Programa de Pós-Graduação em Lingüística

PPGL

INVESTIGAÇÃO DE ESTRATÉGIAS DE SUMARIZAÇÃO HUMANA MULTIDOCUMENTO

Renata Tironi de Camargo

SÃO CARLOS
2013

Universidade Federal de São Carlos

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

INVESTIGAÇÃO DE ESTRATÉGIAS DE SUMARIZAÇÃO HUMANA MULTIDOCUMENTO

RENATA TIRONI DE CAMARGO

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.

Orientadora: Profa. Dra. Ariani Di Felippo
Coorientador: Prof. Dr. Thiago A. S. Pardo

São Carlos - São Paulo - Brasil
2013

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

C172ie

Camargo, Renata Tironi de.
Investigação de estratégias de sumarização humana multidocumento / Renata Tironi de Camargo. -- São Carlos : UFSCar, 2013.
132 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2013.

1. Linguística. 2. Sumarização automática. 3. Sumarização humana multidocumento. 4. Estratégias de seleção de conteúdo. I. Título.

CDD: 410 (20ª)

UNIVERSIDADE FEDERAL DE SÃO CARLOS
Centro de Educação e Ciências Humanas
Programa de Pós-Graduação em Linguística

BANCA EXAMINADORA DA DISSERTAÇÃO DE MESTRADO DE RENATA TIRONI DE CAMARGO

Profª. Drª. Ariani Di Felippo
Orientadora e Presidente
UFSCar/São Carlos

Profª. Drª. Sandra Maria Aluisio
Membro titular
USP/São Carlos

Profª. Drª. Flávia Bezerra de Menezes Hirata Vale
Membro titular
UFSCar/São Carlos

Submetida a defesa pública em sessão realizada em: 30/agosto/2013.
Homologada na 60ª reunião da CPGL, realizada em 27/09/2013.

Carlos Piovezani
Coordenador
PPGL/UFSCar

À minha família: fonte de inspiração e amor incondicional.

AGRADECIMENTOS

À minha mãe, **Edina**, e ao meu pai, **Waldir**, pelo amor incondicional e por sempre confiarem em mim e acreditarem nos meus sonhos.

À minha irmã, **Mariana**, pela amizade, pelo exemplo de amor fraterno, pela confiança e pelos ensinamentos.

Ao meu namorado, **Péricles**, pelo amor imensurável, pelo cuidado, pela paciência, pelo incentivo e por participar das minhas conquistas.

À minha orientadora, **Ariani Di Felippo**, e ao meu coorientador, **Thiago A. S. Pardo**, pelas palavras de sabedoria, pela paciência, sobretudo por me inserirem no surpreendente mundo da Linguística Computacional.

À minha orientadora de estágio, **Diana Santos**, por me acolher de forma carinhosa na Noruega e, principalmente, por me mostrar um mundo de ensinamento/aprendizado completamente diferente.

A todos os integrantes do NILC, pelo apoio, pelo companheirismo e pelas divertidas viagens de natureza acadêmica. À **Paula** e **Lucía**, pelos conhecimentos computacionais compartilhados. À **Verônica**, pela amizade e pelos incansáveis dias de anotação de *corpus*.

A todos os amigos, professores e funcionários da UFSCar.

Aos amigos da República Camorra e agregados, pelo acolhimento e por toda a diversão proporcionada.

Às minhas amigas de confinamento diário em casa, pela companhia. À **Thamara**, pela acolhida carinhosa e pela amizade.

Aos meus inestimáveis amigos de São Carlos, **Ana Paula**, **Fabricio** e **Duane**, com quem eu tenho dividido momentos que ficarão para sempre no coração.

Às minhas madrinhas, **Jacira** e **Angelina**, pela torcida e energias positivas.

Aos meus amigos de Cambará, por nunca me abandonarem e pelos momentos de distração.

À FAPESP, pelo apoio financeiro.

RESUMO

A sumarização humana multidocumento (SHM), que consiste na produção manual de um sumário a partir de uma coleção de textos, provenientes de fontes-distintas, que abordam um mesmo assunto, é uma tarefa linguística até então pouco explorada. Tomando-se como motivação o fato de que sumários monodocumento são compostos por informações que apresentam características recorrentes, a ponto de revelar estratégias de sumarização, objetivou-se investigar sumários multidocumento com o objetivo de identificar estratégias de SHM. Para a identificação das estratégias de SHM, os textos-fonte (isto é, notícias) das 50 coleções do *corpus* multidocumento em português CSTNews (CARDOSO et al., 2011) foram manualmente alinhados em nível sentencial aos seus respectivos sumários humanos, relevando, assim, a origem das informações selecionadas para compor os sumários. Com o intuito de identificar se as informações selecionadas para compor os sumários apresentam características recorrentes, as sentenças alinhadas (e não-alinhadas) foram caracterizadas de forma semiautomática em função de um conjunto de atributos linguísticos identificados na literatura. Esses atributos traduzem as estratégias de seleção de conteúdo da sumarização monodocumento e os indícios sobre a SHM. Por meio da análise manual das caracterizações das sentenças alinhadas e não-alinhadas, identificou-se que as sentenças selecionadas para compor os sumários multidocumento comumente apresentam certos atributos, como localização das sentenças no texto e redundância. Essa constatação foi confirmada pelo conjunto de regras formais aprendidas por um algoritmo de Aprendizado de Máquina (AM) a partir das mesmas caracterizações. Tais regras traduzem, assim, estratégias de SHM. Quando aprendidas e testadas no CSTNews pelo AM, as regras obtiveram precisão de 71,25%. Para avaliar a pertinência das regras, 2 avaliações intrínsecas foram realizadas, a saber: (i) verificação da ocorrência das estratégias em outro *corpus*, e (ii) comparação da qualidade de sumários produzidos pelas estratégias de SHM com a qualidade de sumários produzidos por estratégias diferentes. Na avaliação (i), realizada automaticamente por AM, as regras aprendidas a partir do CSTNews foram testadas em um *corpus* jornalístico distinto e obtiveram a precisão de 70%, muito próxima da obtida no *corpus* de treinamento (CSTNews). Na avaliação (ii), a qualidade, avaliada de forma manual por 10 linguistas computacionais, foi considerada superior à qualidade dos demais sumários de comparação. Além de descrever características relativas aos sumários multidocumento, este trabalho, uma vez que gera regras formais (ou seja, explícitas e não-ambíguas), tem potencial de subsidiar a Sumarização Automática Multidocumento (SAM), tornando-a mais linguisticamente motivada. A SAM consiste em gerar sumários multidocumento de forma automática e, para tanto, baseava-se na adaptação das estratégias identificadas na sumarização monodocumento ou apenas em indícios, não comprovados sistematicamente, sobre a SHM. Com base neste trabalho, a seleção de conteúdo em métodos de SAM poderá ser feita com base em estratégias identificadas de forma sistemática na SHM.

Palavras-chave: sumarização humana multidocumento, estratégia de seleção de conteúdo, sumarização automática multidocumento.

ABSTRACT

The multi-document human summarization (MHS), which is the production of a manual summary from a collection of texts from different sources on the same subject, is a little explored linguistic task. Considering the fact that single document summaries comprise information that present recurrent features which are able to reveal summarization strategies, we aimed to investigate multi-document summaries in order to identify MHS strategies. For the identification of MHS strategies, the source texts sentences from the CSTNews corpus (CARDOSO et al., 2011) were manually aligned to their human summaries. The corpus has 50 clusters of news texts and their multi-document summaries in Portuguese. Thus, the alignment revealed the origin of the selected information to compose the summaries. In order to identify whether the selected information show recurrent features, the aligned (and non-aligned) sentences were semi automatically characterized considering a set of linguistic attributes identified in some related works. These attributes translate the content selection strategies from the single document summarization and the clues about MHS. Through the manual analysis of the characterizations of the aligned and non-aligned sentences, we identified that the selected sentences commonly have certain attributes such as sentence location in the text and redundancy. This observation was confirmed by a set of formal rules learned by a Machine Learning (ML) algorithm from the same characterizations. Thus, these rules translate MHS strategies. When the rules were learned and tested in CSTNews by ML, the precision rate was 71.25%. To assess the relevance of the rules, we performed 3 different kinds of intrinsic evaluations: (i) verification of the occurrence of the same strategies in another corpus, and (ii) comparison of the quality of summaries produced by the HMS strategies with the quality of summaries produced by different strategies. Regarding the evaluation (i), which was automatically performed by ML, the rules learned from the CSTNews were tested in a different newspaper corpus and its precision was 70%, which is very close to the precision obtained in the training corpus (CSTNews). Concerning the evaluating (ii), the quality, which was manually evaluated by 10 computational linguists, was considered better than the quality of other summaries. Besides describing features concerning multi-document summaries, this work has the potential to support the multi-document automatic summarization, which may help it to become more linguistically motivated. This task consists of automatically generating multi-document summaries and, therefore, it has been based on the adjustment of strategies identified in single document summarization or only on not confirmed clues about MHS. Based on this work, the automatic process of content selection in multi-document summarization methods may be performed based on strategies systematically identified in MHS.

Keywords: multi-document human summarization, content selection strategy, multi-document automatic summarization.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE GRÁFICOS

LISTA DE SIGLAS

AM – Aprendizado de Máquina

CST – *Cross-document Structure Theory*

DUC – *Document Understanding Conference*

LC – Linguística de *Corpus*

NILC – Núcleo Interinstitucional de Linguística Computacional

PLN – Processamento de Linguagem Natural

ROUGE – *Recall-Oriented Understudy of Gisting Evaluation*

RST – *Rethorical Structure Theory*

SA – Sumarização Automática

SAM – Sumarização Automática Multidocumento

SHM – Sumarização Humana Multidocumento

SUCINTO – *Summarization for Clever Information Acces*

SUMMAC - *Text Summarization Evaluation Conference*

TAC – *Text Analysis Conference*

SUMÁRIO

1 INTRODUÇÃO

1.1 Contextualização

A sumarização humana monodocumento pode ser entendida como o processo de seleção de conteúdo de um texto-fonte para produzir uma versão mais curta do mesmo visando determinado usuário/tarefa (MANI; MAYBURY, 1999).

Essa atividade é importante para uma gama de profissionais, como educadores, pesquisadores e editores no geral. Para os educadores, a sumarização é atividade comum em sala de aula. Os pesquisadores, por sua vez, podem produzir sumários dos principais textos da literatura relacionados ao trabalho que estão desenvolvendo e sumários de seus próprios textos a serem publicados, como artigos, monografias, dissertações e teses. Para os editores do mercado jornalístico, a sumarização é atividade constante na medida em que precisam produzir versões condensadas de notícias.

Nos Estados Unidos, a sumarização constitui uma profissão, em que as atividades dos "sumarizadores profissionais" (do inglês, *professional asbtractors*) são regulamentadas por uma associação[1] própria.

A produção de um sumário é influenciada por vários fatores. Além do tamanho desejado do sumário, que determina o quanto a informação do texto-fonte deve ser condensada, a sumarização é influenciada pela audiência a que se destina o sumário, a função e o tipo do sumário, etc. (MANI, 2001). Com base na audiência, pode-se produzir um sumário genérico, que veicula a informação principal do texto-fonte, substituindo a leitura do mesmo, ou um sumário que veicula a informação de interesse para um tipo específico de usuário. Quanto à função, pode-se objetivar a produção de sumários informativos, isto é, contém as informações principais de um texto-fonte ao ponto de dispensar a leitura do original, indicativos, ou seja, não substitui o texto-fonte, apenas diz do que ele trata (p.ex.: índices de livros) ou críticos, os quais apresentam, além da informação principal do texto-fonte, avaliações sobre ele (p.ex.: resenhas). Quanto ao tipo, têm-se os sumários extrativos (ou extratos), ou seja, compostos por trechos inalterados dos textos-fonte, e os abstrativos (ou *abstracts*), isto é, construídos pela reescrita das informações dos textos-fonte.

[1] A associação em questão é a *National Federation of Abstracting and Information Services* (NFAIS). Mais informações podem ser encontradas no endereço http://www.nfais.org/.

A sumarização profissional tem sido alvo de vários estudos que buscam compreendê-la. Na maioria, os trabalhos investigam o processo de produção de sumários que são ao mesmo tempo genéricos, informativos e abstrativos, ou seja, destinados a uma audiência ampla e compostos pela informação principal do texto-fonte que foi alvo de reescrita.

Cremmins (1996), por exemplo, identificou que os profissionais produzem sumários acadêmicos[2] em 4 estágios: (i) interpretação, ou seja, leitura e identificação das principais características (do inglês, *features*) do texto, (ii) seleção, ou seja, identificação da informação pertinente a ser levada para o sumário, (iii) extração, organização e redução da informação selecionada e (iv) refinamento da informação relevante.

Estágios similares aos identificados por Cremmins (1996) também foram verificados por Endres-Niggemeyer (1998) ao analisar os protocolos de sumarização (no caso, registros que evidenciam o passo a passo da produção textual) gerados por 6 profissionais.

Com relação ao estágio de seleção, estudos que investigaram a produção de sumários extrativos (genéricos e informativos) mostraram que os humanos diferem quanto às informações extraídas dos textos originais, evidenciando, portanto, baixa concordância com relação às informações selecionadas.

O estudo desenvolvido por Rath et al. (1961), por exemplo, envolveu 10 artigos científicos e 6 humanos (sujeitos). Diante de cada um dos 10 textos, os 6 sujeitos selecionaram 20 sentenças para compor o sumário correspondente. Em média, os humanos concordaram na escolha de 1.6 sentenças por texto, ou seja, em 8% das 20 selecionadas. Ao considerar apenas 5 dos 6 sujeitos, essa média passou para 6.4 sentenças (32%).

Em Salton et al. (1997), 2 humanos produziram sumários extrativos para 50 artigos compilados de uma enciclopédia em língua inglesa. Para compor os sumários, os sujeitos selecionaram parágrafos dos textos-fonte. No caso, os autores verificaram que a sobreposição de conteúdo entre os sumários foi em média de 46%, ou seja, os extratos gerados por um dos sujeitos cobrem em média 46% da informação contida no sumário do outro.

A baixa concordância na seleção de conteúdo deve-se ao fato de que a produção textual em questão é uma atividade intelectual e como tal é quase sempre influenciada pelo conhecimento prévio, atitude e disposição do escritor. Assim, a seleção da informação pode depender também de: (i) objetivos do autor do sumário, (ii) objetivos ou interesses de seus

[2] Ou seja, *abstracts* genéricos e informativos produzidos a partir de textos do gênero científico.

possíveis leitores e (iii) importância relativa (e subjetiva) que o próprio sumarizador atribui às informações textuais (LUHN, 1958).

Se, por um lado, os humanos pouco concordam sobre toda a informação a ser extraída, estes parecem concordar quanto à informação principal (JOHNSON, 1970).

Marcu (1997a), com base em 13 humanos e 5 textos da revista *Scientific American*, identificou uma concordância de 71% na seleção da informação principal para compor sumários extrativos. Em outro trabalho, Marcu (1999) confirmou a evidência anterior ao verificar maior consistência na identificação da informação importante e menos consistência na identificação da informação menos importante.

Jing et al. (1998), a partir de um conjunto de 40 textos e 5 sujeitos, identificaram que os sujeitos concordaram em 96% quando os sumários extrativos eram compostos por 10% do texto-fonte e em 90% quando os sumários eram compostos por 20%, evidenciando que o tamanho do sumário tem influência. No caso, quanto menor o sumário, maior a concordância com relação à informação principal que deve ser selecionada.

Quando os humanos concordam quanto à informação principal que deve compor o sumário, eles selecionam as sentenças com base em certas características textuais superficiais e profundas, identificadas no estágio de interpretação dos textos-fonte.

Dentre as superficiais, encontram-se o título e os subtítulos (do inglês, *headings*), as expressões-chave ou indicativas (do inglês, *key phrases* e *cue phrases*, respectivamente) e a posição da informação no parágrafo ou texto (MANI, 2001).

Além das superficiais, os sumarizadores humanos também se baseiam na identificação da macroestrutura discursiva subjacente aos textos. Liddy (1991), ao estudar 276 sumários acadêmicos, verificou que estes comumente contêm as informações de *contexto, objetivo, metodologia, resultado* e *conclusão*. Tais informações constituem os "componentes discursivos" que garantem a transmissão do conteúdo principal dos textos do gênero científico sem perda de informatividade.

Ao se identificar que os humanos selecionam de forma recorrente a informação importante com base em certos atributos textuais, identificaram-se consequentemente "estratégias de seleção de conteúdo", a saber (CREMMINS, 1996; ENDRES-NIGGEMEYER, 1998):

(i) selecionar informação que se relaciona com as palavras contidas no título/subtítulo do texto-fonte;

(ii) selecionar informação por meio da identificação de expressões-chave do texto-fonte; no caso de artigos científicos, "o objetivo de este artigo é" é um exemplo de expressão-chave que indica o componente discursivo *objetivo*;

(iii) selecionar informação localizada em certas posições dos textos, as quais também são dependentes de gênero; no caso de textos jornalísticos, as informações localizadas no início dos textos expressam o fato principal de uma notícia e, por isso, são selecionadas para compor o sumário;

(iv) selecionar informação referente a determinados componentes discursivos; em textos, acadêmicos, por exemplo, os humanos comumente selecionam as informações que se referem aos componentes "contexto", "objetivo", "metodologia", "resultado" e "conclusão".

Se, por um lado, a sumarização humana monodocumento foi alvo de inúmeros estudos linguísticos a ponto de se delinear estágios gerais que compõem essa atividade e estratégias[3] específicas para cada um desses estágios, a sumarização multidocumento ainda não foi investigada de forma sistemática. Essa atividade consiste em selecionar conteúdo de uma coleção de textos-fonte que tratam de um mesmo assunto, advindos de fontes distintas, e apresentá-la na forma de um texto coeso e coerente.

Apesar de pouco intuitiva, essa atividade é relativamente comum em alguns cenários (MANI, 2001). O "*clipping*" é um exemplo de sumarização multidocumento realizada por humanos. Essa atividade é bastante antiga na área da comunicação e se caracteriza pela pesquisa e seleção contínua de notícias relacionadas a determinados assuntos, atendendo a um público direcionado (TEIXEIRA, 2001).

Os *clippings* de mídia impressa, enquanto resultados do processo, podem ser veiculados nas formas: (i) clássica, ou seja, por meio de um conjunto de recortes de notícias, reportagens, artigos, etc., (ii) sinopse, isto é, por meio de um texto que contempla as

---

[3] Além das estratégias de seleção, algumas estratégias de produção dos sumários também foram identificadas com base nos sumários produzidos pelos sumarizadores profissionais, as quais caracterizam a reescrita do texto-fonte. Jing e Mckeown (1999) identificaram 6 operações de "recorta e cola" (do inglês, *cut-and-paste*) do texto-fonte: (a) redução sentencial (do inglês, *sentence reduction*), (ii) combinação sentencial (do inglês, *sentence combination*); (iii) transformação sintática (do inglês, *syntactic transformation*); (iv) paráfrase lexical (do inglês, *lexical paraprhasing*); (v) generalização/especificação (do inglês, *generalisation/specification*) e (vi) reordenação (do inglês, *reordering*). Hasler (2007) identificou 5 classes de operações divididas em 2 tipos: (a) operações atômicas, como deleção (do inglês, *deletion*) e inserção (do inglês, *insertion*), e (b) complexas, como substituição (do inglês, *replacement*), reordenação (do inglês, *reordering*) e amálgama (do inglês, *merging*).

principais notícias de interesse do cliente, ou (iii) análise, ou seja, por meio de um texto que contém uma interpretação crítica das informações coletadas (TEIXEIRA, 2001).

No caso do *clipping* eletrônico (*e-clipping* ou *web clipping*), a atividade consiste em selecionar, coletar e organizar as informações veiculadas por diversas fontes da *web* a respeito de determinado tópico, pessoa, instituição etc. Os resultados da seleção podem ser organizados e veiculados não só nas mesmas formas do *clipping* de mídia impressa, mas também nos formatos de: (i) lista de *hyperlinks*, em que cada *link* leva a um *site* ou documento específico, e (iii) lista de excertos de documentos. Atualmente, várias empresas têm oferecido o serviço de *clipping* eletrônico, como a Associação Brasileira das Empresas de Monitoramento de Informação (ABEMO)[4], o Grupo Info4[5], o Armazém Digital[6], entre outras.

