.

Tabela 13 – Médias das avaliações ROUGE para cada método

Método	ROUGE-1			ROUGE-2		
	Cobertura	Precisão	Medida-F	Cobertura	Precisão	Medida-F
LCFSummN	0.415082	0.407554	0.41079	0.20653	0.20392	0.20497
LCFSummN-pond	**0.433772**	**0.423922**	**0.42852**	**0.21298**	**0.20755**	**0.21009**
LCFSummN-V	0.425422	0.413144	0.41868	0.20916	0.20378	0.20617
LCFSummN-V-pond	0.420848	0.401462	0.41051	0.19979	0.19025	0.19469

Com base nas médias apresentadas na Tabela 13, destaca-se que o método LCFSummN-pond foi o que obteve resultados superiores para todas as medidas (cobertura, precisão e medida-f) de ROUGE-1 e ROUGE-2. Em outras palavras, esses resultados indicam que o referido método obteve melhor desempenho, pois gera extratos cujas sentenças cobrem de forma mais adequada o conteúdo dos sumários humanos.

Para fins de comparação, considerou-se outro método profundo da literatura, o CSTSumm (CASTRO JORGE; PARDO, 2010), considerado estado da arte dessa

abordagem. Os extratos gerados por esse sumarizador para as coleções C1, C2, C3, C4 e C5 estão atualmente disponíveis na interface *online* do CSTNews. Para a avaliação, tais extratos foram submetidos à interface *online* da ROUGE juntamente com os extratos produzidos pelos métodos léxico-conceituais. Os resultados obtidos estão descritos na Tabela 14.

Tabela 14 – Comparação geral das avaliações ROUGE

Método	ROUGE-1			ROUGE-2		
	Cobertura	Precisão	Medida-F	Cobertura	Precisão	Medida-F
LCFSummN	0.415082	0.407554	0.41079	0.20653	0.20392	0.20497
LCFSummN-pond	0.433772	0.423922	**0.42852**	0.21298	0.20755	**0.21009**
LCFSummN-V	0.425422	0.413144	0.41868	0.20916	0.20378	0.20617
LCFSummN-V-pond	0.420848	0.401462	0.41051	0.19979	0.19025	0.19469
CSTSumm	0.48052	0.427008	**0.451854**	0.243502	0.216736	**0.229146**

Considerando a medida-f como parâmetro principal, posto que esta é uma média ponderada das medidas de precisão e cobertura, destaca-se, com base nos resultados da Tabela 14, que o método LCFSummN-pond apresenta os melhores resultados entre os métodos léxico-conceituais, mas não supera o estado da arte. Embora não supere os valores obtidos pelo CSTSumm, as medidas-f de LCFSummN-pond são bastante próximas às medidas do método baseado em conhecimento discursivo, indicando que os extratos gerados por LCFSummN-pond têm bom nível de informatividade.

6. CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

Neste trabalho, investigou-se a aplicação de conhecimento linguístico profundo de nível léxico-conceitual na SAM para o português. Especificamente, testaram-se várias versões do método genérico denominado LCFSumm em 5 coleções do *corpus* CSTNews para a produção de sumários extrativos em português.

No geral, os métodos baseados em conhecimento léxico-conceitual, denominados LCFSummN, LCFSummN-pond, LCFSummN-V e LCFSummN-V-pond, geram extratos com "boa" qualidade linguística Dentre os métodos LCFSumm, o método LCFSummN, o qual se baseia exclusivamente na frequência dos conceitos nominais na coleção, obteve as melhores médias na maioria dos critérios da TAC, indicando que os conceitos verbais não contribuem para a identificação da informação principal da coleção. A combinação da frequência simples dos conceitos nominais e verbais na seleção de conteúdo, realizada pelo método LCFSummN-V, melhorou apenas a gramaticalidade dos extratos. Quanto à informatividade, o método LCFSummN-pond apresentou os melhores resultados de medida-f quanto à ROUGE-1 e 2, aproximando-se do CSTSumm, mas não o superando.

Vale ressaltar que os resultados ora apresentados são apenas indícios sobre a relevância da utilização de conhecimento léxico-conceitual na SAM monolíngue. Diz-se isso porque tais resultados precisam ser validados em um *corpus* maior, posto que somente 5 coleções multidocumento foram utilizadas neste trabalho. O tamanho do *corpus* utilizado, aliás, pode ser considerado uma limitação desta pesquisa.

Além da investigação dos métodos, este trabalho contribui para as pesquisas no PLN com a sistematização de diretrizes para a anotação conceitual de nomes e com a extensão da anotação conceitual dos nomes de 5 coleções do CSTNews via WN.Pr, enriquecendo o referido recurso linguístico-computacional. Ademais, produziu-se neste trabalho um manual de utilização da ferramenta automática NILC-WISE. Como trabalho futuro, destaca-se a intenção de se aplicar os métodos baseados em conhecimento léxico-conceitual a mais coleções do CSTNews, o que implicará na ampliação da anotação conceitual dos nomes nas coleções a serem selecionadas.

REFERÊNCIAS BIBLIOGRÁFICAS

AKABANE, A. T.; PARDO, T. A. S.; RINO, L. H. M. Explorando medidas de redes complexas para sumarização multidocumento. In: STIL STUDENT WORKSHOP ON INFORMATION AND HUMAN LANGUAGE TECHNOLOGY, 2011, Cuiabá. **Proceedings…** Cuiabá, 2011. p. 1-3.

BARZILAY, R.; MCKEOWN, K.; ELHADAD, M. Information fusion in the context of multi-document summarization. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 37., 1999, Maryland. **Proceedings…** Maryland, 1999. p. 550-7.

BAXENDALE, P. B. Machine-made index for technical literature-an experiment. **IBM Journal of Research and Development**, v. 2, n. 4, p. 354-361, 1958.

BOSSARD, A.; RODRIGUES, C. Combining a Multi-Document Update Summarization System –CBSEAS– with a Genetic Algorithm. **Systems and Technologies**, v. 8, p. 71-87, 2011.

CAMARGO, R. T. **Investigação de estratégias de sumarização humana multidocumento**. 2013. 133 p. Dissertação (Mestrado, Programa de Pós-Graduação em Linguística) - Universidade Federal de São Carlos, São Carlos, SP, 2013.

CAMARGO, R. T.; DI FELIPPO, A.; PARDO, T. A. S. On Strategies of Human Multi-Document Summarization. In: *10th BRAZILIAN SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY* - STIL, 2015, Natal. ***Proceedings***… Natal, 2015. p. 141-150.

CARBONELL, J.; GOLDSTEIN, J. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: ANNUAL INTERNATIONAL ACM SIGIR, 21, 1998, Melbourne. **Proceedings…** Melbourne, 1998. p. 335-336.

CARDOSO, P. C. F. **Exploração de métodos de sumarização automática multidocumento com base em conhecimento semântico-discursivo**. 2014. 182 p. Tese (Doutorado, Instituto de Ciências Matemáticas e de Computação) - Universidade de São Paulo, São Carlos, SP, 2014.

CARDOSO, P. C. F.; et al. CSTNews -A discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese. In: RST BRAZILIAN MEETING, 3, 2011, Cuiabá. **Proceedings…** Cuiabá, 2011. p. 88-105.

CARDOSO, P. C. F.; PARDO, T. A. S.; NUNES, M. G. V. Métodos para sumarização automática multidocumento usando modelos semântico-discursivos. In: RST BRAZILIAN MEETING, 3, 2011, Cuiabá. **Proceedings...** Cuiabá, 2011. p. 59-74.

CARDOSO, P.; PARDO, T. A. S. Multi-document summarization using semantic discourse models. **Processamiento de Lenguaje Natural**, v. 56, p. 57–64, 2016.

CARDOSO, P. C. F.; PARDO, T. A. S. Joint semantic discourse models for automatic multi-document summarization. In: BRAZILIAN SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY (STIL), 10, 2015, Natal. **Proceedings...** Natal, 2015. p. 81-90.

CASTRO JORGE, M. L. R. **Modelagem gerativa para sumarização automática multidocumento**. 2015. 151 p. Tese (Doutorado, Instituto de Ciências Matemáticas e de Computação) – Universidade de São Paulo, São Carlos, SP, 2015.

CASTRO JORGE, M. L. R.; PARDO, T. A. S. Experiments with CST-based Multi-document summarization. In: ACL WORKSHOP TEXTGRAPHS-5: GRAPH-BASED METHODS FOR NATURAL LANGUAGE PROCESSING, 2010, Uppsala. **Proceedings…**Uppsala, Sweden, 2010. p. 74-82.