Além dessas empresas prestadoras de serviços, inúmeras instituições e associações também oferecem *clippings* eletrônicos a seus associados e/ou consulentes, como a Universidade Federal de São Carlos (UFSCar)[7], Universidade de São Paulo (USP)[8], Associação dos Advogados de São Paulo (AASP)[9], entre outras.

No mercado editorial, os sumários multidocumento constituem, por exemplo, as introduções de coletâneas de artigos e de livros, nas quais as informações principais de cada artigo ou capítulo são fornecidas aos leitores.

Além de ser influenciada pelos mesmos fatores que afetam a tarefa monodocumento, a produção de um sumário a partir de uma coleção de textos que tratam de um mesmo assunto engloba uma seleção de conteúdo que também é afetada por outros fenômenos, como (i) informações redundantes, complementares e contraditórias, (ii) estilos de escrita variados, (iii) eventos/fatos em tempos distintos e (iv) perspectivas e focos também diferentes.

Diante desses fenômenos típicos do cenário multidocumento, os humanos discordam ainda mais sobre as informações a serem selecionadas para os sumários em relação à tarefa monodocumento. Apesar disso, há indícios de que eles também concordam sobre a informação principal, delimitando-a com base principalmente na redundância do conteúdo na coleção (MANI, 2001; NENKOVA, 2006).

---

[4] http://www.abemo.org/
[5] http://www.info4.com.br/info4/novosite/site2/
[6] http://www.adigital.com.br/
[7] http://www.ccs.ufscar.br/clipping
[8] http://www.usp.br/agen/clip/pdf/
[9] http://www.aasp.org.br/aasp/imprensa/clipping/index.asp

Diante da ampla investigação e sistematização das estratégias de sumarização humana monodocumento, métodos linguisticamente motivados de sumarização automática (SA) têm sido desenvolvidos pelos pesquisadores da área do Processamento Automático das Línguas Naturais (PLN), em especial, da subárea denominada Sumarização Automática (SA) (GUPTA; LEHAL, 2010).

Apesar de a estratégia de seleção não ser o único fator que garante a qualidade de um sumário automático, reconhece-se que a investigação dos sumários humanos permitiu a identificação de estratégias de seleção de conteúdo que vêm sendo utilizadas em métodos de SA, tornando-os mais linguisticamente motivados.

Nos métodos mais simples de SA monodocumento, as estratégias humanas são traduzidas em atributos linguísticos superficiais, os 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 área de SA, Baxendale (1958), por exemplo, propôs um método em que um sumário é gerado a partir da seleção das sentenças localizadas no início e final dos parágrafos do seu respectivo texto-fonte. Vale ressaltar que os métodos automáticos de seleção de conteúdo podem se basear em um ou vários atributos linguísticos, ou seja, em uma ou mais estratégias.

Nos métodos mais sofisticados, especificamente aqueles que se baseiam em uma modelagem discursiva do texto-fonte, a seleção é feita por meio da correlação entre as estratégias humanas e a modelagem discursiva. Por exemplo, ao se modelar um texto-fonte de acordo com a teoria *Rhetorical Structure Theory* (RST) (MANN; THOMPSON, 1987), gera-se uma árvore retórica em que as unidades de conteúdo (p.ex.: sentenças) são representadas por nós e as relações semântico-discursivas (p.ex.: *Circumstance, Background, Justify, Concession,* etc.) entre as unidades são representadas por arestas entre os nós. Quando da SA de um texto-jornalístico, a primeira sentença é geralmente a mais nuclear em uma árvore RST bem construída do mesmo texto e, por isso, comumente selecionada para compor o sumário. Nesse caso, por uma árvore RST codificar conhecimento semântico-discursivo, a localização no topo dessa árvore é tida como um atributo profundo da sentença.

Para a proposição de métodos de SA multidocumento (SAM), os pesquisadores do PLN contam apenas com indícios sobre o processo de sumarização humana multidocumento (SHM), principalmente com o indício de que a seleção de conteúdo se baseia na redundância.

Nos métodos mais simples de SAM, os quais, assim como os de SA monodocumento, geram extratos informativos e genéricos, a redundância é traduzida para o atributo estatístico "frequência". Especificamente, a seleção das sentenças dos textos-fonte é feita em função das

palavras de conteúdo mais recorrentes na coleção (SPARCK JONES, 1999). No caso, as sentenças constituídas pelas palavras mais frequentes são selecionadas para o sumário.

Nos métodos mais sofisticados, a redundância é correlacionada à modelagem discursiva dos textos-fonte de uma coleção, como é feito nos métodos em que as sentenças dos textos-fonte são conectadas, quando pertinente, por meio das relações previstas na teoria *Cross-document Structure Theory* (CST) (RADEV, 2000). Especificamente, as sentenças advindas de textos distintos são conectadas por relações como *Identity*, *Equivalence*, *Subsumption*, entre outras (as quais, aliás, buscam capturar os fenômenos linguísticos típicos da multiplicidade de textos-fonte), sendo que a redundância é capturada pelo atributo profundo: número de relações CST que as sentenças possuem com as demais da coleção. No caso, as sentenças que possuem maior número de relações são classificadas como relevantes e, portanto, selecionadas para o sumário.

Diante do cenário exposto sobre a SHM, identificaram-se algumas lacunas de pesquisa e formularam-se objetivos e hipóteses, os quais são apresentados na sequência.

1.2 Lacunas, objetivos e hipóteses

Apesar do avanço pelo qual a SAM tem passado nos últimos anos, destaca-se que os pesquisadores do PLN dispõem apenas de indícios sobre a SHM para subsidiar o desenvolvimento de métodos e sistemas.

Isso se deve ao fato de que não há conhecimento sistematizado sobre a SHM, já que as características dos sumários multidocumento produzidos por humanos ainda não foram amplamente descritas e analisadas.

Diante desse cenário, traçaram-se 2 objetivos: (i) caracterizar um *corpus* composto por sumários humanos multidocumento em português, descrevendo atributos linguísticos superficiais e profundos, e (ii) identificar, formalizar e avaliar estratégias de seleção de conteúdo a partir da caracterização em (i), de tal forma que estas possam subsidiar métodos automáticos de seleção de conteúdo linguisticamente motivados para a SAM.

Os objetivos formulados motivaram em parte a proposição do projeto denominado SUSTENTO (FAPESP 2012/13246-5/ CNPq 483231/2012-6), que objetiva gerar conhecimento linguístico que possa subsidiar o enriquecimento de métodos existentes e/ou a

proposição de novos métodos, principalmente no que tange ao processamento do português. Os resultados do projeto SUSTENTO, que foram incluídos posteriormente, poderão ser utilizados em outro projeto, denominado SUCINTO[10] (FAPESP 2012/03071-3), cujo objetivo é produzir recursos, ferramentas e sistemas de SA que, além da contribuição científica, possam ser disponibilizados para uso de pesquisadores e usuários finais. Ambos estão sendo desenvolvidos no NILC[11].

Para especificar os objetivos desta pesquisa, formularam-se 4 hipóteses a respeito da SHM: (i) há atributos linguísticos que são utilizados com frequência para selecionar o conteúdo principal de dada coleção de textos, advindos de fontes distintas, que abordam um mesmo assunto, (ii) esses atributos, por serem utilizados com recorrência, revelam estratégias de seleção de conteúdo na SHM e (iii) a localização e a redundância são atributos recorrentes que, por isso, confirmam-se como estratégias.

1.3 Metodologia

Para alcançar os objetivos traçados, tomou-se como ponto de partida a metodologia genérica apresentada por Dias-da-Silva (1996, 2006), na qual os sistemas de PLN são vistos como "sistemas especialistas" (do inglês, *expert systems*) ou "sistemas baseados em conhecimento" (do inglês, *knowledge-based systems*).

Segundo essa concepção, a construção de um sistema de PLN, ou parte dele, envolve uma "engenharia do conhecimento linguístico", que é equacionada em função das etapas previstas por Hayes-Roth (1990) para o desenvolvimento dos sistemas especialistas, a saber: "extração do solo" (isto é, explicitação dos conhecimentos e habilidades), "lapidação" (isto é, representação formal desses conhecimentos e habilidades) e "incrustação" (isto é, o programa de computador que codifica essa representação).

Especificamente, Dias-da-Silva (1996), com base em Hayes-Roth, propôs uma metodologia que decompõe a construção de um sistema, ferramenta (p.ex.: um analisador sintático) ou recurso (p.ex.: as bases de conhecimento lexical) em um conjunto de atividades

[10] A página eletrônica do projeto SUCINTO (*Summarization for Clever Information Acess*) está disponível em: http://www.icmc.usp.br/~taspardo/sucinto/.

[11] A página eletrônica do NILC (Núcleo Interinstitucional de Linguística Computacional) está disponível em: http://www.nilc.icmc.usp.br/nilc/.

sucessivas e complementares, agrupadas, segundo sua natureza, em três domínios: o linguístico, o representacional e o implementacional.

No domínio linguístico, as atividades ficam concentradas na investigação dos fatos da língua natural em diferentes dimensões (morfológica, sintática, semântico-conceitual e até mesmo pragmático-discursiva) de acordo com a especificidade do sistema, ferramenta ou recurso que se queira desenvolver.

No domínio representacional, por sua vez, estudam-se modelos formais de representação para os conhecimentos reunidos no domínio linguístico que sejam computacionalmente tratáveis.

E, por fim, no domínio implementacional, as atividades ficam concentradas nas questões relativas à implementação do sistema de PLN.

Para a realização da pesquisa aqui apresentada, as tarefas do domínio linguístico consistiram em (i) a seleção e anotação de *corpus*, (ii) a caracterização dos sumários multidocumento e (iii) a identificação de estratégias humanas de seleção de conteúdo.

Quanto ao domínio representacional, ressalta-se a formalização (ou seja, representação explícita e não-ambígua) e validação das estratégias humanas de seleção de conteúdo identificadas no domínio linguístico.

1.4 Estrutura da dissertação

Em termos formais, esta dissertação organiza-se, além desta introdução, em 7 capítulos.

Tendo em vista o objetivo de caracterizar sumários, identificar, formalizar e avaliar estratégias de sumarização humana multidocumento, apresenta-se, no Capítulo 2, uma revisão da literatura sobre: (i) sumarização humana e automática, mono e multidocumento, (ii) os *corpora* como fonte de conhecimento para a SA e, (iii) avaliação de estratégias/métodos de SA.

No Capítulo 3, relatam-se as tarefas de seleção e anotação do *corpus* sob pesquisa. A anotação consistiu na indexação ou alinhamento dos textos-fonte das coleções do *corpus* a seus respectivos sumários humanos.

No Capítulo 4, apresenta-se a caracterização dos sumários humanos multidocumento do *corpus* selecionado no Capítulo 3. Especificamente, com base na indexação realizada no Capítulo 3, as sentenças dos textos-fonte alinhadas (e não-alinhadas) aos sumários foram descritas em função dos atributos selecionados na literatura, de tal forma que as sentenças dos sumários foram, assim, caracterizadas indiretamente.

No Capítulo 5, descreve-se o processo de identificação e formalização de estratégias de seleção de conteúdo. Esses processos foram realizados a partir da caracterização dos sumários realizada no Capítulo 4.

No Capítulo 6, apresenta-se o processo de avaliação das estratégias identificadas no Capítulo 5.

Finalmente, no Capítulo 7, apresentam-se as considerações finais sobre este trabalho, destacando-se contribuições e trabalhos futuros.

2 REVISÃO DA LITERATURA

Na Seção 2.1, apresenta-se uma revisão sobre o processo humano de sumarização, sobretudo, sobre as estratégias e indícios a respeito da seleção de conteúdo nos cenários mono e multidocumento. As estratégias/indícios de como os humanos selecionam a informação, seja a partir de um ou de mais textos-fonte, permitiram a identificação de certos atributos linguísticos dos textos-fontes que são relevantes para a seleção realizada por humanos, os quais foram descritos na etapa de caracterização dos sumários.

Nas Seções 2.2 e 2.3, apresentam-se noções gerais de SA e de SAM, respectivamente. Tais noções são relevantes porque os trabalhos que investigaram a sumarização humana estão direta ou indiretamente relacionados às aplicações computacionais. Além disso, a SAM foi motivação para a realização deste trabalho e cenários para possível aplicação de seus resultados.

2.1 O processo humano de sumarização

Como mencionado na Introdução, a sumarização monodocumento pode ser entendida como o processo de seleção de conteúdo de um texto-fonte para produzir uma versão mais curta do mesmo visando determinado usuário/tarefa (MANI; MAYBURY, 1999).

Vários autores têm buscado compreender e sistematizar o modo como os humanos geram versões condensadas a partir de um único documento.

De modo geral, reconhece-se que a sumarização humana envolve alguns subprocessos (CREMMINS, 1996; ENDRE-NIGGEMEYER, 1998). Cremmins (1996), especificamente, identificou que os profissionais produzem sumários acadêmicos em 4 estágios, os quais são detalhados no Quadro 1, elaborado com base em Mani (2001).

Quadro 1 - Estágios da sumarização humana

Estágio	Técnica	Resultado
Identificação das principais características (do inglês, *features*) do texto.	Classificar a forma e o conteúdo dos materiais.	Identificação do tipo de sumário a ser produzido, o tamanho e o nível de dificuldade.
Seleção, ou seja, identificação da informação pertinente a ser levada para o sumário.	(a) Identificar informações relevantes do texto-fonte, buscando por palavras-chave ou expressões sinalizadoras, palavras do título e sentenças topicais. (b) Expandir a busca baseada nos resultados em (a).	Identificação de uma porção de informação relevante para posterior extração.
Extração, organização e redução da informação relevante selecionada.	Organizar e escrever as informações relevantes extraídas na etapa anterior na forma de um sumário, seguindo um formato final padrão.	Preparação de um sumário conciso e unificado, porém não editado.
Refinamento da informação relevante.	O autor do sumário ou um revisor técnico deve editar ou revisar o sumário produzido.	Finalização do sumário, sendo ele informativo ou indicativo.

Fonte: Cremmins (1996) (tradução nossa)

Estágios similares aos identificados por Cremmins (1996) também foram verificados por Endres-Niggemeyer (1998) ao analisar os protocolos de sumarização (ou seja, registros que evidenciam o passo a passo da produção textual gerados por 6 profissionais). Com base nos procedimentos desses profissionais, a autora descreve os objetivos de cada um dos estágios, os quais estão descritos no Quadro 2.

Quadro 2 - Etapas humanas de sumarização

Etapa	Objetivo	Resultado
Exploração do documento	Identificar as características básicas do material: examinar o título do documento, seu perfil, a disposição e formato das informações, a estrutura global e o gênero do documento, familiarizar-se com o conteúdo, etc.	Contrução do "esquema" (do ingles, *scheme*), isto é, de um conhecimento inicial sobre o tipo de documento e estrutura da informação.
Avaliação da relevância	Identificar no texto unidades textuais relevantes, representar o texto em nível discursivo e combinar os elementos da etapa anterior, associando-os aos elementos da etapa em questão.	Construção do "tema" (do inglês, *theme*), ou seja, de uma representação mental estruturada sobre o conteúdo do texto.
Produção do sumário	Transportar itens textuais relevantes do documento original ao sumário e reorganizá-los em uma nova estrutura, envolvendo operações de "recorta e cola" (do inglês, *cutting and pasting operations*)	Produção de um texto condensado a partir dos pontos mais importantes de um documento.

Fonte: Endres-Niggemeyer (1998) (tradução nossa).

Com relação ao estágio denominado “avaliação de relevância” nas tarefas de sumarização mono e multidocumento realizadas pelos humanos, algumas estratégias de seleção de conteúdo foram identificadas.

A seguir, apresentam-se as estratégias mais difundidas na literatura.

2.1.1 As estratégias de seleção de conteúdo

Uma das estratégias é a seleção de informação (comumente, sentenças) que se relaciona com as palavras contidas no título/subtítulo do texto-fonte. Essa estratégia, especificamente, pressupõe a identificação das palavras que compõem o título, subtítulo e tópicos para selecionar trechos que contenham as ideias principais do texto.

No caso, a existência de subtítulos e tópicos depende do tipo/gênero e do tamanho do texto a ser sumarizado. Em textos jornalísticos, por exemplo, é comum a presença de um título e subtítulos (CREMMINS, 1996; ENDRES-NIEGGEMEYER, 1998). A hipótese de que palavras do título e subtítulo são relevantes foi estatisticamente aceita com 99% de significância (EDMUNDSON, 1969).

Além de selecionar conteúdo a partir das palavras do título/subtítulos, os humanos também selecionam informação por meio da identificação de expressões-chave ou expressões-indicativas no texto-fonte (CREMMINS, 1996; ENDRES-NIEGGEMEYER, 1998).

Nessa estratégia, o escritor do sumário pauta-se no tipo/gênero textual, já que textos-fonte de tipos/gêneros diferentes são compostos por expressões também diferentes. Diz-se que tais expressões são “indicativas” porque estas indicam a expressão de certos conteúdos que caracterizam componentes específicos da estrutura discursiva dos gêneros.

Um texto científico, por exemplo, enquadra-se, em geral, dentro de um esquema que já se tornou clássico por abranger aspectos essenciais de um texto desse gênero. Essa estrutura é composta pelos componentes “resumo”, “introdução”, “materiais/métodos”, “resultados”, “discussão” e “conclusão”, os quais são introduzidos nos textos por certas expressões, que, para a seleção de conteúdo, funcionam como pistas; a expressão “o objetivo deste trabalho é”, por exemplo, indica a expressão do conteúdo “meta/objetivo”.

Quanto à macroestrutura dos textos, aliás, Liddy (1991) verificou que, no caso de textos acadêmicos, as informações comumente selecionadas pelos escritores são provenientes

das seções que expressam os componentes *contexto, objetivo, metodologia, resultado* e *conclusão.*

Outra estratégia é a de selecionar a informação localizada em certas posições dos textos. Novamente, o tipo/gênero dos textos influencia a seleção da informação com base no atributo localização. Alguns gêneros específicos possuem uma estrutura mais padronizada que outros, sendo possível identificar informações principais com mais facilidade.

Segundo Endres-Nieggemeyer (1998), as informações expressas no início e final dos parágrafos de um texto-fonte científico são importantes para a seleção de conteúdo. No caso dos textos do gênero jornalístico, as informações mais importantes selecionadas por humanos para compor os sumários são expressas no início do texto.

Isso se deve ao fato de que tais textos são compostos pelos componentes (i) título, (ii) *lead*, o qual corresponde ao primeiro ou aos dois primeiros parágrafos do texto, e (iii) corpo, que abrange os demais parágrafos, os quais desenvolvem os elementos informativos referidos no *lead*, o qual é frequentemente a informação principal expressa com o intuito de instigar o leitor. Ao contrário das narrativas literárias ou cinematográficas, nos textos jornalísticos o clímax não se guarda para o fim. Logo no início, o leitor tem o essencial da informação. O último parágrafo do corpo do texto evidencia uma informação suficientemente interessante para não desapontar o leitor frente à expectativa criada pelo *lead.* Essa estrutura típica dos textos jornalísticos recebe a denominação de "pirâmide invertida" (LAGE, 2002).

Por expressar o conteúdo principal de um texto jornalístico, a informação localizada no início é selecionada para compor o sumário.

Todas as estratégias descritas até o momento foram sistematizadas no cenário da sumarização humana monodocumento, ou seja, são frutos de trabalhos que investigaram o processo humano de produção de sumários.

Outra estratégia também foi delineada em trabalhos que objetivavam efetivamente o desenvolvimento de sumarizadores automáticos que, ao compararem os sumários automáticos a manuais, evidenciaram a relevância de certas características textuais.

Assim, além de selecionar conteúdo com base nas palavras do título/subtítulo, expressões indicativas, localização e componentes discursivos, os humanos, ao produzirem sumários extrativos, também selecionam informação com base no tamanho ou extensão (em número de palavras) das sentenças dos textos-fonte (KUPIEC et al., 1995). Em geral, os humanos selecionam sentenças não muito longas, nem muito curtas para compor um sumário.

A seleção de conteúdo com base nas palavras-chave foi identificada como estratégia na sumarização humana monodocumento (p.ex.: LHUN, 1958) e também na SHM (NENKOVA, 2006).

Nenkova (2006), em particular, identificou que os humanos comumente selecionam as sentenças que contêm as palavras mais frequentes de uma coleção para compor seu respectivo sumário, evidenciando que a frequência de ocorrência das palavras é uma característica textual (superficial) relevante no processo de seleção de conteúdo.

Especificamente, Nenkova (2006) utilizou o *corpus* fornecido pela conferência internacional DUC´2003[12] (*Document Understanding Conferences*). Esse *corpus* possui 30 coleções, cada uma composta por 10 textos jornalísticos compilados do jornal *The New York Times*. Para cada coleção, os 10 sujeitos do experimento produziram um sumário multidocumento formado por aproximadamente 100 palavras.

Desse experimento, a autora comprovou que em média 94,66% das palavras mais frequentes de uma coleção estão presentes no respectivo sumário. Considerando-se as 12 palavras mais frequentes de uma coleção, estas ocorreram em média 85,25% no sumário correspondente.

A identificação de conteúdo relevante com base na macroestrutura discursiva dos textos se mostra relevante nas atividades humanas de sumarização mono e multidocumento (CREMMINS, 1996, ENDRES-NIEGGEMEYER, JAIDKA, 2010).

Quanto à SHM, Jaidka et al. (2010), em particular, identificaram que na produção de sumários multidocumento de textos científicos os humanos selecionam preferencialmente as informações contidas nas seções “resumo” (ou *abstract*), “conclusão” e “metodologia”.

Outras duas maneiras de se identificar as informações mais importantes para compor os sumários são indícios específicos da SHM. Dessa forma, Mani (2001) observou 2 indícios de estratégia de seleção de conteúdo multidocumento.

Segundo o autor, os humanos escolhem um texto de sua preferência como base para selecionar as informações principais e, na sequência, recorrem aos demais textos da coleção para complementar as informações do sumário. Para a escolha do texto-fonte “base”, diversos fatores podem influenciar, como a (i) data de publicação, isto é, o sumarizador pode partir do texto mais recente ou mais antigo, dependendo de seu interesse; (ii) prestígio do veículo de

[12] http://duc.nist.gov/duc2003/tasks.html

comunicação, ou seja, o autor do sumário pode identificar o texto-fonte de partida pela relevância do veículo no qual o texto foi publicado; (iii) autoria; (iv) textualidade (coesão e coerência), etc.

Além de selecionar um texto-fonte "base", Mani (2001) observou que a seleção da informação no cenário multidocumento está pautada na redundância, ou seja, a informação mais repetida entre os textos de uma coleção que trata de um mesmo assunto é escolhida como "principal".

No Quadro 3, as principais estratégias (comprovadas) e indícios encontrados na literatura estão sistematizados, elucidando-se o cenário de sumarização no qual foram investigados, o nível do conhecimento linguístico envolvido nas estratégias e a autoria original da identificação.

Quadro 3 - As principais estratégias de sumarização humana

Sumarização humana	**Estratégia/Indício**	**Nível**	**Autor**
Monodocumento	Selecionar conteúdo com base no tamanho da sentença	Superficial	Kupiec et al. (1995)
	Selecionar conteúdo com base nas palavras mais frequentes da coleção	Superficial	Luhn (1958)
	Selecionar conteúdo com base nas palavras do título/subtítulo	Superficial	Edmundson (1969), Cremmins (1996), Endress-Niggemeyer (1998), Lage (2002)
	Selecionar conteúdo com base em expressões sinalizadoras	Superficial	
	Selecionar conteúdo com base na posição que ocupa no texto/parágrafo	Superficial	
	Selecionar conteúdo com base na macroestrutura discursiva do texto	Profundo	Liddy (1991)
Multidocumento	Selecionar conteúdo com base nas palavras mais frequentes da coleção	Superficial	Nenkova (2006)
	Selecionar conteúdo com base na macroestrutura discursiva do texto	Profundo	Jaidka (2010)
	Selecionar conteúdo com base na redundância	Profundo	Mani (2001)
	Selecionar um dos textos-fonte como "base"	Linguístico ou Extralinguístico	

Fonte: Elaborado pelo autor

A seguir, apresentam-se algumas noções básicas do processo automático de sumarização, isto é, os processos de um sistema de SA e os parâmetros que influenciam essas etapas.

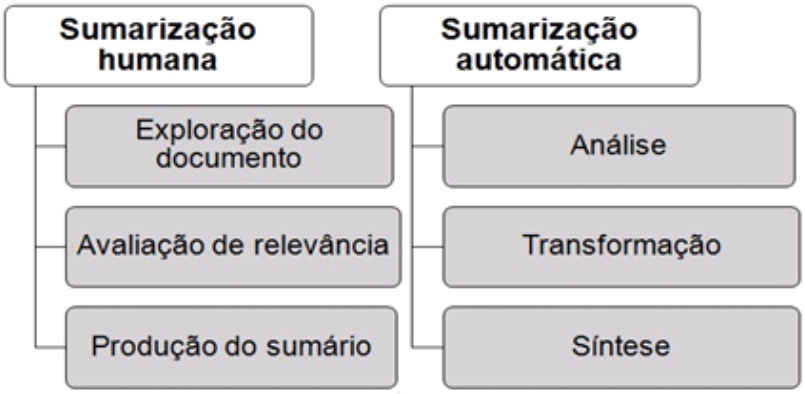
2.2 Noções básicas de Sumarização Automática

O PLN é uma área multidisciplinar que busca desenvolver sistemas capazes de realizar tarefas linguísticas específicas, como a correção ortográfica e gramatical, a tradução, a extração de informação, a sumarização automática (SA), entre outras.

Na SA, subárea do PLN, busca-se automatizar a produção de sumários a partir de um ou mais documentos, ou seja, a geração de versões condensadas de um ou mais textos (MANI, 2001). Os sistemas que realizam tal tarefa são denominados sumarizadores automáticos.

Buscando emular na máquina os processos humanos, Mani e Maybury (1999) sugerem que a SA envolva idealmente os 3 processos: (i) análise dos textos-fonte, (ii) transformação e (iii) síntese. O paralelo entre as etapas humanas e automáticas de sumarização é apresentado na Figura 1.

Figura 1 - Etapas de sumarização humana e automática



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

A primeira delas é a “análise”, que corresponde à interpretação dos textos-fonte e gera uma representação do conteúdo linguístico expresso em termos computáveis.

A segunda é a “transformação”, etapa em que o conteúdo formalizado dos textos-fonte é selecionado e condensado em uma representação computável, ou seja, não-textual. O ponto central da seleção de conteúdo é reconhecer as unidades de significado do texto-fonte (p.ex.: palavras, sintagmas, orações, sentenças, etc.) que contêm as ideias centrais do mesmo para compor o sumário (MANI, 2001).

A terceira (e última) etapa é a “síntese”. Nela, o conteúdo condensado é expresso em língua natural na forma de um sumário. Para tanto, métodos de justaposição, ordenação, fusão

e correferenciação dos segmentos textuais selecionados podem ser utilizados (SPARCK JONES, 1993).

Tais etapas são guiadas pela taxa de compressão, ou seja, o tamanho desejado do sumário; um sumário com taxa de compressão de 70% apresenta tamanho equivalente a 30% do tamanho do texto original (em geral, medido em número de palavras).

Quanto ao número de textos-fonte sob processamento, a SA pode ser monodocumento ou multidocumento.

Na SA monodocumento, produz-se um sumário de um único texto-fonte. Na SAM, produz-se um sumário a partir de uma coleção de textos-fonte que abordam um mesmo tópico (MCKEOWN; RADEV, 1995; MANI, 2001).

A SA monodocumento é uma aplicação "tradicional" do PLN, sendo muito explorada e discutida por inúmeros autores há várias décadas (p.ex.: LUHN, 1958; EDMUNDSON, 1969; O'DONNELL, 1997a; SALTON et al., 1997; MARCU, 2000; PARDO; RINO, 2002; PARDO et al., 2003; RINO et al., 2004; UZÊDA et al., 2010; CLARKE; LAPATA, 2010; LOUIS et al., 2010; entre outros).

O interesse pela SAM, como se vê com mais detalhes na próxima seção, é mais recente e tem se fortalecido com o aumento do volume de informação disponível na *web* e pelo pouco tempo que os usuários têm para absorvê-la.

Assim como na sumarização humana, vários parâmetros influenciam a realização das etapas de SA. Além da taxa de compressão, a qual determina o quanto a informação do texto-fonte deve ser condensada e transposta para o sumário, a audiência determinará a produção automática de um sumário genérico ou focado nos interesses do usuário e a função determinará a geração de sumários informativos, indicativos ou críticos.

Um sumário genérico não considera um usuário específico e, por isso, ele simplesmente oferece a informação mais relevante contida no(s) texto(s)-fonte. Por outro lado, os sumários focados no interesse do usuário, ou especializados, são produzidos a partir de uma consulta e buscam englobar as informações que satisfaçam essa consulta (JURAFSKY, 2007). A sumarização genérica é feita de forma mais direta, ou seja, modela-se o documento, dá-se importância aos segmentos textuais e, por fim, selecionam-se os mais salientes para compor o sumário final. Na sumarização focada, esses processos são feitos de forma diferente. Quando se parte de uma palavra-chave, por exemplo, a pontuação dos segmentos é feita em função dela. Assim, os segmentos textuais que contiverem todo ou parte do requisito da consulta serão levados em consideração para uma seleção posterior mais cuidadosa.

O sumário informativo, ou autocontido, contém as informações principais de um texto-fonte de forma coerente e coesa ao ponto de dispensar a leitura do original. O indicativo, ou indexador, não substitui o texto-fonte, apenas diz do que ele trata, sendo utilizado, por exemplo, quando o leitor deseja fazer uma busca por um documento referente a determinado tópico, decidindo qual documento merece mais atenção. O sumário crítico, ou avaliativo, apresenta, além da informação principal do texto-fonte, avaliações sobre ele (MANI; MAYBURY, 1999).

A SA, além de considerar a audiência e a função do sumário, também pode fazer uso de diferentes níveis de conhecimento linguístico, os quais caracterizam suas abordagens (MANI, 2001).

Caso a escolha seja pela utilização de pouco ou nenhum conhecimento linguístico, a abordagem de SA é dita superficial, pois o conhecimento que se utiliza é empírico/estatístico. Por exemplo, uma abordagem que produz um sumário a partir da seleção e justaposição das sentenças do texto-fonte que apresentam as palavras mais frequentes do texto é classificada como superficial.

Caso a opção seja pelo uso de teorias ou modelos linguísticos, a abordagem é classificada como profunda (SPARCK JONES, 1999; MANI, 2001).

A abordagem, aliás, determina a formação do sumário a ser gerado. A partir do processo de SA superficial, originam-se necessariamente extratos (ou seja, sumários compostos por trechos inalterados dos textos-fonte), ao passo que, a partir da sumarização profunda, podem-se produzir extratos ou *abstracts* (isto é, sumários desenvolvidos a partir da manipulação linguística dos textos-fonte, os quais sofrem operações de reescrita).

Apesar das inúmeras pesquisas em SA, a qualidade dos sumários automáticos ainda deixa a desejar, principalmente quanto à coesão/coerência e informatividade.

Na sequência, discorre-se com mais detalhes sobra a SAM.

2.3 A Sumarização Automática Multidocumento

A SAM estabeleceu-se como uma aplicação de destaque no PLN em resposta à demanda por novas formas/tecnologias de gerenciamento do enorme volume de informação que está em constante crescimento e mudança na *web* e cujo processamento na íntegra é praticamente impossível para o humano, principalmente pela falta de tempo frente à rotina agitada do cotidiano (MANI, 2001).

Especificamente, a SAM iniciou nos anos de 1990 (p.ex.: MCKEOWN; RADEV, 1995; RADEV; MCKEOWN, 1998; CARBONELL; GODLSTEIN, 1998) e tem adquirido relevância nos últimos anos (p.ex.: RADEV et al., 2000; ZHANG et al., 2002; OTTERBACHER et al., 2002; MCKEOWN et al., 2005; NENKOVA, 2005a, 2005b; WAN; YANG, 2006; AFANTENOS et al., 2004, 2008; WAN, 2008; HAGHIGHI; VANDERWENDE, 2009; CASTRO JORGE; PARDO, 2010, 2011; CELIKYILMAZ; HAKKANI-TUR, 2011, entre outros).

Para o português, as pesquisas em SAM são incipientes, iniciando-se oficialmente em 2005 com uma extensão bastante simples do sistema superficial monodocumento denominado GistSumm (PARDO, 2005).

As primeiras investigações realmente relevantes datam de 2011 e consistem nos resultados do projeto SUCINTO, já mencionado anteriormente. Desse projeto, resultam, especificamente, os sistemas:

(i) CSTSumm (CASTRO JORGE; PARDO, 2010), o qual se baseia na teoria/modelo linguístico-computacional CST,

(ii) ACSumm (CASTRO JORGE; AGOSTINI; PARDO, 2011), baseado na abordagem de Aprendizado de Máquina (AM),

(iii) RSumm (RIBALDO; PARDO; RINO, 2011) e (iv) CNSumm (AKABANE et al., 2012), os quais se pautam em grafos e redes complexas.

De fato, o CSTSumm apresenta, atualmente, os resultados mais consistentes para o processamento do português.

Além desses sistemas, o SUCINTO também gerou:

(i) CSTNews (CARDOSO et al., 2011a), um *corpus* de referência multidocumento;

(ii) CSTParser (MAZIERO; PARDO, 2011), um *parser* discursivo baseado na teoria CST (RADEV, 2000).

Enquanto que a finalidade da sumarização monodocumento é extrair conteúdo de um documento fonte e apresentar a porção mais importante ao usuário de forma condensada, o objetivo da SAM pode ser caracterizado como uma especificação da primeira.

No caso, o texto condensado a ser gerado deve apresentar a informação mais relevante de uma coleção de textos de mesmo gênero e que discorram sobre um mesmo tópico. Dessa forma, na SAM, é preciso lidar com questões como: redundância, informações complementares, contradição, etc.

Os sumarizadores multidocumento diferem dos monodocumento na quantidade de textos de entrada a serem processados. Assim, a arquitetura genérica de um sumarizador multidocumento é modelada como a ilustrada na Figura 2 (SPARCK JONES,1993).

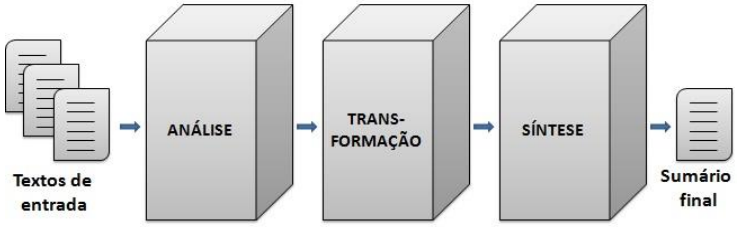


Figura 2 - Arquitetura genérica de um sistema de SAM

Fonte: Sparck Jones (1993).

Nessa arquitetura, a transformação é etapa central, pois, a partir da representação gerada na análise, o conteúdo dos textos-fonte é condensado em uma representação interna do sumário.

A transformação, como mencionado, engloba a seleção do conteúdo que irá compor o sumário. Para tanto, é necessário ranquear os segmentos dos textos-fonte em função de sua relevância e selecionar os de maior pontuação, que devem conter as ideias centrais do texto.

Na sequência, destaca-se o como as estratégias humanas têm subsidiado o processo de seleção de conteúdo nos métodos superficiais e profundos de SAM.

2.3.1 A aplicação das estratégias humanas na SAM

A SAM tem sido guiada pelas estratégias da sumarização humana monodocumento e, sobretudo, pelo indício de SHM, segundo o qual os humanos selecionam as informações redundantes de uma coleção para compor o sumário (KUMAR; SALIM, 2012).

Nos métodos superficiais que se baseiam em atributos linguísticos (do inglês, *feature-based methods*), estes buscam codificar as estratégias mono e/ou indício multidocumento, sendo que esses atributos podem variar em número e combinação (p.x.: LIN; HOVY, 2002; SCHILDER, KONDADADI, 2008) e apresentar pesos diferentes em função do tipo/gênero dos textos-fonte.

Para codificar a redundância, identificada na literatura como o critério segundo o qual os humanos selecionam conteúdo, esses métodos comumente estipulam um atributo superficial do tipo “frequência”. Nesse caso, a análise consiste na segmentação sentencial e no cálculo da frequência de ocorrência de cada palavra dos textos-fonte na coleção. A

transformação consiste em pontuar e ranquear as sentenças em função da soma da frequência de suas palavras constitutivas. Na sequência, as sentenças mais pontuadas apresentam conteúdo redundante/relevante, as quais são selecionadas para o sumário até que se atinja a taxa de compressão.

Para o português, Pardo (2005) desenvolveu o sumarizador superficial GistSumm (PARDO, 2005), cujo método pode ser caracterizado como "baseado em atributo linguístico". O GistSumm pontua e ranqueia as sentenças dos textos de uma coleção com base na frequência de ocorrência de suas palavras na coleção. A sentença de maior pontuação é considerada a *gist sentence* (isto é, sentença que expressa o conteúdo principal da coleção) e selecionada para iniciar o sumário. As demais sentenças que compõem o sumário satisfazem a dois critérios: (i) conter pelo menos um radical em comum com a *gist sentence* e (ii) ter pontuação maior que a média das pontuações de todas as sentenças.

Nos métodos profundos baseados em conhecimento sintático, o indício da redundância também é utilizado. Nesses métodos, os textos-fonte são analisados sintaticamente[13] e as estruturas predicativas (predicado-argumento) similares são agrupadas, pois teoricamente expressam um mesmo tópico (BARZILAY et al., 1999). As estruturas mais recorrentes são selecionadas e reordenadas para gerar as sentenças de sumários abstrativos.

Nos métodos profundos baseados em conhecimento semântico-conceitual, por sua vez, cada texto-fonte é modelado em um grafo, no qual as palavras são representadas por nós e a similaridade distribucional entre elas por arestas. No caso, dois nós com arestas similares representam palavras sinônimas e, portanto, expressam um conceito. Diante dos conceitos mais importantes da coleção, selecionam-se as sentenças que contêm as palavras que os expressam para compor o sumário (MANI et al., 1999).

Outros métodos profundos de SAM utilizam conhecimento semântico-discursivo, os quais primordialmente partem da análise[14] dos textos-fonte com base na teoria/ modelo linguístico-computacional CST (RADEV, 2000). A Figura 3 ilustra um esquema genérico de relacionamento entre as sentenças de textos sobre um mesmo tópico.

---

[13] A análise em questão pode ser feita por um analisar automático sintático (ou *parser*).

[14] A análise é feita por um analisador automático discursivo, como o CSTParser (MAZIERO, PARDO, 2011).

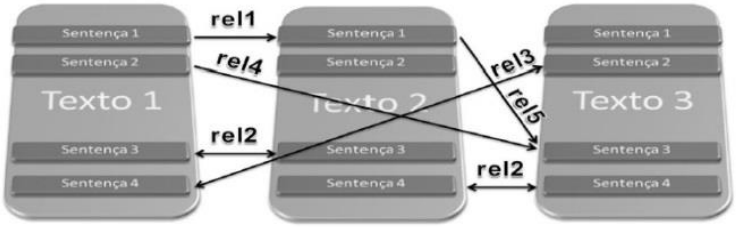


Figura 3 – Esquema genérico de análise multidocumento.
Fonte: Maziero e Pardo (2011).

Inspirada na *Rhetorical Structure Theory* (RST) (MANN; THOMPSON, 1987), a CST permite estruturar o discurso pela conexão das sentenças (ou outras unidades textuais) advindas de diferentes documentos (RADEV, 2000). Originalmente, propôs-se um conjunto de 24 relações intertextual (Quadro 4).

Quadro 4 - 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).

Trabalhos posteriores refinaram as relações, produzindo conjuntos mais compactos (p.ex.: ZHANG et al., 2003; MAZIERO et al., 2010). O Quadro 5, por exemplo, apresenta o conjunto refinado de relações CST proposto por Maziero et al. (2010) e suas respectivas características.

Quadro 5 – Conjunto de relações CST e suas características.

Nome da Relação: *Identity*
Tipo da Relação: Conteúdo->Redundância->Total
Direcionalidade: Nula
Restrições: As sentenças devem ser idênticas
Nome da Relação: *Equivalence*
Tipo da Relação: Conteúdo->Redundância->Total
Direcionalidade: Nula
Restrições: As sentenças apresentam o mesmo conteúdo, mas expresso de forma diferente.