CASTRO JORGE, M. L. R.; PARDO, T. A. S. A Generative approach for multi-document summarization using the Noisy Channel model. In: RST BRAZILIAN MEETING, 3, 2011, Cuiabá. **Proceedings…** Cuiabá, 2011. p. 75-87.

CRUSE, D. **Lexical semantics.** Cambridge: Cambridge University Press, 1986. 310 p.

DANG, H. T. Overview of the TAC 2008 opinion question answering and summarization tasks. In: TAC, 2008, Gaithersburg. **Proceedings…** Gaithersburg, 2008.

DIAS-DA-SILVA, B. C. Wordnet.Br: An Exercise of Human Language Technology Research. In: THIRD INTERNATIONAL WORDNET CONFERENCE, 2005, Brno. **Proceedings**... Brno: Masarykova Univ, 2005. p. 301-303.

DI FELIPPO, A.; TOSTA, F. E. S.; PARDO, T. A. S. Applying Lexical-Conceptual Knowledge for Multilingual Multi-document Summarization. In: *XII INTERNATIONAL CONFERENCE ON THE COMPUTATIONAL PROCESSING OF PORTUGUESE* - PROPOR (LNAI 9727), 2016, Tomar/Portugal. ***Proceedings…*** Tomar/Portugal, 2016. p. 38-49.

EDMUNDSON, H. P. New methods in automatic extracting. **Journal of the ACM**, v. 16, n. 2, p. 264-285, 1969.

ENDRES-NIGGEMEYER, B. **Summarization Information**. Berlin: Springer, 1998. 374 p.

FELLBAUM, C; Miller, G. A. (Ed.). **Wordnet: an electronic lexical database** (Language, speech and communication). Cambridge, Massachusetts: **MIT Press**, 1998. 442 p.

FILLMORE, C. J. Frame Semantics. In: **Linguistics in the Morning Calm.** Seoul, South Korea: Hanshin Publishing Co., 1982. p. 111–137.

FILLMORE, C. J.; JOHNSON, C. R.; PETRUCK, M. R. L. Background to framenet. **International Journal of Lexicography**, v. 16, n. 3, p. 235-250, 2003.

FIORIN, J. L. **Introdução à Linguística**. vol. II. Princípios de análise. São Paulo: Contexto, 2003.

GUPTA, V; LEHAL, G. S. A survey of text summarization extractive techniques. **Journal of Emerging Technologies in Web Intelligence**, v. 2, n. 3, p. 258-268, 2010.

HENNIG, L., UMBRATH, W., WETZKER, R. An ontology-based approach to text summarization. In: WORKSHOP ON NATURAL LANGUAGE PROCESSING AND ONTOLOGY ENGINEERING (NLPOE 2008), 3, 2008, Toronto. **Proceedings…** Toronto, Canada. 2008. p. 291-294.

JURAFSKY, D; MARTIN, J. **Speech and Language Processing**: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. New Jersey: Prentice Hall, 2007. 1024 p.

KUMAR Y. J.; SALIM N.; RAZA B. Cross-document structural relationship identification using supervised machine learning. **Applied Soft Computing**, v. 12, p. 3124-3131, 2012.

LI, L., et al Ontology-enriched multi-document summarization in disaster management. In: ACM SPECIAL INTEREST GROUP ON INFORMATION RETRIEVAL (SIGIR), 2010, Geneva. **Proceedings…** Geneva, Switzerland, 2010. p. 819-820.

LIMA, J. B. P; PARDO, T. A. S. Ordenação de Sentenças em Sumários Multidocumento. **Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação**. São Carlos: NILC-TR-12-02, 2012. 37 p.

LIN, C. Y. ROUGE: a Package for Automatic Evaluation of Summaries. In: Workshop ON TEXT SUMMARIZATION BRANCHES OUT (WAS 2004), 8, 2004, Barcelona. **Proceedings…** Barcelona, Spain, 2004. p. 74-81.

LIN, C. Y.; HOVY, E. Automated multi-document summarization in NeATS. In: HUMAN LANGUAGE TECHNOLOGY CONFERENCE, 2, 2002, San Diego. **Proceedings**… San Diego, California: Morgan Kaufmann Publishers Inc., 2002. p. 59-62.

LÓPEZ CONDORI, R. E. **Sumarização automática de opiniões baseada em aspectos.** 2015, 160 p. Dissertação (Mestrado, Instituto de Ciências Matemáticas e de Computação) - Universidade de São Paulo, São Carlos, SP, 2015.

LOUIS, A.; NENKOVA, A. Automatically assessing machine summary content without a gold standard. **Computational Linguistics**, Cambridge, MA, v. 39, n. 2, p. 267-300, 2013.

LUHN, H. P. The automatic creation of literature abstracts. **IBM Journal of Research and Development**, v. 2, p. 159-165, 1958.

LYONS, J. **Introdução à linguística teórica**. São Paulo: Nacional, 1979.

MANI, I. **Automatic Summarization**. Amsterdam: John Benjamins Publishing Co., 2001. 285 p.

MANI, I.; BLOEDORN, E. Multi-document summarization by graph search and matching. In: NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE (AAAI), 14, 1997, Rhode Island. **Proceedings…** Rhode Island, 1997. p. 622-628.

MANI, I.; MAYBURY, M. T. **Advances in automatic text summarization**. Cambridge, MA: MIT Press, 1999. 442 p.

MANN, W. C.; THOMPSON, S. A. **Rhetorical Structure Theory: a theory of text organization.** 1987. (Technical Report ISI/RS-87-190).

MARCU, D. **The theory and practice of discourse Parsing and Summarization**. Cambridge, Massachusetts: The MIT Press, 2000. 248 p.

MAZIERO, E. G., PARDO, T. A. S. Multi-document discourse parsing using traditional and hierarchical machine learning. In: BRAZILIAN SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY, 8, 2011, Cuiabá. **Proceedings…** Cuiabá, 2011. p. 1-10.

MAZIERO, E. G.; JORGE, M. L. C.; PARDO, T. A. S. Identifying multi-document relations. In: INTERNATIONAL WORKSHOP ON NATURAL LANGUAGE PROCESSING AND COGNITIVE SCIENCE, 7, 2010, Funchal. **Proceedings.**.. Funchal, 2010. p. 60-9.

MCKEOWN, K., RADEV, D. R. Generating summaries of multiple news articles. In: ANNUAL INTERNATIONAL ACM-SIGIR, 18, 1995, Seattle. **Proceedings…** Seattle, 1995. p. 74-82.

MIHALCEA, R. Knowledge-Based Methods for WSD. In: Agirre, E.; Edmonds, P. (Eds) **Word Sense Disambiguation: Algorithms and Applications.** Dordrecht: Springer, 2007. p. 107-131.

MIHALCEA, R.; TARAU, P. An algorithm for language independent single and multiple document Summarization. In: INTERNATIONAL JOINT CONFERENCE ON NATURAL LANGUAGE PROCESSING (IJCNLP), 2005, Korea. **Proceedings…** Korea, 2005. p. 19-21.

MILLER, G. A., FELLBAUM, C. Semantic networks of English. **Cognition.** v. 41, n. 1-3, p. 197–229, 1991.

NAYEEM, M. T.; CHALI, Y. Extract with Order for Coherent Multi-Document Summarization. In: TEXTGRAPHS-11: THE WORKSHOP ON GRAPH-BASED METHODS FOR NATURAL LANGUAGE PROCESSING, ACL, 2017, Vancouver. **Proceedings…** Vancouver, 2017. p. 51-56.

NENKOVA, A.; PASSONNEAU, R. Evaluating content selection in summarization: The pyramid method. In: CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (HLT/NAACL), 2004, Boston. **Proceedings...** Boston, MA, 2004. p. 1-8.

NENKOVA, A.; McKeown, K. Automatic Summarization. **Information Retrieval.** v. 5, n. 2-3, p. 103–233, 2011.

NÓBREGA, F. A. A. **Desambiguação lexical de sentidos para o português por meio de uma abordagem multilíngue mono e multidocumento.** 2013. 126 p. Dissertação (Mestrado, Instituto de Ciências Matemáticas e de Computação) - Universidade de São Paulo, São Carlos, SP, 2013.

NÓBREGA, F. A. A.; PARDO, T. A. S. General Purpose Word Sense Disambiguation Methods for Nouns in Portuguese. In: PROPOR 2014 PHD AND MSC/MA DISSERTATION CONTEST / 11ST INTERNATIONAL CONFERENCE ON COMPUTATIONAL PROCESSING OF PORTUGUESE - PROPOR (LNAI 8775), 2014, São Carlos. ***Proceedings…*** São Carlos, 2014. p. 94-101.

O'DONNELL, M. Variable-length on-line document generation. In: EUROPEAN WORKSHOP ON NATURAL LANGUAGE GENERATION, 6, 1997, Duisburg. **Proceedings…** Duisburg, Gerhard-Mercator University: 1997.

OLIVEIRA, H. G.; ANTON PEREZ, L.; GOMES, P. Integrating lexical-semantic knowledge to build a public lexical ontology for Portuguese. In: 17th INTERNATIONAL CONFERENCE ON APPLICATIONS OF NATURAL LANGUAGE PROCESSING AND INFORMATION SYSTEMS, 2012, Berlin. **Proceedings**… Berlin, 2012. p. 210-215.

PAICE, C. D. The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases. In: SIGIR, 4, 1980, Berkeley, California. **Proceedings…** Berkeley, California, 1981. p. 172–191.

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

PARDO, T. A. S.; ALEIXO, P. **CSTNews**: **um córpus de textos jornalísticos anotados segundo a teoria discursiva multidocumento CST (*Cross-documentStructureTheory*).** São Carlos: NILC-ICMC, 2008. (Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional).

PARDO, T. A. S.; RINO, L. H. M., NUNES, M. G. V. GistSumm: a summarization tool based on a new extractive method. In: MAMEDE, N.J., BAPTISTA, J., TRANCOSO, I., NUNES, M.G.V. (Eds.). **Computational Processing of the Portuguese Language – 6th International Workshop, Written and Spoken (PROPOR) (Lecture Notes in Artificial Intelligence 2721),** Faro/Portugal, Proceedings, 2003. p. 210-218.

RADEV, D. R. A common theory of information fusion from multiple text sources, step one: Cross-document structure. In: ACL SIGDIAL WORKSHOP ON DISCOURSE AND DIALOGUE, 1, 2000, Hong Kong. **Proceedings…**Hong Kong, 2000. p. 74-83.

RADEV, D. R.; MCKEOWN, K. Generating natural language summaries from multiple on-line sources. **Computational Linguistics**, v. 24, n. 3, p. 469-500, 1998.

RADEV, D.; et al. MEAD - A Platform for Multi Document Multilingual Text Summarization. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC 2004), 4, 2004, Lisbon. **Proceedings…** Lisbon, Portugal, 2004. p. 1-4.

RATNAPARKHI, A. A maximum entropy model for part-of-speech tagging. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 1996, Philadelphia. **Proceedings**… Philadelphia, 1996. p. 133-142.

RIBALDO, R. **Investigação de Mapas de Relacionamento para Sumarização Multidocumento.** 2013. 61 p. Monografia de Conclusão de Curso (Graduação, Instituto de Ciências Matemáticas e de Computação) - Universidade de São Paulo (ICMC/USP). São Carlos, SP, 2013.

RIBALDO, R.; CARDOSO, P. C. F.; PARDO, T. A. S. Exploring the subtopic-based relationship map strategy for multi-document summarization. ***Journal of Theoretical and Applied Computing* - RITA**, v. 23, n. 1, p. 183-211, 2016.

RIBALDO, R.; RINO, L. H. M.; PARDO, T. A. S. Graph-based Methods for Multi-document Summarization: Exploring Relationship Maps, Complex Networks and Discourse Information. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL PROCESSING OF PORTUGUESE, 10, 2012, Coimbra. **Proceedings...** Coimbra: Universidade de Coimbra, 2012. p. 260-271.

SAGGION, H; et al. Developing infrastructure for the evaluation of single and multi-document summarization systems in a cross-lingual environment. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC 2002), 3, 2002, Las Palmas. **Proceedings...** Las Palmas: ELRA, 2002. p. 747-754.

SAGGION, H.; LAPALME, G. Concept identification and presentation in the context of technical text summarization. In: NAACL-ANLP WORKSHOP ON AUTOMATIC SUMMARIZATION, 2000, Seattle. **Proceeedings...** Seattle: Association for Computational Linguistics, 2000. p. 1-10.

SALTON, G.; *et al.* Automatic text structuring and summarization. **Information Processing & Management**, v. 33, n, 2, p. 193-207, 1997.

SALTON, G. Automatic Text Processing. Reading, MA: Addison-Wesley, 1998.

SCHIFFMAN, B.; NENKOVA, A.; MCKEOWN, K. Experiments in multi-document summarization. In: INTERNATIONAL CONFERENCE ON HUMAN LANGUAGE TECHNOLOGY RESEARCH, 2, 2002, San Francisco. **Proceedings…** San Francisco, 2002. p. 52-58.

SCHILDER, F.; KONDADADI, R. FastSum: fast and accurate query-based multi-document summarization. In: MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (ACL), 46, 2008, Columbus. **Proceedings…**Columbus, 2008. p. 205–208.

SCHLUTER, N.; SØGAARD, A. Unsupervised extractive summarization via coverage maximization with syntactic and semantic concepts. In: 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), 2015, Beijing. **Proceedings…** Beijing, 2015. p. 840–844.

SHANNON, C. E. A Mathematical Theory of Communication. **Bell System Technical Journal**, vol. 27, p. 379–423, 1948.

SINCLAIR, J. Corpus and text: basic principles. In: WYNNE, M. (Ed.). **Developing linguistic corpora: a guide to good practice.** Oxford: Oxbow Books, 2005. p. 1-16.

SOBREVILLA CABEZUDO, M. A. **Investigação de métodos de desambiguação lexical de sentidos de verbos do português do Brasil.** 2015, 158p. Dissertação (Mestrado, Instituto de Ciências Matemáticas e de Computação) - Universidade de São Paulo, São Carlos, SP, 2015.

SOBREVILLA CABEZUDO, M.A.; PARDO, T.A.S. Exploring Classical and Linguistically Enriched Knowledge-based Methods for Sense Disambiguation of Verbs in Brazilian Portuguese News Texts. ***Procesamiento del Lenguaje Natural,*** v. 59, p. 83-90, 2017.

SPARCK JONES, K. **Automatic summarising**: a review and discussion of the state of the art. Cambridge: University of Cambridge, 2007. (Technical Report UCAM-CL-TR-679).

SPARCK JONES, K. Automatic summarizing: factors and directions. In: MANI, I.; MAYBURY, M. T. (Eds.). **Advances in automatic text summarization**. Cambridge, Massachusetts: MIT Press, 1998. p. 1-12.

SPARCK JONES, K. Discourse modeling for automatic summarisation. **Tech. Report No. 290**. University of Cambridge, UK, 1993. 30 p.

SPARCK JONES, K.; GALLIERS, J. R. **Evaluating Natural Language Processing systems: An analysis and review.** Berlin: Springer-Verlag, 1996.

SUANMALI, L.; SALIM, N.; BINWAHLAN, M.S. Fuzzy genetic summarization based text summarization. In: INTERNATIONAL CONFERENCE ON DEPENDABLE AUTOMATIC AND SECURE COMPUTING, 9, 2001, Sydney. **Proceedings...** Sydney: IEEE, 2011. p. 1184-1191.

TOSTA, F. E. S. **Aplicação de conhecimento léxico-conceitual na Sumarização Automática Multidocumento Multilíngue.** 2014. 116 p. Dissertação (Mestrado, Programa de Pós-Graduação em Linguística) - Universidade Federal de São Carlos, São Carlos, SP, 2014.

VARGAS, F. A. **Agrupamento semântico de aspectos para mineração de opinião.** 2017. 152 p. Dissertação (Mestrado, Instituto de Ciências Matemáticas e de Computação) - Universidade de São Paulo, São Carlos, SP, 2017,

VAN-HALTEREN, H.; TEUFEL, S. Examining the consensus between human summaries: initial experiments with factoid analysis. In: HLT-NAACL DUC WORKSHOP, 2003, Edmonton. **Proceedings**… Edmonton, 2003. p. 57-64.

VOSSEN, P. Introduction to EuroWordNet. **Computers and the Humanities,** v. 32, p. 73-89, 1998.

WAN, X. An exploration of document impact on graph-based multi-document summarization. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 2008, Honolulu. **Proceedings…** Honolulu, 2008. p. 755-762.

WHITE, J.; DOYON, J.; TALBOTT, S. Task tolerance of MT output in integrated text processes. In: ANLP-NAACL 2000 WORKSHOP: EMBEDDED MT SYSTEMS WORKSHOP, 2000, Seattle. **Proceedings…** Seattle, 2000, p. 9-16.