Nome da Relação: *Summary*
Tipo da Relação: Conteúdo->Redundância->Total
Direcionalidade: S1<-S2
Restrições: S2 apresenta o mesmo conteúdo que S1, mas de forma mais compacta.
Comentários: *Summary* é um tipo de *equivalence*, mas *summary* deve haver diferença significativa de tamanho entre as sentenças.
Nome da Relação: *Subsumption*
Tipo da Relação: Conteúdo->Redundância->Parcial
Direcionalidade: S1->S2
Restrições: S1 apresenta as informações contidas em S2 e informações adicionais.
Comentários: S1 contém X e Y, S2 contém X.
Nome da Relação: *Overlap*
Tipo da Relação: Conteúdo->Redundância->Parcial
Direcionalidade: Nula
Restrições: S1 e S2 apresentam informações em comum e ambas apresentam informações adicionais distintas entre si.
Comentários: S1 contém X e Y, S2 contém X e Z.
Nome da Relação: *Historical background*
Tipo da Relação: Conteúdo->Complemento->Temporal
Direcionalidade: S1<-S2
Restrições: S2 apresenta informações históricas/passadas sobre algum elemento presente em S1.
Comentários: O elemento explorado em S2 deve ser o foco de S2; se forem apresentadas informações repetidas, considere outra relação (por exemplo, *overlap*); se os eventos em S1 e S2 forem relacionados, pondere sobre a relação *follow-up.*
Nome da Relação: *Follow-up*
Tipo da Relação: Conteúdo->Complemento->Temporal
Direcionalidade: S1<-S2
Restrições: S2 apresenta acontecimentos que acontecem após os acontecimentos em S1; os acontecimentos em S1 e em S2 devem ser relacionados e ter um espaço de tempo relativamente curto entre si.
Nome da Relação: *Elaboration*
Tipo da Relação: Conteúdo->Complemento->Atemporal
Direcionalidade: S1<-S2
Restrições: S2 detalha/refina/elabora algum elemento presente em S1, sendo que S2 não deve repetir informações presentes em S1.
Comentários: O elemento elaborado em S2 deve ser o foco de S2; se forem apresentadas informações repetidas, considere outra relação (por exemplo, *overlap*); se forem apresentadas informações temporais, pondere sobre a relação *historical background*.
Nome da Relação: *Contradiction*
Tipo da Relação: Conteúdo->Contradição
Direcionalidade: Nula
Restrições: S1 e S2 divergem sobre algum elemento das sentenças .
Nome da Relação: *Citation*
Tipo da Relação: Apresentação/Forma->Fonte/Autoria
Direcionalidade: S1<-S2
Restrições: S2 cita explicitamente informação proveniente de S1.

Comentários: Dada a natureza desta relação, ela não pode co-ocorrer com relações de redundância total.
Nome da Relação: *Attribution*
Tipo da Relação: Apresentação/Forma->Fonte/Autoria
Direcionalidade: S1<-S2
Restrições: S1 e S2 apresentam informação em comum e S2 atribui essa informação a uma fonte/autoria.
Comentários: Dada a natureza desta relação, ela não pode co-ocorrer com relações de redundância total.
Nome da Relação: *Modality*
Tipo da Relação: Apresentação/Forma->Fonte/Autoria
Direcionalidade: S1<-S2
Restrições: S1 e S2 apresentam informação em comum e em S2 a fonte/autoria da informação é indeterminada/relativizada/amenizada
Comentários: Dada a natureza desta relação, ela não pode co-ocorrer com relações de redundância total.
Nome da Relação: *Indirect speech*
Tipo da Relação: Apresentação/Forma->Estilo
Direcionalidade: S1<-S2
Restrições: S1 e S2 apresentam informação em comum; S1 apresenta essa informação em discurso direto e S2 em discurso indireto.
Nome da Relação: *Translation*
Tipo da Relação: Apresentação/Forma->Estilo
Direcionalidade: Nula
Restrições: S1 e S2 apresentam informação em comum em línguas diferentes.

Fonte: Maziero et al. (2010).

Os trechos de notícia do Quadro 6, coletados de fontes distintas, relatam um mesmo acidente aéreo. Neles, algumas relações CST do conjunto de Maziero et al. (2010) podem ser identificadas.

Quadro 6 - Exemplos de relações CST

Texto 1
[1] Ao menos 17 pessoas morreram após a queda de um avião de passageiros na República Democrática do Congo.
[2] Segundo uma porta-voz da ONU, o avião, de fabricação russa, estava tentando aterrissar no aeroporto de Bukavu em meio a uma tempestade.
[3] A aeronave se chocou com uma montanha e caiu, em chamas, sobre uma floresta a 15 quilômetros de distância da pista do aeroporto.

Texto 2
[1] Um acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo (RDC), matou 17 pessoas na quinta-feira à tarde, informou nesta sexta-feira um porta-voz das Nações Unidas.
[2] As vítimas do acidente foram 14 passageiros e 3 membros da tripulação.
[3] 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.

Fonte: http://www2.icmc.usp.br/~taspardo/sucinto/cstnews.html.

Por exemplo, a sentença [1] do texto 1 e a sentença [1] do texto 2 estão ligadas pela relação *Attribution*, pois tais sentenças apresentam informação em comum, sendo que a sentença [1] do texto 2 atribui essa informação a uma fonte/autoria (porta-voz das Nações Unidas). Outra relação entre as mesmas unidades também pode ser identificada. No caso, a relação é a *Subsumption*, já que a sentença [1] do texto 2 apresenta, além do mesmo conteúdo da sentença [1] do texto 1, informações adicionais.

Com relação à seleção de conteúdo, busca-se comumente capturar o indício da redundância com base no número de relações que uma sentença possui com as demais da coleção. Nesses trabalhos, os textos-fonte são modelados em um grafo, em que as sentenças são representadas por nós e as relações CST por arestas. As sentenças são então pontuadas e ranqueadas com base no número de conexões no grafo, sendo que as sentenças com mais conexões ocupam o topo do ranque (MANI, 2001). No Quadro 6, vê-se que a sentença [1] do texto 1 e a sentença [1] do texto [2] estão relacionadas por 2 relações CST (*Attribution* e *Subsumption*), o que indica sobreposição de conteúdo e, consequentemente, maior relevância dos segmentos em detrimento dos demais com menor número de relações.

A seleção também pode ser guiada pelo tipo de relação CST identificada entre as sentenças de textos-fonte distintos. A partir de um conjunto refinado de 14 relações CST, Maziero et al. (2010) propuseram uma tipologia, a qual consta na Figura 4[15].

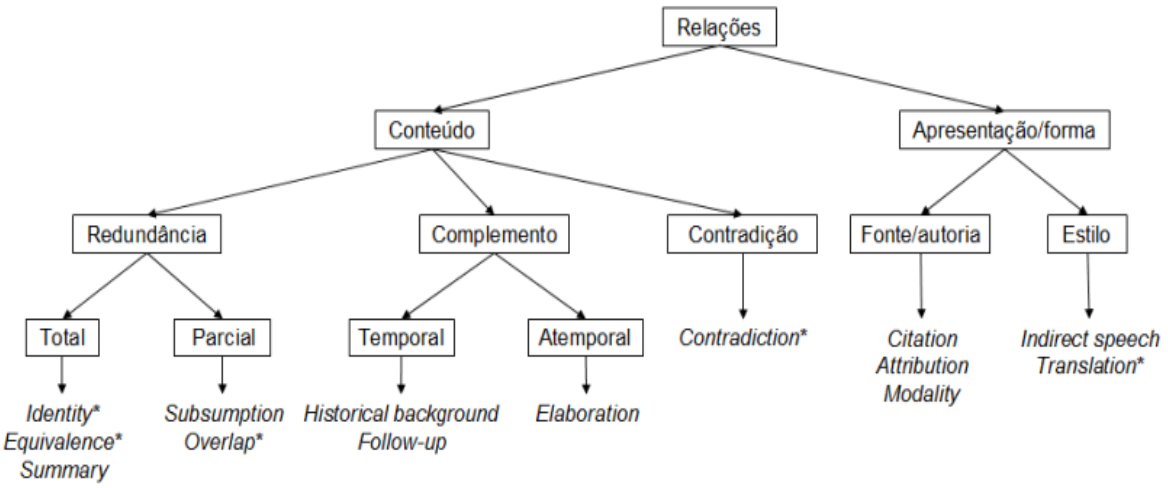


Figura 4 - Tipologia das relações CST.

Fonte: Maziero et al. (2010).

[15] As relações com asterisco não têm direcionalidade.

Essa tipologia classifica as relações em dois grandes grupos: relações de conteúdo (isto é, que ligam o conteúdo das sentenças) e relações de forma (ou seja, relações que ligam sentenças com base na forma). Cada grupo apresenta subdivisões. As relações de conteúdo podem ser classificadas nas categorias "redundância", "complemento" e "contradição". As relações da categoria "redundância", em especial, podem ser parciais ou totais, e as da categoria "complemento" podem ser temporais ou atemporais. As relações de forma, por sua vez, podem ser do tipo "fonte/autoria" ou "estilo".

Com base nessa tipologia, vê-se que das relações identificadas entre a sentença [1] do texto 1 e a sentença [1] do texto 2, *Attribution* é uma relação de forma e *Subsumption,* de conteúdo, em especial, que indica redundância parcial.

Caso o sistema de SAM selecione as sentenças apenas com base no tipo de relação, este pode dar maior pontuação às sentenças relacionadas por relações de conteúdo da categoria "redundância". No caso, a sentença [1] do texto 2 que subsume o conteúdo da sentença [1] do texto 1 é possível candidata a compor o sumário em detrimento de outras com apenas relações de forma, por exemplo. Aparentemente, quanto mais relações do tipo redundância uma sentença possui, mais chances ela tem de ser selecionada na etapa de transformação.

O primeiro trabalho a utilizar relações discursivas na SAM foi o de Radev e McKeown (1998), no qual as bases da CST foram formuladas. Zhang et al. (2002) propuseram a troca de sentenças pouco relacionadas por sentenças que apresentam maior número de relações CST, o que resultou em melhora significativa na qualidade dos sumários.

Para o português, desconhecem-se trabalhos que utilizam informações sintáticas e semântico-conceituais. Os trabalhos para o português utilizam majoritariamente conhecimento semântico-discursivo, codificado nas relações da CST (CASTRO JORGE; PARDO, 2010; CASTRO JORGE; PARDO, 2011).

Quanto a métodos de SAM híbridos, Schiffman et al. (2002) caracterizam-se por unificar informações superficiais (localização da sentença nos textos-fonte e tamanho da sentenças) e conhecimento léxico-conceitual. Além dos atributos superficiais, os autores relacionam as palavras dos textos-fonte pela sinonímia e hiponímia para delimitar os conceitos mais representativos da coleção. As relações são identificadas pela indexação das palavras à WordNet de Princeton (WN.Pr), base léxico-conceitual do inglês norte-americano (FELLBAUM, 1998).

O método de Ribaldo et al. (2012), por sua vez, desenvolvido para o português, explora medidas estatísticas aplicadas a grafos e redes em combinação com as relações CST.

Especificamente, esse método seleciona os nós mais densos para compor o sumário, ou seja, com maior número de conexões com o objetivo de garantir a cobertura dos conceitos principais dos textos.

A seguir, apresentam-se o conceito de *corpus*, suas utilidades para a pesquisa linguística e os vários níveis de anotação de um *corpus*.

2.4 Os *corpora:* fontes de conhecimento para a SA

A Linguística de *Corpus* (LC) pode ser definida como uma área de pesquisa que se ocupa da "coleta e exploração de *corpora*, ou conjunto de dados linguístico-textuais que foram coletados criteriosamente com o propósito de servirem para a pesquisa de uma língua ou variedade linguística" (BERBER SARDINHA, 2004, p.3).

Uma razão importante para o uso de *corpora* em pesquisas é a possibilidade de extração de padrões linguísticos de uma língua, isto é, ocorrências de itens lexicais ou sintagmas que podem ser identificados por meio de um conjunto vasto de textos que caracterizam uma língua ou variedade dela. Em alguns casos, para se extrair as informações a partir de um *corpus*, uma análise linguística precisa ser inserida previamente. O processo de adição de tais informações linguísticas a um *corpus* eletrônico que contém dados de língua falada e/ou escrita é chamado de "anotação de *corpus*" (LEECH, 1997a).

Conforme Aluísio e Almeida (2006) postulam, os fenômenos anotados e as representações usadas variam muito, mas, para que os *corpora* anotados sejam usados em aplicações de PLN, é necessário que os vários esquemas de anotação sejam compatíveis uns com os outros para permitir que essas informações sejam manipuladas por meio de um ambiente computacional único. Essa compatibilização é ainda mais necessária quando as diferentes anotações são realizadas sobre um mesmo *corpus*.

Para a língua portuguesa, existem diversos projetos que realizaram a anotação de *corpus* em diversos níveis, como: (i) a anotação morfossintática do *corpus* Mac-Morpho, (ii) a anotação sintática do *corpus* Floresta Sintá(c)tica, (iii) a anotação de papéis semânticos do Propbank-Br, (iv) anotações de discurso (RST e CST) do *corpus* CSTNews, dentre outras (ALUÍSIO; ALMEIDA, 2006).

Além das anotações já citadas, pode-se considerar que a tarefa de alinhamento também configura um tipo de anotação. Especificamente, alinhamento é o processo de se relacionar segmentos textuais (por exemplo, itens lexicais, orações, sentenças, parágrafos) de diferentes textos. Essa tarefa é utilizada em muitas aplicações desenvolvidas no PLN: tradução,

perguntas e respostas, simplificação textual, sumarização, entre outras. Na maioria das aplicações, o alinhamento (também referido por "indexação") geralmente subsidia a aquisição de conhecimentos sobre a tarefa/fenômeno em estudo, a fim de permitir a sua descrição linguística e/ou automação.

Pode-se citar uma série de autores que realizaram alinhamentos entre textos sobre o mesmo tema, sendo que cada um deles usou um critério diferente para relacionar as unidades textuais. Marcu (1999) e Jing e MacKeown (1999), por exemplo, basearam-se na intuição para realizar a tarefa, isto é, na interpretação das unidades textuais para decidir o que deveria ser alinhado. Hatzivassiloglou et al. (1999), por sua vez, definiram duas unidades como semelhantes quando ambas focavam um mesmo conceito (ator, objeto ou ação). Barzilay e Elhadad (2003) alinharam 2 sentenças quando estas continham pelo menos 1 oração que expressasse o mesmo conteúdo.

Especificamente para a SA, os primeiros esforços datam do final dos anos 90, com poucas obras que realizaram o alinhamento entre textos e sumários.

A tarefa de alinhamento é reconhecidamente relevante para a SA monodocumento, pois o alinhamento de sumários humanos ou manuais a textos-fonte evidencia a origem das informações que compõem o sumário, permitindo investigar suas características e a forma por meio da qual foram transpostas para o sumário (NENKOVA, 2006). Desse tipo de investigação, é possível obter estratégias de sumarização humana que podem subsidiar a sumarização automática, tornando-a mais linguisticamente motivada.

Para a sumarização monodocumento, Marcu (1999) relata que 14 anotadores selecionaram aleatoriamente 10 textos originais do *corpus* Ziff-Davis[16] e os alinharam aos seus correspondentes sumários humanos monodocumento, considerando orações e sentenças como unidades textuais. Os anotadores também anotaram o tipo de sobreposição entre as unidades alinhadas, considerando uma das 5 possibilidades: (i) relacionam-se perfeitamente, (ii) relacionam-se perfeitamente, mas a unidade do resumo apresenta mais informações, (iii) relacionam-se perfeitamente, mas a unidade do texto-fonte apresenta mais informações, (iv) relacionam-se por meio de inferência, e (v) relacionam-se de forma distinta das outras (i, ii, iii e iv).

[16] Uma coleção de textos jornalísticos que anunciam produtos relacionados com computadores, em inglês.

Nesse experimento, os autores mediram a concordância entre os anotadores usando a medida *kappa* (CARLETTA, 1996), que é a medida de concordância mais sofisticada que se tem conhecimento. O valor dessa medida varia de 0 a 1, sendo que 0 indica a não concordância entre os anotadores e 1 indica total concordância entre eles. Assim, os autores avaliaram tanto a concordância em nível sentencial, quanto em nível oracional, resultando em 0,52 e 0,48, respectivamente.

Semelhante a esse estudo, pode-se destacar o trabalho de Jing e McKeown (1999), os quais seguiram as mesmas diretrizes que Marcu (1999) e, por isso, selecionaram alguns textos-fonte a partir do mesmo *corpus*. Assim, 14 anotadores alinharam 10 textos coletados aleatoriamente do *corpus* Ziff-Davis e o alinhamento foi baseado na interpretação das unidades textuais.

Na obra de Barzilay e Elhadad (2003), 2 anotadores alinharam textos a seus sumários. Eles alinhavam 2 sentenças que contivessem pelo menos uma oração que expressasse a mesma informação. Segundo as autoras, houve concordância para a maioria dos casos e, para os casos em que houve discordância, decidiu-se por um terceiro juiz. O acordo entre os anotadores não foi computado.

Outros estudos que também realizaram o alinhamento entre textos e seus sumários monodocumento são os de Daumé III e Marcu (2004, 2005). Nesses 2 estudos, 2 anotadores alinharam manualmente 45 textos-fonte aos seus sumários em nível sentencial. Os textos utilizados foram selecionados do *corpus* Ziff-Davis, o mesmo utilizado por Marcu (1999) e Jing e McKeown (1999). Como treinamento, 5 dos 45 textos-fonte foram anotados individualmente e as anotações foram discutidas na sequência por ambos os especialistas. Os outros 40 foram anotados de forma independente. Os anotadores tinham de verificar se os alinhamentos eram "*possible*" ou "*sure*". Além disso, eles usaram a medida *kappa* para calcular a concordância entre os anotadores e o resultado foi de 0,63.

Quanto à sumarização multidocumento, pode-se destacar o trabalho de Hirao et al. (2004). Nesse trabalho, os autores utilizaram o *corpus* TSC (*Text Summarization Challenge*) (OKUMURA et al., 2003). Esse *corpus* é composto por 30 textos-fonte monodocumento e 30 coleções de textos-fonte (224 no total) que versam sobre um mesmo tópico. Para cada coleção de textos, 3 anotadores criaram 3 sumários curtos e 3 longos. Em seguida, as sentenças dos sumários foram alinhadas com as sentenças dos textos-fonte. Os autores observaram que o processo de alinhamento considerando sumários multidocumento era um fenômeno complexo, uma vez que as sentenças desses sumários podiam resultar da compactação,

combinação e integração de outras sentenças. O acordo entre os anotadores não foi computado.

Como se pode notar, há alguns autores que atribuem tipos aos alinhamentos. Especificamente, a tipificação consiste em classificar os alinhamentos de acordo com alguns critérios e, para expressar/codificar os tipos de alinhamentos, etiquetas ou rótulos são comumente utilizados (p.ex.: MARCU, 1999; DAUMÉ III e MARCU, 2004, 2005).

Quanto à tipificação, cita-se também o trabalho realizado por Clough et al. (2002), no qual os autores tentaram descobrir os tipos de operações de reescrita que podem ocorrer entre um texto original proveniente de uma agência jornalística e suas diferentes versões publicadas posteriormente. Considerando o texto como um todo, os jornalistas atribuíram um dos 3 rótulos: (i) integralmente derivado, (ii) parcialmente derivado ou (iii) não derivado. Em nível sentencial ou lexical, eles tipificaram o alinhamento em: (i) literal, (ii) reescrito ou (iii) novo. Novamente, a concordância entre os anotadores não foi computada.

A seguir, apresenta-se a revisão literária sobre como a avaliação de sistemas e estratégias de seleção de conteúdo na área de sumarização tem sido realizada.

2.5 Avaliação na SA

A avaliação de sistemas de PLN foi bastante explorada na última década, pois permite verificar o avanço do "estado da arte" das aplicações como a sumarização automática. Aliás, a realização de conferências internacionais dedicadas somente à avaliação de sumarizadores automáticos, como a SUMMAC[17] (*Text Summarization Evaluation Conference*) e a TAC[18] (*Text Analysis Conference*), evidencia a importância e a necessidade da avaliação e das dificuldades inerentes à SA.

De um modo geral, a avaliação de sistemas de SA pode ser classificada como intrínseca ou extrínseca. Na primeira, avalia-se o desempenho dos sistemas por meio da análise de seus resultados (sumários). Na segunda, avalia-se a utilidade dos sumários em alguma tarefa específica, por exemplo, na recuperação de informação (SPARCK JONES; GALLIERS, 1996).

---

[17] http://www-nlpir.nist.gov/related_projects/tipster_summac/
[18] http://www.nist.gov/tac/about/index.html

Na literatura, reconhece-se que a avaliação extrínseca é uma tarefa demorada, cara e que requer um planejamento cuidadoso (HALTEREN; TEUFEL, 2003) e que a intrínseca deve focar a qualidade e a informatividade dos sumários (MANI, 2001). A avaliação intrínseca, aliás, é a mais frequentemente realizada nos trabalhos de SA.

A avaliação da qualidade dos sumários automáticos é realizada por humanos, pois o foco reside na análise de aspectos relativos à gramaticalidade (p.ex.: ortografia e gramática) e à textualidade (p.ex.: coesão e coerência) (p.ex.: SAGGION; LAPALME, 2000; WHITE et al., 2000), os quais não podem ser avaliados automaticamente.

Na SAM, Pitler et al. (2010), por exemplo, propuseram uma série de atributos linguísticos que buscam avaliar a qualidade dos sumários em função de gramaticalidade, não-redundância, clareza referencial, foco, estrutura e coerência. Sobre a SAM que envolve o português, Castro Jorge e Pardo (2011) realizaram a avaliação humana da qualidade dos sumários gerados pelo CSTSumm por um conjunto de juízes que analisaram 5 aspectos textuais bastante semelhantes aos de Pitler et al. (2010): informatividade, coerência, coesão, redundância e gramaticalidade.

A avaliação da informatividade consiste em identificar o quanto de informação relevante dos textos-fonte o sumário automático incorpora. Essa identificação é comumente feita pela comparação automática entre os sumários automáticos e os sumários humanos ("sumários de referência"). Para tanto, utiliza-se com frequência o pacote de medidas denominado *Recall-Oriented Understudy of Gisting Evaluation* (ROUGE), que calcula a informatividade por meio da coocorrência de n-gramas entre os sumários automáticos e os humanos e a expressa pelas medidas de precisão, cobertura e medida-f (LIN; HOVY, 2003).

Na SAM, Castro Jorge e Pardo (2011), por exemplo, também avaliaram a informatividade dos sumários gerados pelo CSTSumm e, nesse caso, utilizam exatamente a medida ROUGE (LIN; HOVY, 2003) para compará-los a sumários humanos de referência.

Além dessas possibilidades, diversos autores têm investigado outras estratégias, já que não há consenso sobre a melhor forma de se avaliar um sistema dessa natureza. Dentre eles, citam-se Sparck Jones (1999), Jing et al. (1998), Donaway et al. (2000), Saggion et al. (2002), Halteren e Teufel (2003), Nenkova e Passonneau (2004), Louis e Nenkova (2013), etc.

Saggion et al. (2002), por exemplo, propuseram 3 métodos de avaliação baseado em conteúdo 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. Halteren e Teufel (2003) especificaram uma abordagem que combina 2 aspectos: (i) comparação entre sumário automático e sumário de referência via

factoids[19], uma representação pseudo-semântica das unidades de informação presentes nos textos-fonte (jornalísticos), e (ii) uso de um sumário consensual de referência, baseado em 50 *abstracts* de um mesmo texto. Nenkova e Passonneau (2004), por sua vez, propuseram um método denominado "pirâmide", o qual atribui valor ao sumário por meio da similaridade entre suas *summarization content units* (SCUs), ou seja, a SCU que aparecer em todos os sumários de referência sob avaliação recebe o maior peso (número de sumários em que apareceu) e ocupa a última camada da pirâmide. Nesse sentido, é possível prever o conteúdo ideal que deve conter em um sumário, visto que no topo se encontram as unidades mais importantes.

Diferentemente dos trabalhos anteriores, em Louis e Nenkova (2013), as autoras apresentam 3 métricas de avaliação para a sumarização baseadas em pouco ou nenhum envolvimento humano, são elas: (i) similaridade entre textos-fonte e sumários, isto é, elas consideram que quanto mais similar o sumário é dos seus textos-fonte, melhor o seu conteúdo, (ii) adição de pseudomodelos, ou seja, aos sumários humanos de referência, acrescentam-se sumários automáticos escolhidos por humanos e (iii) sumários automáticos como modelo, isto é, as autoras consideram que os sumários automáticos são bons o suficiente para servirem de sumários de referência e, portanto, não utilizam nenhum esforço humano para a criação dos mesmos. Dessa forma, as autoras revelam que as avaliações humanas podem ser reproduzidas por essas métricas totalmente automáticas com alta precisão.

Os diferentes métodos de avaliação aqui apresentados, originalmente propostos para a avaliação de sistemas de SA, podem ser utilizados/adaptados para a avaliação das estratégias de seleção de conteúdo da SHM. Assim, a análise intrínseca das estratégias pode ser feita por meio de: (i) a verificação da ocorrência das estratégias em outro *corpus*, distinto do utilizado para o aprendizado das estratégias, o qual é comumente denominado "*corpus* de teste", (ii) a comparação da informatividade de sumários produzidos pelas estratégias em questão com a informatividade de sumários de referência, e (iii) comparação da qualidade de sumários produzidos pelas estratégias em questão com a qualidade de sumários produzidos por

[19] Neste caso, *factoids* correspondem a expressões interpretadas a partir de uma única unidade de significado. Por exemplo, a partir da sentença "A polícia prendeu um homem holandês branco", é possível identificar 5 *factoids*: (i) um suspeito foi preso; (ii) a polícia realizou a prisão; (iii) o suspeito era branco; (iv) o suspeito era holandês, e (v) o suspeito era do sexo masculino.

estratégias diferentes. A avaliação extrínseca, por sua vez, pode ser feita pela análise da utilidade, em alguma tarefa específica (p.ex.: recuperação de informação), dos sumários produzidos pelas estratégias que se quer avaliar.

Na sequência, apresentam-se a seleção do *corpus* utilizado como fonte linguística para a descrição do objeto de estudo em questão, no caso, sumários multidocumento, e o subsequente processo de anotação do *corpus* selecionado.

3 SELEÇÃO E ANOTAÇÃO DO *CORPUS*

3.1 Seleção do *corpus*

Por definição, um *corpus* é um conjunto de dados linguísticos em formato eletrônico que tenham sido sistematizados de acordo com determinados critérios para representar, na medida do possível, uma língua ou uma variedade da língua e que, por isso, possam servir como fonte para pesquisa linguística (SINCLAIR, 2005)[20]. Por essa definição, *corpus* é um artefato produzido para pesquisa e, por isso, a maioria de suas características é dependente dos objetivos da pesquisa. Tendo em vista o objetivo ora apresentado, o *corpus* tinha de apresentar as seguintes características:

- Monolíngue: essa característica advém do objetivo de se investigar a SHM a partir de textos em português, buscando, assim, gerar descrições linguísticas sobre essa língua que possam contribuir para o crescimento do corpo de trabalhos interdisciplinares em PLN e Linguística Descritiva.
- Multidocumento: essa característica advém do fato de este trabalho focar a SHM; portanto, necessitava-se de um *corpus* composto por coleções (ou *clusters*) de textos que versam sobre um mesmo assunto, advindos de fontes distintas.

Com base nos objetivos do trabalho, identificaram-se 2 corpora multidocumento na literatura, a saber: o CSTBank[21] (RADEV et al., 2004) e o CSTNews (CARDOSO et al., 2011a).

O CTSBank foi o primeiro corpus multidocumento anotado com relações CST que se tem registro. Especificamente, ele é um *corpus* em língua inglesa composto por 6 coleções de textos jornalísticos anotados manualmente por relações CST. Os textos das coleções foram agrupados manual e automaticamente e as coleções, por sua vez, organizadas em "famílias", isto é, de acordo com a fonte jornalística da qual os textos foram extraídos (p.ex.: CNN, USAToday, ABCNews, FoxNews). No total, o CSTBank possui 41 textos com em média 28 sentenças.

---

[20] "*A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research*" (SINCLAIR, 2005).

[21] http://tangra.si.umich.edu/clair/CSTBank/

Por ter sido construído para pesquisas sobre a língua inglesa, o CSTBank não cumpria com os requisitos necessários traçados neste trabalho.

O outro *corpus* identificado na literatura foi o CSTNews (CARDOSO et al., 2011). Tendo em vista que o *corpus* para este projeto devia ser monolíngue (português) e multidocumento, selecionou-se o CSTNews como fonte de dados linguísticos para a realização da pesquisa.

3.1.1 O *corpus* CSTNews

O CSTNews é um *corpus* composto por 50 coleções ou grupos de textos, sendo que cada coleção versa sobre um mesmo tópico. Os textos são do gênero discursivo "notícias jornalísticas", pertencentes à ordem do relatar[22] (DOLZ; SCHNEWLY, 2004). As principais características do gênero "notícias" são: (i) documentar as experiências humanas vividas (domínio social) e (ii) representar pelo discurso as experiências vividas, situadas no tempo (capacidade da linguagem) (BARBOSA, 2001; LAGE, 2004).

Especificamente, cada coleção do CSTNews contém: (i) 2 ou 3 textos sobre um mesmo assunto ou tema compilados de diferentes fontes jornalísticas; (ii) sumários humanos (*abstracts*) mono e multidocumento; (iii) sumários automáticos multidocumento; (iv) extratos humanos multidocumento; (v) anotações CST e RST dos textos, dentre outros dados.

As fontes jornalísticas das quais os textos foram compilados correspondem aos principais jornais *online* do Brasil, a saber: *Folha de São Paulo, Estadão, Jornal do Brasil, O Globo* e *Gazeta do Povo*. A coleta manual foi feita durante aproximadamente 60 dias, de agosto a setembro de 2007. As coleções possuem em média 42 sentenças (de 10 a 89) e os sumários humanos multidocumento possuem em média 7 sentenças (de 3 a 14).

Ademais, as coleções estão categorizadas pelos rótulos das "seções" dos jornais dos quais os textos foram compilados. Assim, o *corpus* é composto por coleções das seguintes

[22] Dolz e Schnewly (2004) classificam os gêneros textuais em 5 categorias de acordo com algumas regularidades linguísticas, a saber: (i) textos da ordem do relatar, (ii) do narrar, (iii) do expor, (iv) do descrever ações e (i) do argumentar. Na categoria "ordem do relatar", estão agrupados os gêneros pertencentes ao domínio social da memorização e documentação das experiências humanas, como diários de viagem, notícias, reportagens, crônicas jornalísticas, relatos históricos, biografias, autobiografias, testemunhos, etc.

categorias: “esporte” (10 coleções), “mundo” (14 coleções), “dinheiro” (1 coleção), “política” (10 coleções), “ciência” (1 coleção) e “cotidiano” (14 coleções).

Quanto aos sumários humanos multidocumento, especificamente, ressalta-se que estes foram construídos manualmente de forma abstrativa, ou seja, com reescrita do conteúdo dos textos-fonte. Além disso, a produção dos mesmos foi guiada por uma taxa de compressão de 70%. Consequentemente, os sumários contêm, no máximo, 30% do número de palavras do maior texto-fonte da coleção. Do ponto de vista da audiência, os sumários do CSTNews são genéricos, pois não foram construídos com o foco em leitores específicos e, do ponto de vista funcional, são informativos, pois contemplam as informações principais de seus textos-fonte, substituindo a leitura dos mesmos.

Os dados completos sobre o CSTNews estão descritos no Quadro 7, elaborado com base em Aleixo e Pardo (2008) e Cardoso et al. (2011a).

Quadro 7 - Estatísticas do CSTNews

Coleção	Categoria	Nº de documentos	Nº de sentenças por	Nº de sentenças por
C1	Mundo	3	24	5
C2	Política	3	51	7
C3	Cotidiano	3	50	10
C4	Cotidiano	3	39	5
C5	Cotidiano	2	23	5
C6	Cotidiano	3	36	5
C7	Ciência	2	23	4
C8	Esportes	3	25	6
C9	Política	3	36	6
C10	Mundo	3	38	10
C11	Cotidiano	3	56	11
C12	Mundo	3	34	4
C13	Mundo	3	37	6
C14	Mundo	3	25	5
C15	Mundo	3	26	6
C16	Política	3	47	6
C17	Política	2	41	6
C18	Mundo	3	70	9
C19	Esportes	2	13	4
C20	Política	3	42	8
C21	Cotidiano	3	41	3
C22	Cotidiano	3	50	9
C23	Mundo	2	25	6
C24	Esportes	3	24	5
C25	Esportes	3	88	8
C26	Mundo	3	58	10
C27	Esportes	3	89	12
C28	Esportes	3	35	4
C29	Mundo	3	48	6
C30	Dinheiro	3	46	4
C31	Esportes	2	10	3
C32	Mundo	3	66	9
C33	Cotidiano	3	68	13
C34	Cotidiano	3	59	8
C35	Mundo	3	36	7
C36	Cotidiano	3	74	14
C37	Cotidiano	2	26	5
C38	Esportes	3	26	3
C39	Cotidiano	3	34	3
C40	Política	3	28	4
C41	Esportes	3	45	6
C42	Política	2	39	5
C43	Política	3	49	7
C44	Política	2	26	9
C45	Cotidiano	3	47	6
C46	Mundo	3	23	5
C47	Mundo	3	43	6
C48	Esportes	2	43	9
C49	Cotidiano	3	23	6
C50	Política	3	62	8
Total	—	140	2067	331
Média	—	**2.8**	**41.34**	**6.62**

Fonte: Elaborado pelo autor

A seguir, apresenta-se o processo de anotação do *corpus* selecionado, que consistiu no alinhamento manual das sentenças dos sumários aos seus respectivos textos-fonte.

3.2 A anotação do *corpus*

Uma vez selecionado o *corpus* CSTNews, realizou-se o alinhamento manual das sentenças dos sumários multidocumento com as sentenças dos textos-fonte das diversas coleções. Ressalta-se que o alinhamento foi realizado manualmente por se tratar de uma tarefa subjetiva. Como já mencionado, essa tarefa evidencia a origem das informações que compõem o sumário, permitindo investigar suas características e a forma por meio da qual foram transpostas para o sumário. A partir desse tipo de investigação, é possível obter estratégias de SHM que podem subsidiar a SAM.

A tarefa em questão foi realizada por 2 anotadores[23] da área de Linguística Computacional durante aproximadamente 2 meses, em reuniões diárias de 1 a 2 horas. Cada pesquisador ficou responsável por alinhar metade das coleções do CSTNews.

Na Figura 5, ilustra-se o alinhamento referente à coleção 31 do CSTNews, que pertence à categoria "esporte" e é composta por 2 documentos ou textos-fonte. Observa-se, por exemplo, que a sentença (S) 1 do sumário foi alinhada à S1 do documento (D) 1 e à S1 e S2 do D2. No caso da coleção 31, nenhuma sentença do sumário foi alinhada à S4 do D1 e à S5 do D2.

O alinhamento foi feito em função de duas diretrizes centrais. A primeira delas diz respeito ao nível dos segmentos a serem alinhados e a segunda refere-se ao critério para a identificação das correspondências.

Quanto ao nível dos segmentos textuais, optou-se pelo sentencial, posto que as sentenças são unidades de informação bem delimitadas.

Sobre o critério de alinhamento propriamente dito, ressalta-se que as correspondências entre os sumários e seus respectivos textos-fonte foram identificadas com base na sobreposição de conteúdo, total ou parcial. Tendo em vista que os sumários do CSTNews são *abstracts*, o alinhamento das informações contidas nos sumários com as dos textos-fonte nem sempre foi simples, posto que houve reescrita do conteúdo dos textos-fonte transpostos para a versão condensada.

---

[23] O alinhamento em questão foi realizado pela própria autora da dissertação em conjunto com Verônica Agostini, mestranda do Instituto de Ciências Matemáticas e de Computação (ICMC-USP) e pesquisadora do NILC, com a supervisão do Prof. Thiago A. S. Pardo (ICMC-USP), que coorientou este trabalho.

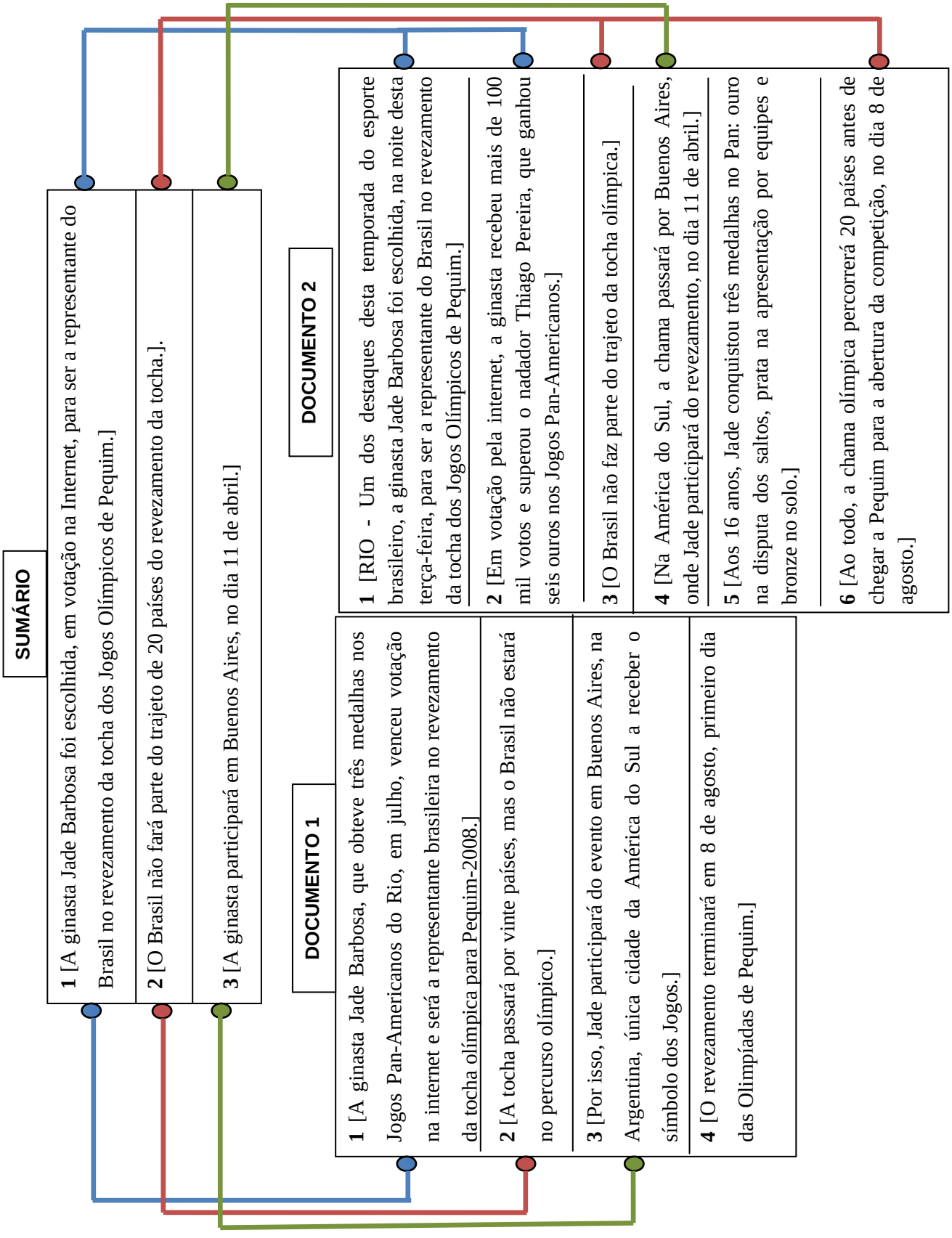


Figura 5 - Exemplo do alinhamento de um sumário a seus textos-fonte

Fonte: Elaborado pelo autor.

Ao se optar por um alinhamento baseado na sobreposição de conteúdo ou informação, o processo de indexação não se baseia na sobreposição de formas, ou seja, de unidades lexicais. Consequentemente, sentenças que continham conteúdo em comum, total ou parcial, com baixa sobreposição lexical (do inglês, *word overlap*), foram alinhadas.

No exemplo do Quadro 8, observa-se que a sentença do sumário e a sentença do texto-fonte apresentam sobreposição parcial de conteúdo. Diz-se "parcial" porque a sobreposição refere-se apenas aos trechos negritados. No caso, o trecho "**se preparando para a passagem do furacão**" do sumário expressa uma informação mais genérica que o trecho "**estocaram alimentos, água, lanternas e velas**" do texto-fonte, já que a "estocagem" pode ser interpretada como uma "espécie de preparação" para a chegada do furacão. Tal sobreposição de conteúdo não seria identificada com base exclusivamente nas unidades lexicais, pois as sentenças em questão não apresentam palavras de conteúdo (nome, verbo, adjetivo e advérbio) em comum.