WU, C.-W.; LIU, C.-L. Ontology-based Text Summarization for Business News Articles. **Computers and Their Applications**, v. 2003, p. 389-392, 2003.

YASUNAGA, M.; et al. Graph-based Neural Multi-Document Summarization. In: 21st Conference on Computational Natural Language Learning (CoNLL 2017), 2017, Vancouver. **Proceedings…** Vancouver, 2017. p. 452–462.

ZACARIAS, A. C. I. **Investigação De Métodos De Sumarização Automática Multidocumento Baseados Em Hierarquias Conceituais.** 2016. 141 p. Dissertação (Mestrado, Programa de Pós-Graduação em Linguística) - Universidade Federal de São Carlos, São Carlos, SP, 2016.

APÊNDICE 1 – Sumários gerados por coleção[33]

CLUSTER C1

Extrato gerado pelo método LCFSummN

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. (D2_S3) 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. (D2_S6)

Extrato gerado pelo método LCFSummN-pond

Ao menos 17 pessoas morreram após a queda de um avião de passageiros na República Democrática do Congo. (D1_S1) O avião, explodiu e se incendiou, acrescentou o porta-voz da ONU em Kinshasa, Jean-Tobias Okala. (D3_S4) Acidentes aéreos são frequentes no Congo, onde 51 companhias privadas operam com aviões antigos principalmente fabricados na antiga União Soviética. (D1_S4)

Extrato gerado pelo método LCFSummN-V

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. (D2_S3) 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. (D2_S6)

Extrato gerado pelo método LCFSummN-V-pond

Ao menos 17 pessoas morreram após a queda de um avião de passageiros na República Democrática do Congo. (D1_S1) O avião, explodiu e se incendiou, acrescentou o porta-voz da ONU em Kinshasa, Jean-Tobias Okala. (D3_S4) Acidentes aéreos são frequentes no Congo, onde 51 companhias privadas operam com aviões antigos principalmente fabricados na antiga União Soviética. (D1_S4)

[33] As sentenças que compõem os extratos estão especificadas pela informação de sua origem (D_documento e S_posição no texto-fonte). Essa informação, no entanto, não costuma compor extratos automáticos reais.

***CLUSTER* C2**

Extrato gerado pelo método LCFSummN

A pesquisa CNI/Ibope realizada em julho e divulgada nesta sexta-feira mostra que o presidente Luiz Inácio da Silva teria 44% dos votos no primeiro turno, enquanto o candidato tucano Geraldo Alckmin teria 25% das intenções de voto. (D2_S1) A CNI explica que a pesquisa não traz a comparação com pesquisas anteriores para primeiro turno porque é a primeira vez que o Ibope utiliza a lista oficial de candidatos a presidente da República. (D2_S5) Naquela ocasião, faziam parte da lista da pesquisa CNI/Ibope Enéas (Prona), que tinha 2%; Pedro Simon (PMDB), também 2%, além de Cristovam Buarque e Eymael, cada com 1% das intenções. (D2_S7) Na hipótese de um segundo turno com a candidata Heloísa Helena, Lula também teve uma redução nas intenções de voto de 57% para 53%. (D2_S10)

Extrato gerado pelo método LCFSummN-pond

A pesquisa CNI/Ibope realizada em julho e divulgada nesta sexta-feira mostra que o presidente Luiz Inácio da Silva teria 44% dos votos no primeiro turno, enquanto o candidato tucano Geraldo Alckmin teria 25% das intenções de voto. (D2_S1) José Maria Eymael, do PSDC, e Rui Pimenta, do PCO, não chegaram a obter 1% das intenções de voto. (D1_S3) A CNI explica que a pesquisa não traz a comparação com pesquisas anteriores para primeiro turno porque é a primeira vez que o Ibope utiliza a lista oficial de candidatos a presidente da República. (D2_S5) Os votos brancos e nulos somaram 10% e aqueles que não sabem ou não opinaram são 4%. (D2_S9) Num eventual segundo turno com Heloísa Helena, a vantagem de Lula também cairia, de acordo com a pesquisa de hoje. (D1_S11) O candidato tucano Geraldo Alckmin, caiu de 34% para 28%. (D2_S14)

Extrato gerado pelo método LCFSummN-V

A pesquisa CNI/Ibope realizada em julho e divulgada nesta sexta-feira mostra que o presidente Luiz Inácio da Silva teria 44% dos votos no primeiro turno, enquanto o candidato tucano Geraldo Alckmin teria 25% das intenções de voto. (D2_S1)

A CNI explica que a pesquisa não traz a comparação com pesquisas anteriores para primeiro turno porque é a primeira vez que o Ibope utiliza a lista oficial de candidatos a presidente da República. (D2_S5)

Embora não permita comparações, vale relembrar que na pesquisa de junho Lula tinha 48% das intenções de voto; Alckmin 18% e Heloísa Helena, 5%. (D2_S6)
No segundo turno, as intenções de voto do presidente Lula caíram de 53% em junho para 50% em julho, enquanto o candidato Alckmin subiu de 29% para 36%. (D2_S8)
Na pesquisa, divulgada no último dia 19 pela Folha, Lula aparece com 44% das intenções de voto e Alckmin, com 28%. (D1_S21)

Extrato gerado pelo método LCFSummN-V-pond

A pesquisa CNI/Ibope realizada em julho e divulgada nesta sexta-feira mostra que o presidente Luiz Inácio da Silva teria 44% dos votos no primeiro turno, enquanto o candidato tucano Geraldo Alckmin teria 25% das intenções de voto. (D2_S1) José Maria Eymael, do PSDC, e Rui Pimenta, do PCO, não chegaram a obter 1% das intenções de voto. (D1_S3) Neste cenário, Lula sobe de 27% para 31% e Alckmin passa de 4% para 14%; Heloísa também cresce, de 1% para 6%. (D1_S7) Agora, Lula teria 50%, contra 36% de Alckmin.(D1_S10) Em junho, Lula tinha 57% e Heloísa 21%. (D1_S13) O candidato tucano Geraldo Alckmin, caiu de 34% para 28%. (D2_S14) A pesquisa ouviu 2.002 pessoas entre os dias 29 e 31 de julho, em 142 municípios do país. (D1_S14) A margem de erro é de dois pontos percentuais, para mais ou para menos. (D1_S15) A candidata do PSOL à Presidência vem em terceiro, com 10%. (D1_S22)

***CLUSTER* C3**

Extrato gerado pelo método LCFSummN

Uma câmera da Rede Globo posicionada estrategicamente em frente ao Palácio do Planalto, pouco antes do início do Jornal Nacional, captou imagens do assessor especial da Presidência para Assuntos Internacionais, Marco Aurélio Garcia, e seu assessor de imprensa, Bruno Gaspar, comemorando com gestos obscenos a notícia de que o reverso direito do avião da TAM - que explodiu na última terça-feira contra um prédio da companhia aérea provocando a morte de pelo menos 198 pessoas - estava com defeito. (D3_S4) A falha no reversor --mecanismo que ajuda o avião a frear-- foi detectada pelo sistema eletrônico de checagem da própria aeronave, que continuou voando nos dias seguintes, com o reversor direito desligado. (D1_S5) Caso a possibilidade de que uma falha na aeronave tenha provocado a tragédia com o Airbus, as suspeitas de que o acidente esteja relacionado a falhas na recém-reformada pista do aeroporto perdem força. (D1_S15) A informação de que o Airbus A320 da TAM prefixo PR-MBK apresentava

um defeito no reverso - peça acionada durante o pouso para inverter o sentido da propulsão das turbinas - direito foi antecipada pela jornalista Hildegard Angel, em sua coluna da edição de ontem do JB e confirmada pelo vice-presidente técnico da companhia, Ruy Amparo. (D3_S15)

Extrato gerado pelo método LCFSummN-pond

O Airbus-A320 da TAM acidentado em São Paulo na última terça-feira (17) tinha um defeito no reverso da turbina direita desde o último dia 13. (D1_S1) A TAM confirmou, na noite desta quinta-feira, que ao airbus da TAM estava com o reverso do lado direito desligado, desde o último dia 13. (D3_S1) De acordo com a empresa aérea, o reversor estava travado, mas argumentou que a aeronave tinha condições de pouso normais, mesmo sem ele. (D3_S2) A falha no reversor --mecanismo que ajuda o avião a frear-- foi detectada pelo sistema eletrônico de checagem da própria aeronave, que continuou voando nos dias seguintes, com o reversor direito desligado. (D1_S5) Em 1996, uma falha no reverso foi a causa do acidente com o Fokker-100 da TAM, ocorrido segundos depois da decolagem, também em Congonhas. (D1_S13) Caso a possibilidade de que uma falha na aeronave tenha provocado a tragédia com o Airbus, as suspeitas de que o acidente esteja relacionado a falhas na recém-reformada pista do aeroporto perdem força. (D1_S15) Ainda segundo o "Jornal Nacional", o mesmo avião teve problemas para pousar, um dia antes do acidente, também em Congonhas. (D1_S17)

Extrato gerado pelo método LCFSummN-V

De acordo com a empresa aérea, o reversor estava travado, mas argumentou que a aeronave tinha condições de pouso normais, mesmo sem ele. (D3_S2) Uma câmera da Rede Globo posicionada estrategicamente em frente ao Palácio do Planalto, pouco antes do início do Jornal Nacional, captou imagens do assessor especial da Presidência para Assuntos Internacionais, Marco Aurélio Garcia, e seu assessor de imprensa, Bruno Gaspar, comemorando com gestos obscenos a notícia de que o reverso direito do avião da TAM - que explodiu na última terça-feira contra um prédio da companhia aérea provocando a morte de pelo menos 198 pessoas - estava com defeito. (D3_S4) A falha no reversor --mecanismo que ajuda o avião a frear-- foi detectada pelo sistema eletrônico de checagem da própria aeronave, que continuou voando nos dias seguintes, com o reversor direito desligado. (D1_S5) O reverso é um instrumento auxiliar na hora de frear o avião, mas não é considerado fator importante para o pouso, portanto, mesmo não

funcionando, ele poderia ter brecado a aeronave sem problemas, diz. (D2_S7) Caso a possibilidade de que uma falha na aeronave tenha provocado a tragédia com o Airbus, as suspeitas de que o acidente esteja relacionado a falhas na recém-reformada pista do aeroporto perdem força. (D1_S15)

Extrato gerado pelo método LCFSummN-V-pond

O Airbus-A320 da TAM acidentado em São Paulo na última terça-feira (17) tinha um defeito no reverso da turbina direita desde o último dia 13. (D1_S1) A TAM confirmou, na noite desta quinta-feira, que ao airbus da TAM estava com o reverso do lado direito desligado, desde o último dia 13. (D3_S1) De acordo com a empresa aérea, o reversor estava travado, mas argumentou que a aeronave tinha condições de pouso normais, mesmo sem ele. (D3_S2) A falha no reversor --mecanismo que ajuda o avião a frear-- foi detectada pelo sistema eletrônico de checagem da própria aeronave, que continuou voando nos dias seguintes, com o reversor direito desligado. (D1_S5) Em 1996, uma falha no reverso foi a causa do acidente com o Fokker-100 da TAM, ocorrido segundos depois da decolagem, também em Congonhas. (D1_S13) Caso a possibilidade de que uma falha na aeronave tenha provocado a tragédia com o Airbus, as suspeitas de que o acidente esteja relacionado a falhas na recém-reformada pista do aeroporto perdem força. (D1_S15) Ainda segundo o "Jornal Nacional", o mesmo avião teve problemas para pousar, um dia antes do acidente, também em Congonhas. (D1_S17)

***CLUSTER* C4**

Extrato gerado pelo método LCFSummN

A cidade tinha oito pontos de alagamento, sendo dois intransitáveis, e a Marginal do Pinheiros registrava o pior ponto de lentidão da capital, com 9,4 km da Rodovia Castelo Branco até a Ponte Cidade Jardim, no sentido Interlagos. (D2_S3) Segundo o CGE, às 9h30 havia alagamentos no viaduto José Colassuono; na pista expressa da avenida Alcântara Machado, sentido bairro, altura da rua Doutor Fomm; na avenida Airton Pretini, sentido centro; na avenida Magalhães de Castro, sentido Castello Branco; na avenida das Nações Unidas, sentido Castello Branco, na rua João Alfredo, sentido centro; na marginal Tietê, sentido centro; e na avenida Otaviano Alves de Lima, sentido Castello Branco. (D1_S4)

Extrato gerado pelo método LCFSummN-pond

Naquele horário, segundo a CET (Companhia de Engenharia de Tráfego), havia 110 km de congestionamento em toda a cidade enquanto a média para o horário era de 76 km. (D1_S2) Segundo o CGE, às 9h30 havia alagamentos no viaduto José Colassuono; na pista expressa da avenida Alcântara Machado, sentido bairro, altura da rua Doutor Fomm; na avenida Airton Pretini, sentido centro; na avenida Magalhães de Castro, sentido Castello Branco; na avenida das Nações Unidas, sentido Castello Branco, na rua João Alfredo, sentido centro; na marginal Tietê, sentido centro; e na avenida Otaviano Alves de Lima, sentido Castello Branco. (D1_S4)

Extrato gerado pelo método LCFSummN-V

Às 9 horas, a cidade tinha 113 km de lentidão, sendo que a média para o horário é de 82 km, segundo a Companhia de Engenharia de Tráfego (CET). (D2_S2) Segundo o CGE, às 9h30 havia alagamentos no viaduto José Colassuono; na pista expressa da avenida Alcântara Machado, sentido bairro, altura da rua Doutor Fomm; na avenida Airton Pretini, sentido centro; na avenida Magalhães de Castro, sentido Castello Branco; na avenida das Nações Unidas, sentido Castello Branco, na rua João Alfredo, sentido centro; na marginal Tietê, sentido centro; e na avenida Otaviano Alves de Lima, sentido Castello Branco. (D1_S4)

Extrato gerado pelo método LCFSummN-V-pond

Às 9 horas, a cidade tinha 113 km de lentidão, sendo que a média para o horário é de 82 km, segundo a Companhia de Engenharia de Tráfego (CET). (D2_S2) Segundo o CGE, às 9h30 havia alagamentos no viaduto José Colassuono; na pista expressa da avenida Alcântara Machado, sentido bairro, altura da rua Doutor Fomm; na avenida Airton Pretini, sentido centro; na avenida Magalhães de Castro, sentido Castello Branco; na avenida das Nações Unidas, sentido Castello Branco, na rua João Alfredo, sentido centro; na marginal Tietê, sentido centro; e na avenida Otaviano Alves de Lima, sentido Castello Branco. (D1_S4) O estado de atenção na cidade foi suspenso às 9h25. (D1_S5)

***CLUSTER* C5**

Extrato gerado pelo método LCFSummN

O ministro da Defesa, Nelson Jobim, informou no fim da noite desta terça-feira que a economista Solange Vieira, de 38 anos, será a nova presidente da Agência Nacional de Aviação Civil (Anac). (D2_S1) Mas, diante da dificuldade para encontrar pessoas que aceitassem assumir uma das diretorias da agência reguladora, após a renúncia de três diretores, Jobim decidiu indicar a economista para o cargo. (D1_S4) Alvo de críticas incisivas da oposição desde o acidente com o Airbus da TAM, o atual presidente da Anac, Milton Zuanazzi, já teria concordado em renunciar e deve entregar o cargo nos próximos dias. (D2_S4) Como os diretores de agências têm mandato de cinco anos, só podem sair por renúncia, decisão judicial ou acusação de improbidade administrativa. (D2_S6)

Extrato gerado pelo método LCFSummN-pond

O ministro da Defesa, Nelson Jobim, deve encaminhar o nome da economista Solange Vieira para assumir uma das diretorias da Agência Nacional de Aviação Civil (Anac). (D1_S1) O ministro da Defesa, Nelson Jobim, informou no fim da noite desta terça-feira que a economista Solange Vieira, de 38 anos, será a nova presidente da Agência Nacional de Aviação Civil (Anac). (D2_S1) Alvo de críticas incisivas da oposição desde o acidente com o Airbus da TAM, o atual presidente da Anac, Milton Zuanazzi, já teria concordado em renunciar e deve entregar o cargo nos próximos dias. (D2_S4) Outros três diretores já entregaram os cargos. (D2_S9) Em 2001, por cerca de oito meses, Solange comandou com mão-de-ferro os fundos de pensão do país, como titular da Secretaria de Previdência Complementar (SPC) do Ministério da Previdência. (D2_S10)