Quadro 8 - Exemplo de alinhamento com base na sobreposição de conteúdo

Sentença do sumário	Sentença do documento
Vários moradores e turistas nas regiões, inclusive brasileiros, foram retirados dos locais, enquanto outros estão **se preparando para a passagem do furacão**.	Na Jamaica, muitos **estocaram alimentos, água, lanternas e velas**.

Fonte: Elaborado pelo autor.

A partir das diretrizes gerais, realizou-se, antes do alinhamento propriamente dito, uma fase de treinamento, na qual 2 coleções foram aleatoriamente selecionadas e alinhadas pelos anotadores, individual e separadamente. Na sequência, os alinhamentos foram comparados e os casos de divergência foram discutidos com o intuito de ajustar a concordância entre os anotadores.

Desse treinamento, algumas regras gerais e específicas foram elaboradas, as quais passaram por um processo de refinamento ao longo da indexação. Assim, ao final, gerou-se um manual de alinhamento de sumários humanos multidocumento e textos-fonte, cujas regras são apresentadas na sequência.

3.2.1 As regras de alinhamento

Ao todo, elaboraram-se 8 regras, sendo 4 gerais e 4 específicas. A seguir, tais regras são descritas e exemplificadas por alinhamentos reais do CSTNews, destacando-se, sobretudo, os critérios linguísticos que subsidiaram a formulação das mesmas. Para tanto, as sentenças dos documentos ou textos-fonte são referenciadas pela abreviação SD e as sentenças dos sumários são referenciadas pela abreviação SS.

a) Regras Gerais

REGRA 1: *Alinhar com base na sobreposição de conteúdo e não de forma*
Essa regra estabeleceu que o alinhamento fosse feito em função da sobreposição de conteúdo entre uma SS e uma ou mais SDs e não em função da ocorrência de unidades lexicais comuns ou mesmo estruturas sintáticas semelhantes. Consequentemente, sentenças que veiculavam certo conteúdo em comum por meio de expressões linguísticas (superficiais) diferentes foram alinhadas. Além disso, ressalta-se que o conteúdo em comum nem sempre foi identificado diretamente, mas sim por meio de inferências.

Em (1) e (2), apresentam-se exemplos de alinhamento com base na Regra 1. No caso de (1), as sentenças compartilham o mesmo conteúdo principal, ou seja, "o número de mortos no acidente aéreo", expresso de forma diferente. Em (2), o alinhamento foi feito por meio de inferência, isto é, a expressão "abrir(ão) mão" da SS focaliza o mesmo conteúdo central da SD, no caso, "as renúncias", já que a expressão "abrir mão" é semanticamente equivalente à unidade lexical "renunciar".

(1) **SS:** 17 pessoas morreram após a queda de um avião na República Democrática do Congo.
SD: Um acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo (RDC), matou 17 pessoas na quinta-feira à tarde, informou nesta sexta-feira um porta-voz das Nações Unidas. (D2_C1)

(2) **SS:** A expectativa de lideranças da Câmara e do Conselho de Ética é que pouco mais de 10% dos 69 deputados denunciados no relatório parcial da CPI **abrirão mão de seus mandatos**.

SD: **As renúncias** têm que ser publicadas até terça-feira, quando o presidente do Conselho de Ética, deputado Ricardo Izar (PTB-SP), vai instaurar os processos de perda de mandato contra os 69 deputados acusados pela CPI dos Sanguessugas de envolvimento com a máfia das ambulâncias. (D1_C16)

REGRA 2: *Alinhar com base na sobreposição da informação principal*

Essa regra estabeleceu que o alinhamento fosse feito em função do conteúdo principal veiculado pelas sentenças. Assim, uma SS foi alinhada a SDs quando houve sobreposição da ideia central, expressa pelo verbo principal.

Em (3), apresenta-se um exemplo em que as sentenças não foram alinhadas em função da Regra 2, apesar da sobreposição dos sujeitos. Os verbos principais "descobrir" e "informar", que expressam a informação principal de cada sentença, não são similares.

(3) **SS:** Usando telescópios do Observatório Europeu Sul (ESO), Ray Jayawardhana, da Universidade de Toronto, e Valentin D. Ivanov, do ESO, **descobriram** um planemo com sete vezes a massa de Júpiter, o planeta mais pesado do Sistema Solar, e outro com o dobro desse peso, que giram um ao redor do outro, denominado Oph 162225-240515, o primeiro planemo duplo.

SD: Os pesquisadores Ray Jayawardhana e Valentin D. Ivanov **informam** a descoberta na edição de quinta-feira do serviço online Science Express, mantido pela revista Science. (D1_C7)

REGRA 3: *Alinhar com base na sobreposição de informação secundária*

Essa regra especificou que as sentenças fossem alinhadas diante da sobreposição de conteúdo ou informação secundária. Assim, uma SS foi alinhada a uma ou mais SDs não somente pelo conteúdo principal, mas também pelo compartilhamento de informação periférica.

Em (4) e (5), apresentam-se alinhamentos que ilustram a aplicação da Regra 3. Em (4), por exemplo, a SS e a SD foram alinhadas porque compartilham a informação secundária expressa pelos trechos "**giram um ao redor do outro**" e "**giram em torno um do outro**", apesar de não haver sobreposição do conteúdo central. Em (5), o alinhamento foi realizado porque SS e SD compartilham um mesmo episódio: "Renan pagar uma pensão informal (despesas pessoais) à jornalista".

(4) **SS:** Usando telescópios do Observatório Europeu Sul (ESO), Ray Jayawardhana, da Universidade de Toronto, e Valentin D. Ivanov, do ESO, descobriram um planemo com sete vezes a massa de Júpiter, o planeta mais pesado do Sistema Solar, e outro com o dobro desse peso, que **giram um ao redor do outro**, denominado Oph 162225-240515, o primeiro planemo duplo.

SD: Ambos os mundos têm massa semelhante à de outros exoplanetas já catalogados, mas não giram em torno de uma estrela - na verdade, **giram em torno um do outro**. (D1_C7)

(5) **SS:** Renan é alvo de um processo por quebra de decoro acusado de receber recursos da construtora Mendes Junior para **pagamento de despesas pessoais, como aluguel e pensão para a jornalista Mônica Veloso**, com quem tem uma filha.

SD: Isso permitiria que os peritos da Polícia Federal pudessem trabalhar durante o período de descanso dos senadores e, no retorno das férias, apresentarem um relatório detalhado sobre o conjunto de documentos - notas fiscais, recibos de vacinação, extratos bancários, guias de transporte de animais - que o senador apresentou para justificar o **pagamento da pensão informal à jornalista Mônica Veloso**. (D3_C43)

REGRA 4: *Alinhar todas as sobreposições de um mesmo conteúdo*

Essa regra estabeleceu que uma SS devia ser alinhada sempre que uma SD com sobreposição de conteúdo fosse identificada, mesmo que a SS já tivesse sido alinhada por causa do compartilhamento desse mesmo conteúdo.

Em (7), ilustra-se a aplicação da Regra 4. No caso, a SS, já alinhada a uma sentença do texto-fonte D1 em função do compartilhamento de informação secundária (cf. (4)), foi alinhada novamente a 2 sentenças distintas do texto-fonte D2 da mesma coleção, posto que a sobreposição de conteúdo foi identificada novamente.

(7) **SS:** Usando telescópios do Observatório Europeu Sul (ESO), Ray Jayawardhana, da Universidade de Toronto, e Valentin D. Ivanov, do ESO, descobriram um planemo com sete vezes a massa de Júpiter, o planeta mais pesado do Sistema Solar, e outro com o dobro desse peso, que **giram um ao redor do outro**, denominado Oph 162225-240515, o primeiro planemo duplo.

SD: Astrônomos do Observatório Europeu Austral, localizado no Chile, anunciaram a descoberta de uma dupla de planetas errantes (sem estrela-mãe) que **giram ao redor deles mesmos** e que vagam livremente pelo espaço. (D2_C7)

SD: O fato extraordinário é que **ele não gira em volta de uma estrela, mas em torno de outro corpo frio** com o dobro de sua massa. (D2_C7)

b) Regras Específicas

O conjunto de regras específicas é composto efetivamente por 4 normas, as quais foram elaboradas em função de casos particulares de sobreposição de conteúdo.

A Regra 5, em especial, foi formulada para lidar com casos de contradição, especificamente quando uma SS e uma ou mais SDs expressavam basicamente o mesmo conteúdo, mas diferiam quanto a dados numéricos referentes a um mesmo fato.

As demais regras foram formuladas para os casos em que uma SS e uma ou mais SDs expressavam o mesmo conteúdo principal, mas diferiam quanto ao: (i) grau de generalização (ou especificação) de uma mesma informação (Regra 6) e (iii) grau de assertividade do falante sobre um mesmo fato (Regra 7).

A Regra 8 é a única que, apesar da similaridade de conteúdo, estabeleceu o não-alinhamento.

REGRA 5: *Alinhar com base na sobreposição da informação principal mesmo diante de dado numérico contraditório*

Essa regra estabeleceu que dada SS devia ser alinhada a uma ou mais SDs em função da sobreposição da ideia central mesmo diante de dados numéricos contraditórios, os quais podiam, por exemplo, ser referentes à hora de ocorrência de determinado fato.

Em (8), o exemplo em questão ilustra um alinhamento feito com base na Regra 5. Especificamente, a SS e a SD compartilham o conteúdo principal, no caso, ambas registram o fato de "a cidade de São Paulo apresentar pontos de alagamento", mas apresentam informação contraditória sobre o horário em que tal fato foi observado/registrado. Diante de contradições desse tipo, as sentenças foram alinhadas.

(8) **SS:** Às **9h**, a cidade tinha oito pontos de alagamento, sendo dois intransitáveis.

SD: O CGE (Centro de Gerenciamento de Emergências) da Prefeitura de São Paulo registrava oito pontos de alagamento na cidade, às **9h30** desta segunda-feira.

REGRA 6: *Alinhar com base na sobreposição da informação principal mesmo diante de diferentes graus de generalização*

A Regra 6 previu que uma SS devia ser alinhada a uma ou mais SDs em função da sobreposição da ideia central, mesmo que essa informação fosse apresentada com graus distintos de generalização.

Em (9), observa-se que o alinhamento foi feito em função do compartilhamento da informação principal entre a SS e a SD, no caso, "o índice de congestionamento (na cidade de São Paulo) acima da média", apesar de a SS especificar essa informação ao registrar (i) as extensões exatas do congestionamento em quilômetros e (ii) os horários de registro dessas extensões.

Em (10), as sentenças do sumário em questão foram alinhadas com as do texto-fonte pela Regra 6 tendo em vista que a SS(2) e a SS(3) apresentam informações mais específicas que a SD. No caso, a SS(2) e a SS(3) especificam em porcentagem a "intensificação da fiscalização", conteúdo principal de SD.

Quanto à (11), ressalta-se que é a SS que contém a informação mais genérica, ao passo que as SDs dos documentos D2 e D3 contêm a informação mais específica.

(9) **SS:** A Companhia de Engenharia de Tráfego (CET) anunciou que o índice de congestionamento era de **54 quilômetros** às 8h, **113 km** às 9h e **110 km** meia hora depois, valores bem acima das médias para os horários, que eram de **36, 82 e 76 quilômetros** respectivamente, mas não havia registro de acidentes graves, apesar de haver feridos.

SD: Com o asfalto molhado, o trânsito ficou mais lento e **o congestionamento ficou o dobro da média.** (D3_C4)

(10) **SS(2):** O balanço divulgado mostra que as autuações **cresceram 316,5%** nos sete primeiros meses deste ano e chegaram a R$ 1,339 bilhão.

SS(3): Foram autuados 208.471 contribuintes, um **crescimento de 104,47%** em relação ao mesmo período do ano passado.

SD: BRASÍLIA - A Receita Federal **intensificou a fiscalização** e o resultado foi um aumento do número de contribuintes que caíram na malha fina (D2_C34).

(11) **SS:** A Receita Federal **intensificou a fiscalização** sobre as declarações das pessoas físicas neste ano.

SD: Balanço da fiscalização, divulgado nesta segunda-feira pela Receita mostra que as autuações cresceram 316,5% nos sete primeiros meses deste ano e **chegaram a R$ 1,339 bilhão**. (D2_C34).

SD: O volume de recursos recolhido com multas **passou de R$ 326,1 milhões para R$ 1,339 bilhão**. (D3_C34)

REGRA 7: *Alinhar sentenças com sobreposição da informação principal e diferença no grau de assertividade*

A Regra 7 estabeleceu que uma SS devia ser alinhada a uma ou mais SDs em função da sobreposição da ideia central mesmo que tais sentenças apresentassem graus diferentes de assertividade (certeza) do falante com relação à informação principal que está sendo veiculada. No exemplo em (12), as sentenças foram alinhadas devido à sobreposição da informação central (no caso, "a autoria das ações criminosas"), mesmo verificando-se que a SS apresentava maior grau de assertividade do falante quanto ao fato principal que a SD. Na SD, o menor grau de assertividade é identificado pela ocorrência do verbo auxiliar modal "poder" ("**podem ter sido ordenadas"**).

(12) **SS: As ações <u>são atribuídas</u> à facção criminosa Primeiro Comando da Capital (PCC)**, que já comandou outros ataques em duas ocasiões.

SD: As ações criminosas <u>podem ter sido ordenadas</u> pelos líderes do Primeiro Comando da Capital (PCC), que haviam prometido retomar os ataques no Estado de São Paulo no Dia dos Pais, no próximo domingo.

REGRA 8: *<u>Não</u> alinhar sentenças com sobreposição da informação principal quando uma expressar um todo e a outra uma parte do todo.*

Essa regra previu que uma SS e uma ou mais SDs não deviam ser alinhadas caso houvesse diferença de intensidade ou quantidade referente à informação principal comum a elas. Em (13), ilustra-se um caso em que as sentenças não foram alinhadas com base na Regra 8. Nesse exemplo, verifica-se que a SS e a SD apresentam informação principal similar, no caso, "internação do senador". No entanto, a SS apresenta um sintagma adverbial que indica a quantidade de vezes em que o fato principal ocorreu, no caso, "**por três vezes**". A SD, por sua vez, apresenta o sintagma adverbial "**em abril**", que indica a pontualidade da "internação".

Assim, vê-se que a SS expressa a repetição de um mesmo fato (sequência), ao passo que a SD descreve uma das ocorrências do fato, resultando no não alinhamento das sentenças.

(13) **SS:** Somente neste ano, o senador **se internou por três vezes** no Incor.
SD: Em abril, o senador foi internado no InCor com insuficiência cardíaca.

3.2.2 Os resultados do alinhamento

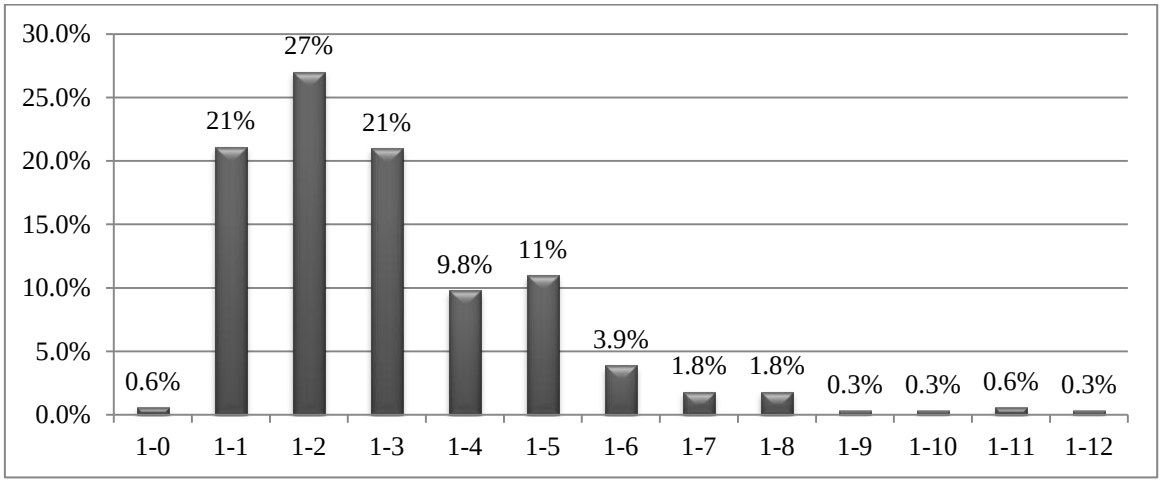
Os resultados provenientes do alinhamento do *corpus* CSTNews englobam especificamente a quantificação dos diferentes tipos de alinhamento e a exemplificação de alguns dos casos. Todos os tipos de alinhamento resultantes podem ser vistos na Tabela 1 e no Gráfico 1.

Tabela 1 - Quantificação numérica dos tipos de alinhamento

	Tipos de alinhamento												
	1-0	1-1	1-2	1-3	1-4	1-5	1-6	1-7	1-8	1-9	1-10	1-11	1-12
Número de alinhamentos	2	71	91	72	33	37	13	6	6	1	1	2	1

Fonte: Elaborado pelo autor.

Gráfico 1 - Quantificação percentual dos tipos de alinhamento



Fonte: Elaborado pelo autor.

Com base na Tabela 1 e no Gráfico 1, observam-se 2 fatos de destaque: (i) sentenças do sumário que não foram alinhadas (alinhamento 1-0) e (ii) sentenças do sumário que foram alinhadas a muitas sentenças dos textos-fonte (alinhamentos do tipo 1-10, 1-11, 1-12).

Os 2 casos de não-alinhamento justificam-se pelo fato de que ambas as sentenças apresentam informação que não está efetivamente presente nos textos-fonte, tendo sido inseridas nos sumários por meio de inferência feita pelos humanos produtores dos resumos. Um desses casos de não-alinhamento ocorreu na coleção C25 do CSTNews, que pertence à categoria "esporte", e cujos textos versam sobre "jogos das seleções brasileiras de vôlei e futebol durante o pan-americano de 2007". No sumário humano multidocumento da C25, a sentença "**Neste domingo, o esporte brasileiro alegrou a torcida verde-amarela**" não foi alinhada a nenhuma sentença dos 3 documentos que compõem a coleção, pois a informação nela contida não está explícita nos textos-fonte, tendo sido inferida ou deduzida pelo humano produtor do sumário.

Na Figura 6, ilustra-se o único caso de alinhamento do tipo 1-12. Esse alinhamento foi feito na coleção C27, pertencente à categoria "esporte" e composta por 3 notícias que relatam "a goleada aplicada pela seleção brasileira de futebol sobre o Equador nas eliminatórias para a Copa do Mundo-2010". No caso, a sentença do sumário "**O jogo contou com belas atuações de craques como Ronaldinho e Kaká**" foi alinhada, no total, a 12 sentenças distintas que compõem os textos-fonte. Esses alinhamentos foram feitos basicamente em função da Regra 6, pois a SS e as 12 SDs compartilham a mesma ideia central com diferentes graus de generalização. No caso, todas as 12 SDs apresentam detalhes sobre a informação generalizada na SS.

Na Tabela 2, apresenta-se a quantificação numérica e percentual das sentenças dos textos-fonte de cada coleção que foram alinhadas.