Extrato gerado pelo método LCFSummN-V

O ministro da Defesa, Nelson Jobim, informou no fim da noite desta terça-feira que a economista Solange Vieira, de 38 anos, será a nova presidente da Agência Nacional de Aviação Civil (Anac). (D2_S1) Mas, diante da dificuldade para encontrar pessoas que aceitassem assumir uma das diretorias da agência reguladora, após a renúncia de três diretores, Jobim decidiu indicar a economista para o cargo. (D1_S4) Como os diretores de agências têm mandato de cinco anos, só podem sair por renúncia, decisão judicial ou acusação de improbidade administrativa. (D2_S6) Ele disse que não está convencido da "participação objetiva" de Zuanazzi nas denúncias contra a agência: - Não podemos

indiciar para agradar à oposição, ao governo ou a quem quer que seja - disse Maia. (D2_S17)

Extrato gerado pelo método LCFSummN-V-pond

O ministro da Defesa, Nelson Jobim, deve encaminhar o nome da economista Solange Vieira para assumir uma das diretorias da Agência Nacional de Aviação Civil (Anac). (D1_S1) O ministro da Defesa, Nelson Jobim, informou no fim da noite desta terça-feira que a economista Solange Vieira, de 38 anos, será a nova presidente da Agência Nacional de Aviação Civil (Anac). (D2_S1) Braço direito do peemedebista desde sua posse, ela é funcionária de carreira do BNDES e comandou a Secretaria de Previdência Complementar no governo Fernando Henrique Cardoso. (D2_S2) A Solange vai ser a nova presidente da Anac - disse Jobim, em jantar que celebrou os 50 anos da Rede RBS em Brasília. (D2_S3) Outros três diretores já entregaram os cargos. (D2_S9) Após vários desentendimentos com o então ministro da pasta, Roberto Brant, foi demitida. (D2_S11)

APÊNDICE 2 –Manual de utilização da interface NILC-WISE

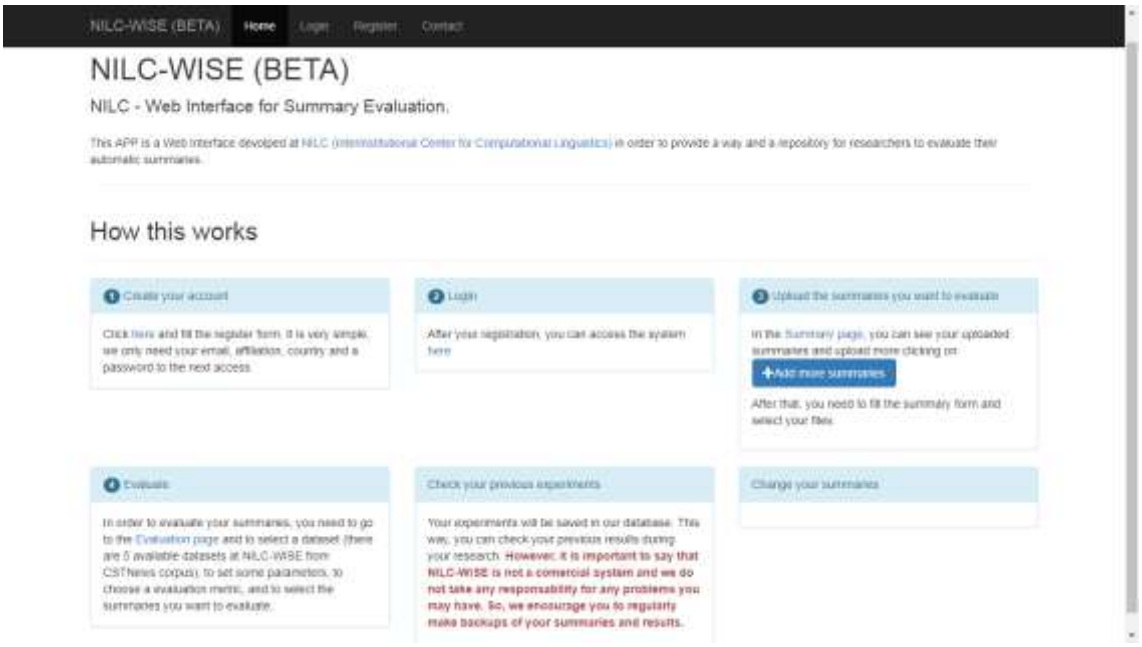
Objetivo: a ferramenta desenvolvida por Nóbrega e Pardo (2016) tem o objetivo de facilitar a avaliação da informatividade de sumários gerados para o *corpus* CSTNews com base no pacote de medidas automáticas da ROUGE.

O pacote de medidas ROUGE: As medidas ROUGE mensuram o nível de informatividade de um sumário por meio do cálculo da coocorrência de n-gramas, ou seja, realiza a comparação da quantidade de palavras em comum entre o sumário automático e um ou mais sumários humanos. Na medida ROUGE, o n-grama pode variar de 1 a 4 palavras. Dessa forma, a medida ROUGE-1 calcula a coocorrência de unigramas, a ROUGE-2 calcula a coocorrência de bigramas, e assim sucessivamente. Nesse pacote de medidas, a avaliação é obtida pelos cálculos de precisão (P), cobertura (C) e medida-f (em inglês, *precision*, *recall* e *f-measure*, respectivamente) (LIN, 2004).

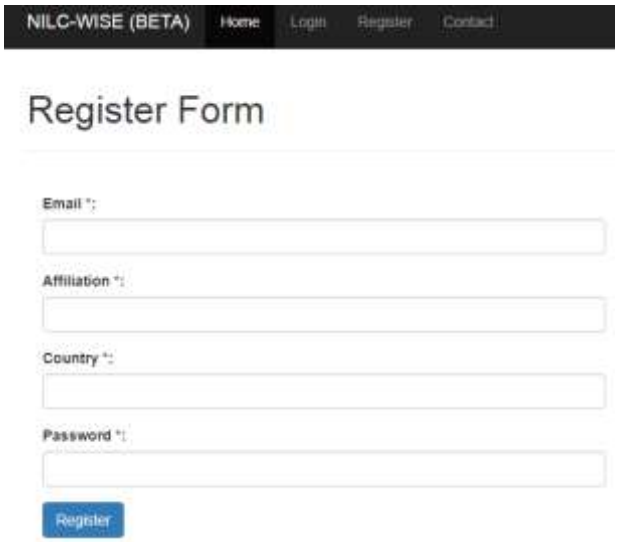
Ferramenta: a interface é denominada NILC-WISE (NILC - Web Interface for Summary Evaluation) e está disponível em http://nilc.icmc.usp.br/nilcwise. A ferramenta possui em sua base de dados 5 conjuntos de textos de referência para o *corpus* CSTNews, sendo (i) *abstracts*, (ii) extratos e (iii) textos-fonte configurados para a Sumarização de Atualização. Assim, os sumários gerados para o CSTNews podem ser comparados a cada um desses conjuntos de textos, conforme a necessidade do usuário.

Metodologia de avaliação: o processo de avaliação por meio da interface NILC-WISE consiste nos seguintes passos:

Acessar a página web: http://nilc.icmc.usp.br/nilcwise que deverá ser exibida da seguinte forma:



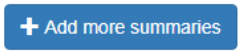
Criar uma conta de acesso: para isso o usuário deve cadastrar-se clicando em "Register" na barra superior. Para o registro é preciso fornecer dados como e-mail, afiliação, país e informar uma senha, conforme figura abaixo. O e-mail e a senha informados serão necessários para acessar o sistema.



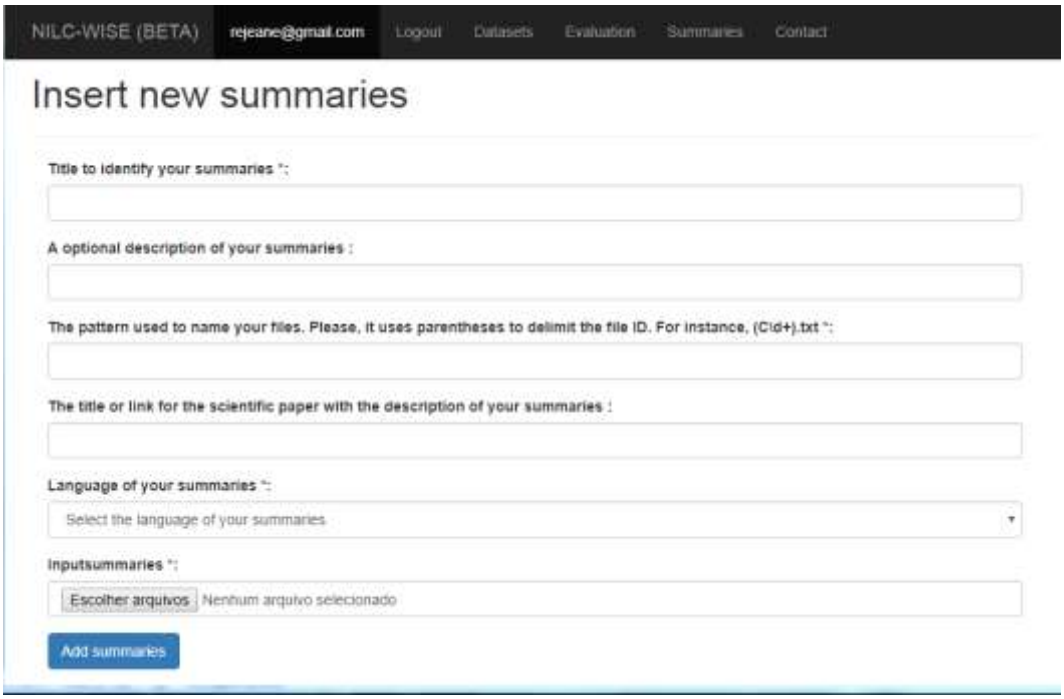
Acessar o sistema: na barra superior o usuário deverá clicar em "Login" e então, informar o e-mail e a senha cadastrados anteriormente e clicar em "Sign in". Após estar "logado" na sua conta, o usuário poderá iniciar o processo de submissão de sumários para avaliação. Esse processo conta com 2 etapas, conforme descritas a seguir.

Etapa 1: Inserir os sumários a serem avaliados

Para inserir os arquivos dos sumários a serem avaliados, o usuário deve clicar em "*Summaries*" na barra superior e então em "*Add more sumaries*"



O usuário será direcionado para a seguinte página:



Nessa página será necessário informar os dados a respeito do sumário a ser avaliado.

É possível inserir apenas 1 (um) arquivo por vez. Os campos indicados com * são obrigatórios.

No campo "*Title to identify your summaries*" é necessário informar o nome dado ao arquivo que contém o sumário a ser avaliado. Esse nome deve seguir o padrão dos arquivos do CSTNews que estão na base de dados da ferramenta NILC-WISE com os quais o sumário em avaliação será comparado. No caso do CSTNews, cada conjunto de textos de referências está no formato C[1-50]_[letras].txt. Assim, se o usuário for submeter à avaliação um sumário gerado para a Coleção 1, por exemplo, o arquivo deve ser nomeado como C1 seguido de _ e uma letra (apenas uma letra), por exemplo: C1_N.

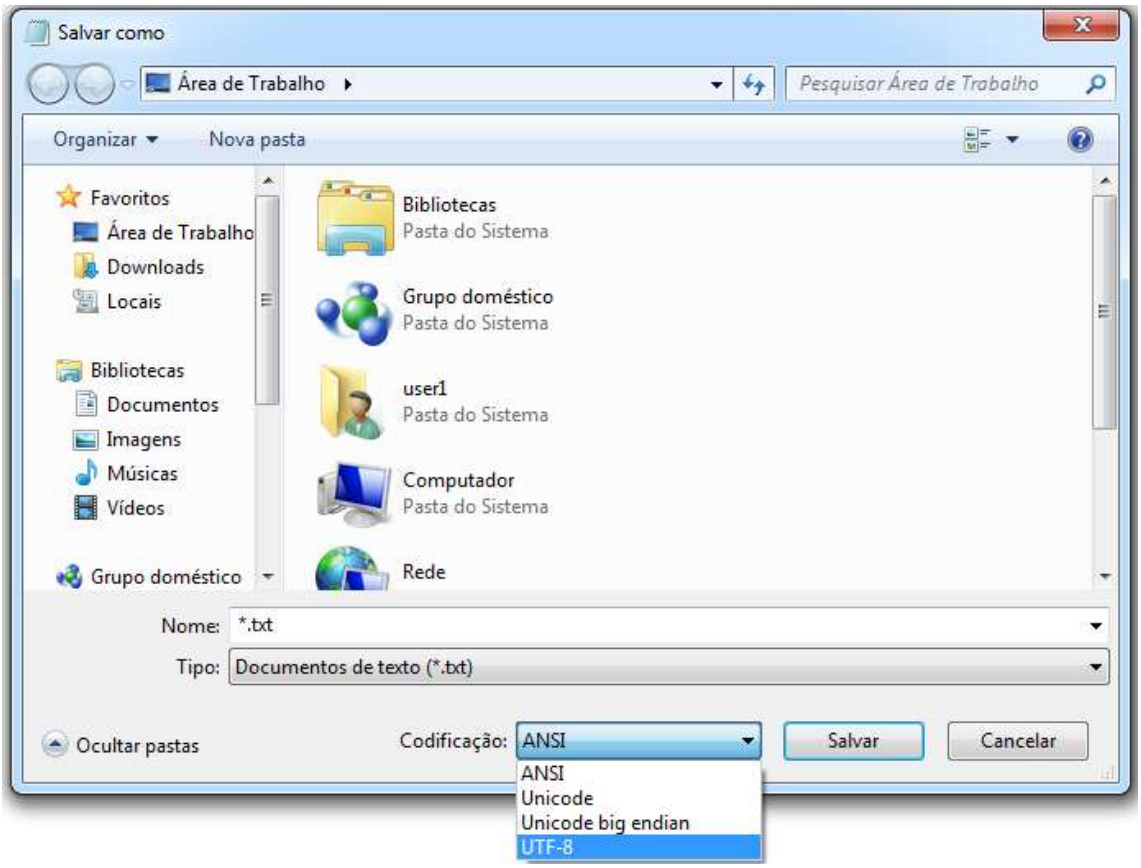
O campo "*A optional description of your summaries*" não é obrigatório e poderá ser preenchido com a descrição do sumário em avaliação, a critério do usuário.

O campo seguinte "*The pattern used to name your files. Please, it uses parentheses to delimit the file ID. For instance, (C\d+).txt*" refere-se ao padrão (utilizado para nomear os arquivos) por meio do qual a ROUGE encontrará o arquivo a ser avaliado. Assim, no caso de um arquivo nomeado como C1_N, por exemplo, esse campo deverá ser preenchido da seguinte forma: (C\d+)_N.txt. Esse padrão especifica para a ROUGE procurar arquivos que:

- Iniciem com a letra **C**
- Seguida de um ou mais números (= **\d+**)
- O caractere **_**
- Finalizado com a letra **N.txt**

O campo "*The title or link for the scientific paper with the description of your summarie*s" não é obrigatório e poderá ser preenchido a critério do usuário com o link do artigo científico que descreve os detalhes sobre o sumário sob avaliação.
No campo "*Language of your summaries*" é necessário indicar a língua na qual o sumário foi gerado. Como a ferramenta possui indexada à sua base de dados apenas arquivos referentes ao *corpus* CSTNews, que é um corpus de referência para o Português, a única língua disponível é o português. Por isso, deve-se selecionar a opção PT-BR

No campo "*Inputsummaries*" o usuário deverá carregar o arquivo do sumário a ser avaliado clicando em "Escolher arquivos". O arquivo deve possuir formato e codificação conforme as especificações da ferramenta NILC-WISE. Nesse caso, os arquivos devem ser criados por um editor de texto que apresente a extensão .txt (o "bloco de notas" e o "Notepad++" são exemplos de editores de texto) e devem ser codificados como UTF-8. No "Bloco de Notas" do Windows, por exemplo, ao solicitar para "salvar" o arquivo, a opção "Codificação" irá aparecer na tela (conforme imagem abaixo) e então, é só selecionar "UTF-8":



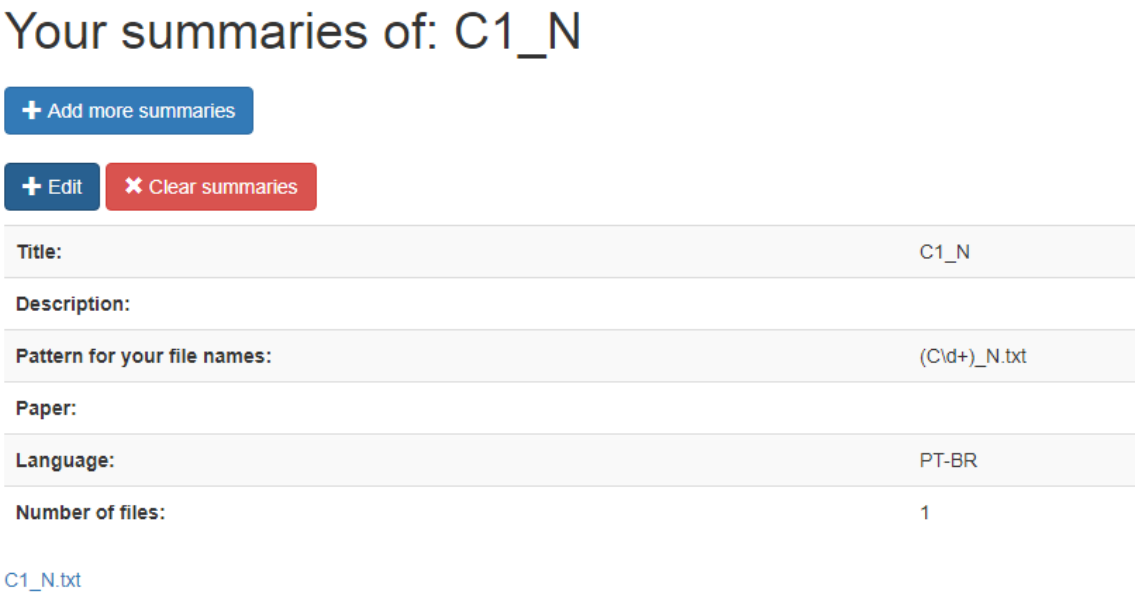
Após selecionar o arquivo a ser carregado é necessário clicar em Add summaries e então, a etapa de inserção de sumários será finalizada.