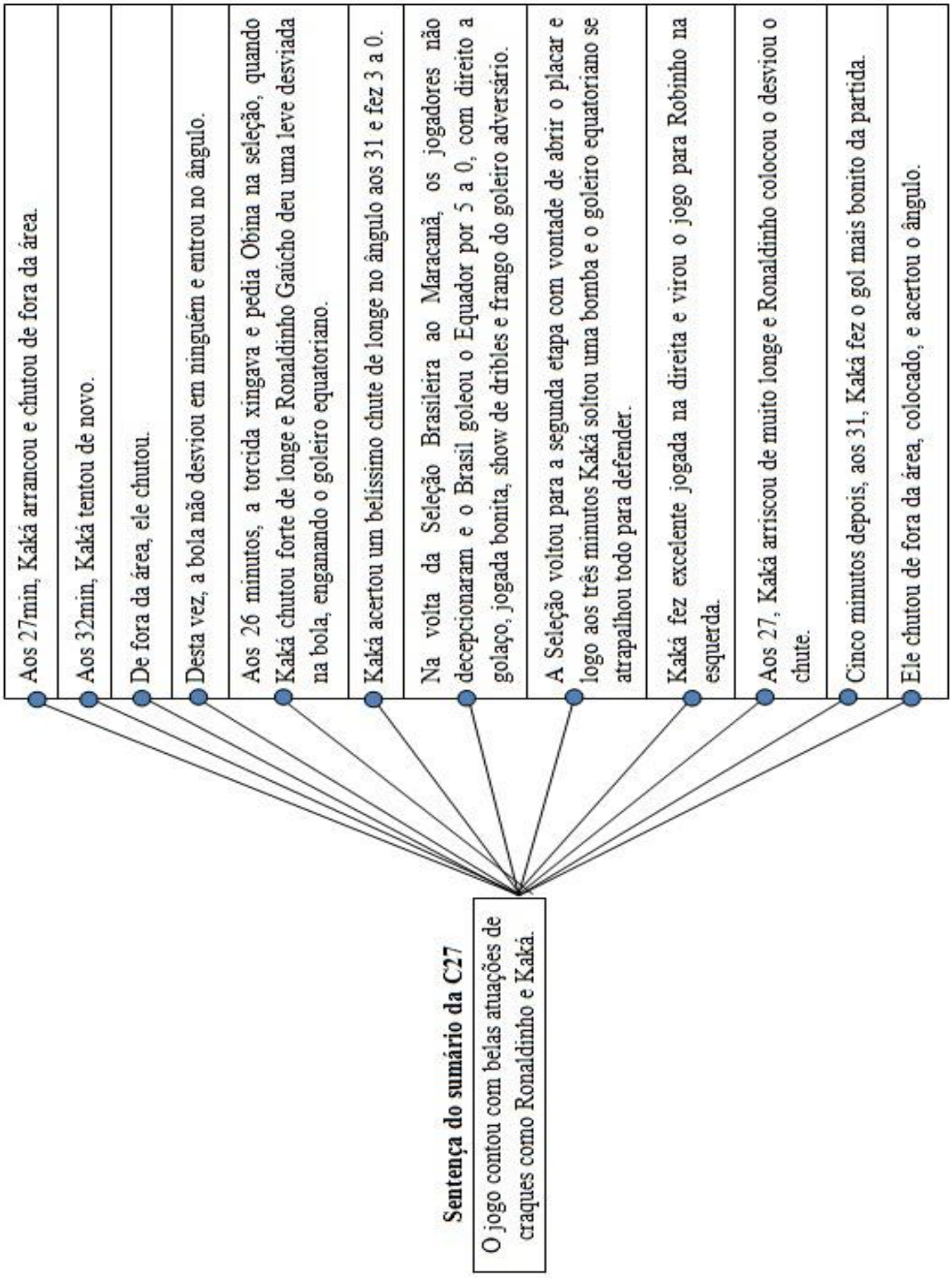


Figura 6 - Exemplo de alinhamento do tipo 1-12.

Fonte: Elaborado pelo autor.

Tabela 2 - Quantidade numérica e percentual das sentenças alinhadas por coleção

Coleção (C)	Nº de sentenças por C	Nº de sentenças alinhadas	% de sentenças alinhadas
C1	24	16	66,66
C2	51	22	43,13
C3	50	18	36
C4	39	16	41,02
C5	23	11	47,82
C6	36	18	50
C7	23	12	52,17
C8	25	16	64
C9	36	18	50
C10	38	19	50
C11	56	22	39,28
C12	34	15	44,11
C13	37	18	48,64
C14	25	19	76
C15	26	17	65,38
C16	47	16	34,04
C17	41	13	31,70
C18	70	25	35,71
C19	13	7	53,84
C20	42	15	35,71
C21	41	7	17,07
C22	50	15	30
C23	25	12	48
C24	24	13	54,16
C25	88	31	35,22
C26	58	28	48,27
C27	89	43	48,31
C28	35	17	48,57
C29	48	13	27,08
C30	46	16	34,78
C31	10	7	70
C32	66	29	43,93
C33	68	29	42,64
C34	59	29	49,15
C35	36	18	50
C36	74	25	33,78
C37	26	10	38,46
C38	26	10	38,46
C39	34	13	38,23
C40	28	10	35,71
C41	45	14	31,11
C42	39	12	30,76
C43	49	12	24,48
C44	26	17	65,38
C45	47	22	46,80
C46	23	10	43,47
C47	43	11	25,58
C48	43	22	51,16
C49	23	19	82,60
C50	62	30	48,38
Total	2067	877	42,43

Fonte: Elaborada pelo autor.

Quanto aos textos-fonte, ressalta-se que, do total de 2067 sentenças que compõem as 50 coleções, 877 (42,43%) foram alinhadas, sendo que a mesma sentença de um sumário pode ter sido alinhada a mais de uma sentença dos textos-fonte. De um modo geral, destaca-se que, das 331 sentenças que compõem o conjunto de 50 sumários do *corpus* CSTNews, 263 (aproximadamente 79%) foram alinhadas a mais de uma sentença dos textos-fonte. Tal fato justifica-se por se tratar de sumários multidocumento, ou seja, versões condensadas de coleções de textos que se estendem sobre um mesmo assunto ou tema.

A confiabilidade dos alinhamentos foi calculada por meio da medida de concordância *kappa*. Com o intuito de calcular a concordância, selecionou-se aleatoriamente 1 coleção por semana durante o período que englobou as últimas 5 semanas de anotação. Assim, 5 coleções foram utilizadas no processo de concordância, sendo de domínios variados (“mundo”, “esporte”, “dinheiro”, “cotidiano” e “política”). A cada semana, os pesquisadores alinharam individualmente uma dessas coleções e compararam os resultados de cada alinhamento para verificar a concordância. A medida *kappa* resultante foi de 0.831, indicando que a tarefa de alinhamento é bem definida. As discordâncias ocorreram por vários motivos; um deles foi a inferência realizada em níveis diferentes (mais ou menos profundo) pelos anotadores, dada a subjetividade da tarefa.

Vale ressaltar que, da comparação entre os alinhamentos individuais de cada coleção, gerou-se um terceiro alinhamento, resultante do consenso entre os pesquisadores e considerado o alinhamento oficial da coleção. Assim, somente os alinhamentos consensuais das 5 coleções utilizadas na concordância foram considerados no cômputo das estatísticas.

Quanto à disponibilização dos alinhamentos, segue-se o formato de anotação XML (do inglês, *Extensible Markup Language*), visto que as demais anotações do CSTNews estão codificadas nessa linguagem. No Quadro 9, o alinhamento do sumário multidocumento e dos textos-fonte da coleção C19 do CSTNews está representado em XML. No esquema em questão, existem quatro blocos de codificação, um para cada sentença do sumário. Esses blocos são delimitados pelas *tags* “<align ‘número da sentença’>” e “</align>”. O primeiro bloco do esquema XML descreve o alinhamento da sentença 1 do sumário (de <align SENT="1"> até </align>). Nesse bloco, a sentença 1 do sumário, codificada como SENT="1", foi alinhada a 3 sentenças dos textos-fonte: SENT="1" do documento 1 (DOC="D1_C19_Folha.txt.seg") e SENT="1" e SENT="2" do documento 2 (DOC="D2_C19_Estadao.txt.seg"). Além da informação sobre a sentença e o texto-fonte, o esquema XML prevê a especificação do tipo de alinhamento (TYPE="none") e do anotador (JUDGE="renata").

Quadro 9 - Exemplo da representação em XML do alinhamento

```
<align SENT="1">
        <DOC="D1_C19_Folha.txt.seg" SENT="1" TYPE="none" JUDGE="renata"/>
        <DOC="D2_C19_Estadao.txt.seg" SENT="1" TYPE="none" JUDGE=" renata "/>
        <DOC="D2_C19_Estadao.txt.seg" SENT="2" TYPE="none" JUDGE=" renata "/>
</align>
<align SENT="2">
        <DOC="D1_C19_Folha.txt.seg" SENT="1" TYPE="none" JUDGE=" renata "/>
        <DOC="D2_C19_Estadao.txt.seg" SENT="2" TYPE="none" JUDGE=" renata "/>
</align>
<align SENT="3">
        <DOC="D1_C19_Folha.txt.seg" SENT="3" TYPE="none" JUDGE=" renata "/>
        <DOC="D2_C19_Estadao.txt.seg" SENT="3" TYPE="none" JUDGE=" renata "/>
</align>
<align SENT="4">
        <DOC="D1_C19_Folha.txt.seg" SENT="4" TYPE="none" JUDGE=" renata "/>
        <DOC="D1_C19_Folha.txt.seg" SENT="5" TYPE="none" JUDGE=" renata "/>
</align>
```

Fonte: Elaborado pelo autor.

O alinhamento em questão do CSTNews está disponível na página eletrônica do projeto SUCINTO para que futuras pesquisas linguístico-computacionais possam ser realizadas.

Quanto às dificuldades encontradas na referida tarefa, destaca-se a necessidade de conhecimento de domínio para realizar o alinhamento dos sumários e dos textos de certas coleções, sobretudo dos que compõem as coleções das categorias “política” e “esporte”.

No Quadro 10, apresenta-se o exemplo da coleção C24, que exigiu dos anotadores conhecimento específico. A marcação entre colchetes no texto corresponde ao número identificador de cada sentença. A coleção C24, proveniente da seção “esporte”, versa sobre a competição da modalidade “salto com vara” nos Jogos Pan-Americanos de 2007. Na sentença [12] do documento 1, diz-se que “a brasileira conseguiu o ouro na segunda tentativa”; na sentença [3] do sumário, narra-se que “a brasileira conseguiu o ouro em três tentativas”. Após pesquisas sobre o esporte em questão, ficou claro que: (i) na sentença do documento, o autor se referiu às tentativas da marca de 4,50m, já que a cada marca o atleta possui três tentativas, e (ii) na sentença do sumário, o autor fez alusão “às tentativas da prova em geral”, ou seja, levou em consideração a quantidade de marcas que a atleta enfrentou para garantir o ouro na competição (4,30m, 4,40m e 4,50m).

Quadro 10 - Exemplo de dificuldade encontrada.

Documento 1
[1]O PRIMEIRO - Murer salta para quebrar recorde pan-americano; primeiro ouro do atletismo.
[2]RIO - Como esperado, a atleta Fabiana Murer conquistou a medalha de ouro - a 29ª brasileira - no salto com vara nos Jogos Pan-Americanos do Rio, nesta segunda-feira, no Estádio João Havelange.
[3]De quebra, esta conquista iguala o número de medalhas de ouro faturadas em Santo Domingo (2003), quando o Brasil também somou 29.
[4]Este recorde deve ser quebrado nesta edição dos Jogos.
[5]Murer - vice-campeã da Copa do Mundo de 2006 - conquistou o lugar mais alto do pódio com a marca de 4m60, contra 4m40 da norte-americana April Steiner, que ficou com a prata.
[6]Já o bronze pertence à cubana Yarisley Silva, com a marca de 4,30m.
[7]Já a outra brasileira que participou da prova, Joana Costa, não subiu ao pódio, uma vez que não alcançou a marca da cubana.
[8]A atleta de Campinas começou a prova na marca de 4,30 m, e seu salto foi perfeito, dando-lhe ânimo inclusive para mandar beijinhos para os torcedores.
[9]O bom salto da brasileira colocou pressão sobre a norte-americana, que falhou em sua primeira tentativa no salto de 4,40m.
[10]Melhor para Fabiana Murer, que alcançou tal marcar em seu primeiro salto, levantando o público.
[11]A atleta norte-americana provou que a disputa seria acirrada, já que bateu a marca de 4,40m, mesmo com as vaias da torcida local.
[12]Fabiana Murer, no entanto, não parecia incomodada, mas errou sua primeira tentativa de alcançar 4,50 m; o erro não se repetiu e a brasileira chegou a tal marca na segunda tentativa, quebrando o recorde pan-americano, garantindo o ouro.
[13]Sem competição, Murer continuou em busca da quebra do seu próprio recorde, e, em sua terceira tentativa, conseguiu alcançar 4,60m.
[14]Sua melhor marca é de 4,66m.

Sumário
[1]A brasileira Fabiana Murer conquistou a medalha de ouro no salto com vara ao saltar 4m60, um novo recorde pan-americano, 20 cm a mais que sua antiga marca.
[2]A medalha de prata ficou com a americana April Steiner com 4m40 e a de bronze com a cubana Yarisley Silva com 4m30.
[3]Fabiana conseguiu o ouro em três tentativas.
[4]Tentou ainda bater o próprio recorde sul-americano de 4m66, mas não conseguiu.
[5]A outra brasileira, Joana Costa, ficou na quinta posição, com 4m20, mostrando que o nervosismo pode atrapalhar as competições em casa.

Fonte: Elaborado pelo autor.

Na sequência, apresenta-se a subsequente tarefa de tipificação dos alinhamentos descritos aqui.

3.3 Tipificação dos alinhamentos

Essa tarefa foi realizada por 2 linguistas computacionais ao longo de um período de 2 meses, com sessões diárias de aproximadamente 1 hora. O processo consistiu em atribuir etiquetas ou rótulos a cada par de sentenças alinhadas do *corpus* CSTNews. Os pares de sentenças alinhadas foram classificados quanto à sobreposição de forma e de conteúdo entre eles, uma

vez que o objetivo da tipificação foi verificar o quanto de material linguístico superficial e profundo dos textos-fonte estavam presentes nos sumários. Assim, os rótulos foram divididos em dois tipos: forma e conteúdo.

Dado um par SD-SS, os anotadores indicaram a sobreposição entre elas com base em itens lexicais em comum, selecionando uma das 3 etiquetas: (i) idêntico, quando as duas sentenças fossem iguais, (ii) parcial, quando elas fossem semelhantes, isto é, tivessem alguns ou vários itens lexicais em comum, e (iii) diferente, quando elas tivessem alguns itens lexicais em comum.

Para indicar a sobreposição de conteúdo, os rótulos poderiam ser: (i) especificação, quando a SS continha alguma informação específica em relação ao conteúdo original da SD, (ii) generalização, quando o SS generalizava o conteúdo da SD, (iii) contradição, quando a SS e a SD apresentavam alguma informação contraditória, (iv) inferência, quando a SS expressava informação que foi inferida a partir da SD correspondente, (v) neutro, quando a SS não se enquadrava em nenhuma das opções anteriores, e (vi) outro, quando os anotadores não concordavam com o alinhamento anterior. A generalização e a especificação, na verdade, são duas operações de fusão entre documentos comumente utilizadas por humanos para condensar conteúdo de textos-fonte e produzir um sumário multidocumento.

Os anotadores também classificaram os alinhamentos com base na ocorrência de elementos onomásticos, ou seja, substantivos próprios. Os aspectos onomásticos foram divididos em dois tipos: (i) topônimos, quando nomes de lugares ocorriam nas sentenças alinhadas, e (ii) antropônimos, quando nomes de pessoas ocorriam nas sentenças alinhadas. O Quadro 11 apresenta todas as etiquetas usadas para caracterizar os alinhamentos.

Quadro 11 - Etiquetas utilizadas para caracterizar os alinhamentos

Tipos	**Subtipos**
Forma	Idêntico Parcial Diferente
Conteúdo	Especificação Generalização Contradição Inferência Neutro Outro
Onomástica	Topônimo Antropônimo

Fonte: Elaborado pelo autor.

Com base no conjunto de etiquetas do Quadro 11, a tipificação foi realizada levando-se em conta primeiramente a forma e posteriormente o conteúdo. Assim, os anotadores comparavam uma sentença inteira do sumário com a sentença do documento alinhada para decidir se elas eram idênticas, semelhantes ou completamente diferentes e escolhiam uma etiqueta para o alinhamento quanto à sua forma. Para as etiquetas de conteúdo, consideravam sintagmas (isto é, n-gramas) para decidir qual/quais subtipo/s estavam presentes, por isso, mais de uma etiqueta era permitida. Finalmente, se havia a presença de itens lexicais referentes à onomástica, colocavam a etiqueta "topônimo" e/ou "antropônimo", indicando a existência de tal relação.

No exemplo do Quadro 12, pode-se notar um alinhamento parcial quanto à forma, uma vez que as 2 sentenças têm algumas palavras em comum, mas não são idênticas. Além disso, identificaram-se algumas transformações de conteúdo entre elas. No caso, houve uma generalização, considerando que "vários estados" da SS é mais geral do que "Amazonas, Distrito Federal, Mato Grosso, Acre e Rondônia", o que é indicado no Quadro 12 em azul. Houve também uma especificação, considerando que “buscas e prisões” da SS especifica o item lexical “investigações” presente na SD, o que é indicado no Quadro 12 em verde. Além das etiquetas que denotam transformação de conteúdo, observou-se que alguns trechos permaneceram neutros, os quais são indicados no Quadro 12 em vermelho. No caso, o trecho “mais de 300 policiais federais” da SD foi transposto na íntegra para a SS. Outro caso de alinhamento neutro foi identificado entre o trecho “fazem parte das” da SD e o trecho “participaram das” da SS. Ademais, os anotadores também identificaram a presença de nomes de lugares (no caso, de estados brasileiros) e marcaram o alinhamento com a etiqueta "topônimo". Ao final, o alinhamento do Quadro 12 recebeu as etiquetas: (i) parcial, (ii) neutro, (iii) generalização, (iv) especificação, e (v) topônimo.

Quadro 12 - Exemplo de tipificação 1

Sentença do Documento
A PF divulgou que **mais de 300 policiais federais do Amazonas, Distrito Federal, Mato Grosso, Acre e Rondônia fazem parte das investigações** da "Operação Dominó”.
Sentença do Sumário
Mais de 300 policiais federais de **vários estados participaram das buscas e prisões** durante a operação.

Fonte: Elaborado pelo autor.

No exemplo apresentado no Quadro 13, o alinhamento recebeu a etiqueta "diferente" quanto à forma, porque ambas as sentenças não têm itens lexicais em comum. O fato de Messi e Riquelme terem sido os principais jogadores foi inferido pelo humano que sumarizou. Além disso, identificou-se a presença dos nomes dos jogadores, resultando na tipificação do alinhamento com a etiqueta “antropônimo”. Por conseguinte, o alinhamento no Quadro 13 recebeu os rótulos: (i) diferente, (ii) inferência e (ii) antropônimo.

Quadro 13 - Exemplo de tipificação 2.

Sentença do Documento	Sentença do Sumário
Após acompanhar o belo futebol apresentado pelos “hermanos” durante toda a Copa América, que foi conduzida pelos habilidosos pés de Riquelme e Messi, o Brasil foi a campo sem o tradicional status de favorito que o acompanha há muito.	O Brasil, mesmo sem duas de suas estrelas, bateu a Argentina, que tinha a melhor campanha do campeonato e contava com seus principais jogadores.

Fonte: Elaborado pelo autor.

No Quadro 14, o alinhamento recebeu: (i) a etiqueta “parcial” quanto à forma, pois ambas as sentenças têm algumas palavras em comum, (ii) o tipo “neutro” quanto ao conteúdo, já que ambas transmitem a mesma informação (o fato de que o Brasil pontuou e, por isso, não importa o ponto exato no jogo), e (iii) a etiqueta “especificação”, já que "4 minutos" é mais específico do que o "início do jogo".

Quadro 14 - Exemplo de tipificação 3.

Sentença do Documento	Sentença do Sumário
É verdade que o Brasil deu sorte de conseguir um gol logo no início da partida.	O Brasil conseguiu um gol logo nos primeiros 4 minutos do jogo, fazendo os argentinos apertarem o ataque no jogo, restando ao Brasil os contragolpes, chegando ao segundo gol, que foi um gol contra.

Fonte: Elaborado pelo autor.

3.3.1 Resultados da tipificação

Como resultado, a partir de um total de 1007 alinhamentos, identificaram-se 867 alinhamentos parciais (86%), 58 idênticos (5,7%) e 82 diferentes (8,1%). Quanto ao conteúdo,

foram identificados 949 alinhamentos neutros (94,2%), 37 de contradição (3,6%), 82 de generalização (8,1%), 48 de especificação (4,7%), 33 de inferência (3,2%) e 6 do subtipo "outro" (0,5%). Considerando os alinhamentos que apresentaram casos relacionados à onomástica, identificaram-se 4 toponímias (0,3%) e 20 antroponímias (1,9%).

Na Tabela 3, pode-se observar a ocorrência de todos os tipos e subtipos no *corpus*.

Tabela 3 - Distribuição dos tipos e subtipos de alinhamento no *corpus*

Tipos	**Subtipos**	**Quantidade no *corpus***	**Porcentagem**
Forma	Parcial	867	86%
	Idêntico	58	5,7%
	Diferente	82	8,1%
Conteúdo	Neutro	949	94,2%
	Contradição	37	3,6%
	Generalização	82	8,1%
	Especificação	48	4,7%
	Inferência	33	3,2%
	Outro	6	0,5%
Onomástica	Antroponímia	20	1,9%
	Toponímia	4	0,3

Fonte: Elaborada pelo autor.