Esta etapa deverá ser repetida a cada novo arquivo a ser submetido.

Todos os arquivos inseridos pelo usuário ficarão salvos no sistema e poderão ser editados quando necessário. Para acessá-los, o usuário deve clicar em "Summaries" na aba superior. Então, todos os arquivos salvos serão exibidos.

Para editar ou excluir um arquivo é necessário clicar no ícone "details" correspondente ao arquivo a ser editado/excluído (conforme a figura abaixo):

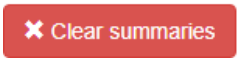
Após clicar em "details" a ferramenta exibirá as informações sobre o sumário em questão bem como arquivo carregado quando da inserção do sumário no sistema. Uma tela parecida com a figura seguinte será exibida:



Para editar as informações, o usuário deve clicar em 

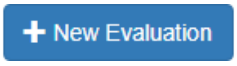
. Todos os campos podem ser alterados, exceto o campo "Title", pois este é imutável. Além disso, um novo arquivo deverá necessariamente ser carregado, pois ao clicar para editar, a ferramenta exclui automaticamente o arquivo carregado anteriormente. Para finalizar a edição, é necessário clicar em 

no final da página.

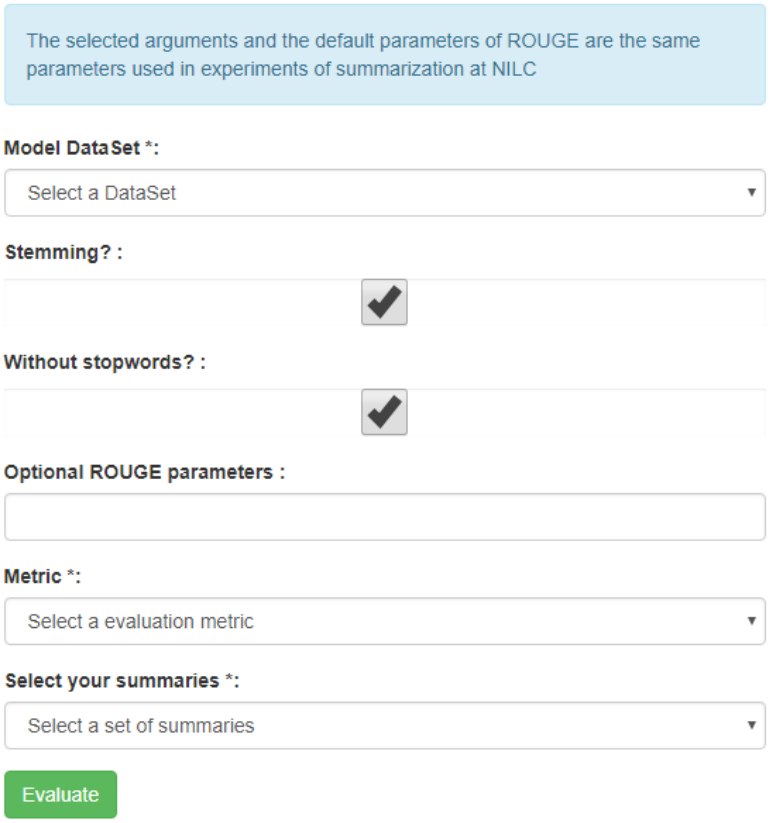
Para excluir um arquivo, após clicar em "details", o usuário deve clicar em 

. O sistema pedirá para o usuário confirmar a exclusão. Ao confirmar, o sumário será excluído do sistema.

Etapa 2: Avaliação dos sumários

Para dar início à avaliação dos sumários, o usuário deve clicar em "*Evaluation*" na barra superior e então em 

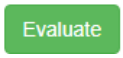
O usuário será direcionado para a página "*Evaluate your summaries*" e o seguinte formulário aparecerá:



Em "*Model DataSet*" o usuário deverá selecionar o conjunto de sumários de referência com o qual o sumário em avaliação deverá ser comparado. A ferramenta NILC-WISE disponibiliza 5 conjuntos de textos de referência para o CSTNews:

Após selecionar o conjunto de sumários de referência com o qual deverá ser feita a comparação, o usuário deverá escolher a métrica a ser utilizada na avaliação. No caso da métrica ROUGE, é necessário selecionar "ROUGE"

No campo “*Select your summaries*”, o usuário deve selecionar o arquivo a ser avaliado, o qual já foi inserido na etapa 1.

Na sequência o usuário deve clicar em 

. O NILC-WISE realizará a avaliação e apresentará uma tela com os dados da avaliação e uma tabela com os cálculos das medidas de Cobertura (*R)*, Precisão (*P*) e Medida-F (*F*) do sumário avaliado para ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4 e ROUGE-L.

Na imagem abaixo apresenta-se um exemplo do resultado da avaliação ROUGE para um extrato gerado para a Coleção 1 (C1) e comparado ao conjunto de *abstracts* (sumários humanos):

---
1 ROUGE-1 Average_R: 0.40851 (95%-conf.int. 0.40851 - 0.40851)
1 ROUGE-1 Average_P: 0.41026 (95%-conf.int. 0.41026 - 0.41026)
1 ROUGE-1 Average_F: 0.40938 (95%-conf.int. 0.40938 - 0.40938)
.....
1 ROUGE-1 Eval 1.1 R:0.40851 P:0.41026 F:0.40938
---
1 ROUGE-2 Average_R: 0.17467 (95%-conf.int. 0.17467 - 0.17467)
1 ROUGE-2 Average_P: 0.17544 (95%-conf.int. 0.17544 - 0.17544)
1 ROUGE-2 Average_F: 0.17505 (95%-conf.int. 0.17505 - 0.17505)
.....
1 ROUGE-2 Eval 1.1 R:0.17467 P:0.17544 F:0.17505
---
1 ROUGE-3 Average_R: 0.10762 (95%-conf.int. 0.10762 - 0.10762)
1 ROUGE-3 Average_P: 0.10811 (95%-conf.int. 0.10811 - 0.10811)
1 ROUGE-3 Average_F: 0.10786 (95%-conf.int. 0.10786 - 0.10786)
.....
1 ROUGE-3 Eval 1.1 R:0.10762 P:0.10811 F:0.10786
---
1 ROUGE-4 Average_R: 0.07373 (95%-conf.int. 0.07373 - 0.07373)
1 ROUGE-4 Average_P: 0.07407 (95%-conf.int. 0.07407 - 0.07407)
1 ROUGE-4 Average_F: 0.07390 (95%-conf.int. 0.07390 - 0.07390)
.....
1 ROUGE-4 Eval 1.1 R:0.07373 P:0.07407 F:0.07390
---
1 ROUGE-L Average_R: 0.38298 (95%-conf.int. 0.38298 - 0.38298)
1 ROUGE-L Average_P: 0.38462 (95%-conf.int. 0.38462 - 0.38462)
1 ROUGE-L Average_F: 0.38380 (95%-conf.int. 0.38380 - 0.38380)
.....
1 ROUGE-L Eval 1.1 R:0.38298 P:0.38462 F:0.38380

Todas as avaliações são salvas na base de dados. Assim, quando o usuário necessitar consultar as informações sobre as avaliações realizadas, deverá acessar a página de avaliação (“Evaluation” na barra superior) e selecionar a avaliação que pretende consultar.