A alta porcentagem de alinhamentos do tipo "parcial" (86%) justifica-se pelo fato de que essa etiqueta fora utilizada para caracterizar todos os pares que não eram efetivamente idênticos e os que não eram completamente diferentes.

Tendo em vista que o alinhamento é um processo por meio do qual se condensa conteúdo, há mais alinhamentos de generalização do que de especificação. Além disso, ressalta-se que dos 867 alinhamentos parciais, 714 foram classificados unicamente como neutro (70,9%), o que caracteriza pouca alteração de conteúdo.

Como no processo de alinhamento, calculou-se a medida *kappa* com base em 5 coleções. Considerando-se todas as etiquetas, obteve-se o valor de 0,452. Considerando-se apenas as etiquetas de forma, obteve-se o resultado da *kappa* de 0,717. Ao final, considerando-se apenas as etiquetas de conteúdo, a *kappa* obtida foi de 0,318. Todos os resultados *kappa* são apresentados na Tabela 4.

Tabela 4 - Resultados da medida kappa

Medida de concordância	**Forma e Conteúdo**	**Forma**	**Conteúdo**
Kappa (CARLETTA, 1996)	0.452	0.717	0.318

Fonte: Elaborada pelo autor.

Quanto à concordância relativa ao tipo conteúdo, o valor baixo da *kappa* (0,318) era esperado, dada a subjetividade da tarefa já mencionada na literatura.

Sobre a tipificação, ressalta-se que foi necessário rever várias vezes os critérios de anotação, pois a tarefa se mostrou mais complexa do que o esperado.

Em geral, pode-se afirmar que o conteúdo dos textos-fonte foi pouco alterado, posto que 818 alinhamentos (81,2%) receberam unicamente a etiqueta de conteúdo "neutro", mudando a forma de conteúdo. Consequentemente, os sumários do CSTNews são compostos por pouco material extraído integralmente dos textos-fonte, uma vez que apenas 58 alinhamentos de 1007 (5,7%) foram anotados como idênticos.

Ademais, pode-se observar que as generalizações são mais frequentes que as especificações, uma vez que podem ser vistas 82 generalizações (8,1%) e 48 especificações (4,7%) nos resultados. É possível explicar essa diferença considerando que generalizar uma informação é uma maneira de remover detalhes desnecessários e reduzir o conteúdo a fim de se criar um sumário.

Embora o processo de tipificação dos alinhamentos tenha sido realizado no âmbito desta pesquisa de mestrado, ressalta-se que ele não foi efetivamente utilizado para a execução dos objetivos traçados. Entretanto, as conclusões ora apresentadas servirão de base para futuras pesquisas que serão realizadas com o objetivo de caracterizar sumários humanos multidocumento.

Na sequência, apresenta-se a caracterização dos sumários quanto à seleção de conteúdo. A caracterização englobou a seleção dos atributos linguísticos e a subsequente identificação dos mesmos no *corpus* alinhado.

4 CARACTERIZAÇÃO DOS SUMÁRIOS MULTIDOCUMENTO

4.1 Seleção e descrição dos atributos

Após o alinhamento manual, todas as sentenças dos textos-fonte alinhadas (e não-alinhadas) aos sumários foram caracterizadas em função de alguns atributos linguísticos, os quais "traduzem" as estratégias/indícios de sumarização humana na SA mono e multidocumento. Consequentemente, os sumários multidocumento foram caracterizados de forma indireta. Para tanto, uma seleção preliminar desses atributos foi feita a partir da revisão literária. No Quadro 15, apresentam-se os atributos linguísticos relativos às estratégias e indícios de estratégias de seleção identificados na literatura que foram efetivamente utilizados neste trabalho para a caracterização dos resumos.

Quadro 15 - Conjunto de atributos linguísticos da literatura sobre sumarização humana

Estratégia/Indício	Atributo	Nível
Selecionar conteúdo com base no tamanho da sentença	Tamanho	Superficial
Selecionar conteúdo com base na redundância	Frequência	
Selecionar conteúdo com base na redundância	Palavra-chave	
Selecionar conteúdo com base na posição que ocupa no texto/parágrafo	Localização	
Selecionar conteúdo com base na redundância	Nº de relações CST	Profundo
Selecionar um dos textos-fonte como "base"	Fonte (de divulgação)	Extralinguístico

Fonte: Elaborado pelo autor.

Os atributos superficiais (i) "tamanho", (ii) "frequência", (iii) "palavra-chave" e (iv) "localização" são comprovadamente relevantes no cenário da sumarização humana monodocumento, com exceção do atributo "frequência" e "palavra-chave", cuja relevância também foi comprovada para a SHM (NENKOVA, 2006). O atributo profundo "nº de relações CST" é específico do cenário da SHM, assim como o atributo "fonte" (de divulgação).

Os 3 atributos identificados na literatura, mas não utilizados, foram: "expressões sinalizadoras", "componentes discursivos" e "frequência das palavras do título". O atributo "expressão sinalizadora" não foi utilizado porque este é mais relevante na seleção de conteúdo em textos científicos. Tendo em vista que o CSTNews é do gênero jornalístico, optou-se pelo descarte do atributo. A mesma justificativa aplica-se à desconsideração do atributo

"componentes discursivos". O atributo "palavra do título" foi excluído porque se constatou que nem todos os textos das coleções do CSTNews possuíam título ou subtítulo.

No caso, foram delimitados 4 atributos profundos com base na tipologia das relações CST: (i) "redundância", (ii) "complemento", (iii) "contradição" e (iv) "forma". Por meio do atributo "redundância", buscou-se especificar a quantidade de relações *Identity, Equivalence, Summary, Subsumption* e *Overlap* que uma sentença estabelece com outras na coleção.

O atributo "complemento" previu a especificação do número de relações *Historical background, Follow up* e *Elaboration* que uma sentença estabelece da coleção.

Por meio do atributo "contradição", por sua vez, buscou identificar a quantidade específica de relações *Contradiction* que a sentença sob análise estabelece.

Por fim, com base no atributo "forma", especificou-se a quantidade de relações C*itation, Attribution, Modality, Indirect speech* e *Translation* que uma sentença mantém com as demais da mesma coleção. Esse atributo, na realidade, busca identificar características textuais que são superficiais e não profundas. No entanto, o atributo "forma" foi considerado profundo porque sua especificação depende aqui das relações do modelo semântico-discursivo CST[24]. Ainda quanto ao atributo "forma", observa-se que, para as relações que expressam fonte/autoria e estilo, não foram previstos atributos específicos. A quantidade dessas relações referente às sentenças do *corpus* está prevista por um único atributo denominado "forma". Isso se deve ao fato de que o interesse maior na caracterização dos sumários reside nas relações de conteúdo do modelo CST.

Em suma, os atributos profundos, em oposição aos atributos superficiais, funcionam com base nas relações semântico-discursivas (de conteúdo) subjacentes ao texto-fonte. Ao final, delimitou-se o conjunto de atributos sistematizados no Quadro 16.

[24] Por essa razão, aliás, o atributo "forma" é marcado com o um asterisco (*) no Quadro 16.

Quadro 16 - Conjunto final de atributos para a caracterização dos sumários

Atributo	Nível
Tamanho Frequência Palavra-chave Localização	Superficial
Redundância Complemento Contradição Forma*	Profundo (via CST)
Fonte (de divulgação)	Extralinguístico

Fonte: Elaborado pelo autor.

Para a identificação dos atributos selecionados, optou-se por uma metodologia automática, que consistiu na utilização de uma ferramenta construída especificamente para essa tarefa pelos cientistas da computação do NILC[25], laboratório no qual este trabalho foi desenvolvido.

Para a identificação dos atributos, essa ferramenta realizou 2 tarefas de pré-processamento do *corpus*: a etiquetação morfossintática e a lematização. A etiquetação morfossintática (do inglês, *part-of-speech tagging* ou *tagging*) consiste em identificar automaticamente a categoria sintática das palavras, associando-se a cada uma delas uma etiqueta morfossintática (VOUTILAINEN, 2004). O etiquetador utilizado na ferramenta de descrição foi o MXPost (RATNAPARCKI, 1996). A lematização (do inglês, *lemmatizer*) é a redução de cada palavra de um texto ao seu lema ou forma canônica (formas não-marcadas) (SPARCK JONES; WILLET, 1997). Na lematização, os verbos são reduzidos ao *infinitivo* (p.ex.: casamos > casar) e os substantivos e adjetivos ao *masculino singular* (p.ex.: latas > lata/ feias > feio). No caso, utilizou-se um lematizador desenvolvido pelo NILC. Além de reunir um etiquetador e um lematizador, a ferramenta de descrição faz uso de uma *stoplist*, ou seja, uma lista de *stopwords*[26], palavras que devem ser excluídas da análise textual.

[25] A ferramenta em questão foi desenvolvida por Maria Lucía Castro Jorge, doutoranda do Instituto de Ciências Matemáticas e de Computação (ICMC-USP) e pesquisadora do NILC, com a supervisão do Prof. Thiago A. S. Pardo (ICMC-USP), que coorientou este trabalho.

[26] As *stopwords* são basicamente palavras funcionais (p.ex.: preposições, artigos, conjunções, etc.).

Além disso, ressalta-se que todos os atributos foram normalizados. A normalização é necessária para representar os valores numéricos igualmente em todas as coleções, visto que cada coleção possui quantidade diferente de palavras. Essa técnica permite reduzir as chances dos dados se tornarem inconsistentes quando comparados entre si. Na normalização, o valor obtido referente ao atributo de uma sentença em dada coleção foi dividido pelo maior valor obtido para esse atributo na mesma coleção. Diante disso, obtiveram-se os atributos "tamanho" e "tamanho-normalizado", por exemplo.

A seguir, apresentam-se as definições de cada um dos atributos. Ademais, descrevem-se especificamente os critérios de identificação dos atributos superficiais nos quais se pautou a ferramenta de descrição, já que a identificação dos atributos profundos e do atributo extralinguístico consistiu na recuperação de dados previamente armazenados no CSTNews.

4.1.1 Os atributos superficiais

a) Tamanho da sentença

Em um texto jornalístico, sentenças longas tendem a detalhar a notícia de forma a incluir informações secundárias, enquanto sentenças muito curtas podem conter somente uma informação secundária (p.ex.: dados numéricos). Uma sentença de tamanho médio tende a conter elementos relevantes de forma concisa e, portanto, geralmente selecionada por humanos para compor um sumário (p.ex.: SCHIFFMAN et al., 2002).

O tamanho das sentenças neste trabalho foi identificado pela quantidade de palavras de conteúdo presentes na sentença, pois estas denotam conteúdo semântico. Assim, esse cálculo requer a identificação da categoria sintática das palavras, realizada por um etiquetador, para que as palavras de classe fechada presentes na *stoplist* possam ser excluídas. Na sequência, apenas as palavras de classe aberta foram consideradas para o cálculo do tamanho da sentença. Por exemplo, para a sentença "Os outros ficarão em Rondônia.", retirada da coleção C9 do *corpus* CSTNews, o valor do atributo "tamanho" foi 2, já que, após a exclusão das *stopwords* "os", "outros" e "em", restaram apenas 2 palavras de conteúdo ("ficar"_verbo e "Rondônia"_nome).

Para o cálculo do valor do atributo "tamanho-normalizado", o valor 2 obtido para o atributo "tamanho" da sentença "Os outros ficarão em Rondônia", por exemplo, foi dividido por 43 (2/43), que é o maior valor do atributo "tamanho" na coleção C3. Assim, obteve-se o valor normalizado 0,046.

b) Frequência

O cálculo da frequência foi feito com base em todas as palavras de conteúdo dos textos. Esse cálculo segue o pressuposto de que a ideia principal de um texto pode ser expressa por alguns itens lexicais, os quais tendem a ocorrer mais vezes em um documento.

Especificamente, a caracterização das sentenças com base no atributo "frequência" consistiu na pontuação das mesmas pela soma do número de ocorrências, na coleção, de cada uma de suas palavras (de conteúdo) constitutivas. Assim, de início, calculou-se a frequência das palavras de todos os textos-fonte na coleção. Dada sentença *x* de um dos textos da coleção, recuperou-se a frequência somente de suas palavras constitutivas na lista de frequência de todas as palavras da coleção. Na sequência, somaram-se as frequências das palavras da sentença, obtendo-se o valor do atributo.

Para a especificação da frequência, a lematização é tarefa essencial, pois a redução das palavras de classe aberta a suas formas canônicas permite identificar automaticamente todas as ocorrências de uma mesma palavra, a qual antes estava marcada por acidentes flexionais. Por exemplo, a sentença do Quadro 17, retirada da coleção C3 do CSTNews, é composta pelas palavras de conteúdo lematizadas "TAM", "negar" e "hipótese", cujas frequências na coleção são 16, 1 e 2, respectivamente. Pela soma dessas ocorrências, obteve-se o valor 19 para o atributo "frequência" relativo à sentença "A TAM negou a hipótese". Caso essa sentença também fosse composta por outras palavras, como "negaram" e "hipóteses", a lematização permite identificar as mesmas como outras ocorrências das palavras "negar" e "hipótese", respectivamente, alterando a frequência das mesmas.

Para calcular o valor normalizado desse atributo para a sentença do Quadro 17, o valor 19 foi dividido pelo maior valor obtido para esse atributo na coleção, no caso, 110 (19/142). Assim, obteve-se o valor 0,13 para a frequência normalizada da sentença em questão.

Quadro 17 - Exemplo de cálculo do atributo "frequência"

Sentença	A TAM negou a hipótese.		
Palavras de classe aberta	TAM	negou	Hipótese
Palavras lematizadas	TAM	negar	Hipótese
Frequência	16	1	2
Pontuação	19		

Fonte: Elaborado pelo autor.

c) Palavra-chave

O atributo “palavra-chave” de uma sentença foi calculado pela soma das frequências de suas palavras (de conteúdo) mais recorrentes na coleção a que se enquadra. No caso, somente 10% das palavras mais frequentes da coleção foram consideradas para a especificação do valor desse atributo. As palavras que compõem o conjunto das 10% mais frequentes são consideradas “palavra-chave”, ou seja, carregam o conteúdo principal dos textos da coleção. Vale ressaltar que essa não é a concepção tradicional de “palavra-chave” da Linguística de *Corpus.* Para a LC, “palavra-chave” são itens lexicais mais distintivos do *corpus* sob análise, isto é, palavras cujas frequências são estatisticamente diferentes em um *corpus* de estudo em relação a um *corpus* de outro domínio ou de referência (BERBER SARDINHA, 2004, 2006). O limite de corte de 10% foi definido pela observação empírica de que as palavras das coleções que comumente não compunham o conjunto das 10% mais frequentes apresentavam frequência bastante baixa nas coleções, de 1 a 3.

Diante dessa especificação, o valor do atributo “palavra-chave” de dada sentença de uma coleção foi calculado levando-se em conta apenas as suas palavras constitutivas que estivessem no conjunto das palavras-chave da coleção. Por exemplo, para a coleção C1 do CSTNews, obteve-se o conjunto de palavras-chave apresentado no Quadro 18, as quais compõem o que se denominou *toplist.* Com base nessa lista, calculou-se, por exemplo, o valor do atributo “palavra-chave” para a sentença S8 “Em março, a União Européia proibiu quase todas as companhias aéreas do Congo de operar na Europa” de um dos textos da coleção. Tendo em vista que, das palavras de classe aberta que constituem essa sentença (“março”, “união”, “europeu”, “proibir”, “quase”, “companhia”, “aéreo”, “congo”, “operar” e “europa”), somente “companhia”, “aéreo” e “congo” estão na lista de palavras-chave ou *toplist,* obteve-se o valor 3 para o atributo em questão.

Para calcular o valor normalizado do atributo “palavra-chave” para a sentença S8 da coleção C1, o valor 3 foi dividido pelo maior valor obtido para esse atributo na coleção C1, no caso, 7 (3/7). Assim, obteve-se o valor 0,42 para a “palavra-chave” normalizada da sentença em questão.

Quadro 18 - *Toplist* da coleção C1 do *corpus* CSTNews

***Toplist* (10%)**	**Frequência**
Avião	10
Congo	9
porta-voz	7
Bukavu	6
Passageiro	4
República	4
Democrática	4
Aeroporto	4
Companhia	4
Haver	4
Aéreo	4
Acidente	4

Fonte: Elaborado pelo autor.

d) Localização

Para especificar o valor do atributo "localização", cuja relevância é comprovada na seleção de conteúdo a partir de textos jornalísticos, como os que compõem o CSTNews, estabeleceram-se 3 valores possíveis para esse atributo: (i) começo; (ii) meio e (iii) fim. O valor "começo" foi atribuído à primeira sentença de cada documento, o valor "fim" foi atribuído à última sentença de cada documento e o valor "meio", por conseguinte, foi atribuído ao restante das sentenças.

Na sequência, as relações do modelo CST cujas anotações no *corpus* permitiram a identificação dos valores relativos aos atributos profundos (i) "redundância", (ii) "complemento", (iii) "contradição" e (iv) "forma" são definidas e ilustradas com exemplos retirados no CSTNews.

4.1.2 Os atributos profundos

a) O atributo "redundância"

As relações discursivas de redundância do modelo CST expressam níveis diferentes de sobreposição de conteúdo e, por isso, codificam redundância total e parcial. As relações CST de redundância total são *identity, equivalence* e *summary* e as de redundância parcial são *subsumption, overlap* e *subsumption.*

Relação *Identity* (total) - as sentenças devem ser idênticas, conforme exemplo (14).

(14) S1: As vítimas do acidente foram 14 passageiros e três membros da tripulação.
S2: As vítimas do acidente foram 14 passageiros e três membros da tripulação.

Relação *Equivalence* (total) - as sentenças apresentam o mesmo conteúdo, mas expresso de forma diferente. Um exemplo dessa relação pode ser visto em (15).

(15) S1: “O dado concreto é que nós vamos fazer deste país um verdadeiro canteiro de obras em se tratando de infra-estrutura", disse.
S2: "Algumas (obras) já estão em andamento, outras vão começar a andar agora, outras ainda precisam de licenciamento".

Relação *Summary* (total), no exemplo (16), a sentença S1 apresenta o mesmo conteúdo que a sentença S2, mas de forma mais compacta. Nessa relação, deve haver diferença significativa de tamanho entre as sentenças.

(16) S1: Lula disse que o critério para o investimento nas cidades será técnico, não partidário.
S2: "O critério é eminentemente técnico, ou seja, eu não quero saber se o prefeito é do PFL, do PT, do PMDB, do PSDB, do PTB, do PR, do PC do B”.

Relação *Subsumption* (parcial) - S2 apresenta as informações contidas em S1 e informações adicionais. Vê-se um exemplo dessa relação em (17).

(17) S1: Os mesmos parlamentares fizeram, também, um conluio com o Ministério Público e com a Justiça do Estado de Rondônia.
S2: A PF (Polícia Federal) prendeu na manhã desta sexta-feira 23 pessoas suspeitas de envolvimento em esquema da Assembléia Legislativa do Estado de Rondônia para desvio de recursos públicos e influência indevida sobre Poder Judiciário, Ministério Público, Tribunal de Contas e Poder Executivo do Estado.

Relação *Overlap* (parcial) - S1 e S2 apresentam informações em comum e ambas apresentam informações adicionais distintas entre si, segundo o exemplo em (18).

(18) S1: SÃO PAULO - A pista principal do Aeroporto Internacional de São Paulo (Cumbica), em Guarulhos, será totalmente reformada em março de 2008, segundo informações do Ministério da Defesa anunciadas nesta segunda-feira, 6.

S2: O Ministério da Defesa anunciou nesta segunda-feira (6) que em março do ano que vem uma das pistas do Aeroporto de Guarulhos será fechada para reformas de seu trecho central.

b) O atributo "complemento"

As relações discursivas da categoria complemento codificam o relacionamento entre sentenças de textos distintos quanto ao conteúdo temporal e atemporal.

As relações de complemento temporal são *Historical background* e *Follow-up* e a relação de complemento atemporal é *Elaboration.*

Relação *Historical Background* (temporal) - S1 apresenta informações históricas/passadas sobre algum elemento presente em S2. Essa relação é explicitada em (19).

(19) S1: Em julho do ano passado, a média foi de 36 km no horário.
S2: O congestionamento esteve ainda maior às 9h, quando chegou a 113 km de extensão para uma média de 32 km.

Relação *Follow-up* (temporal) - S2 apresenta fatos que ocorrem após os acontecimentos em S1; os acontecimentos em S1 e em S2 devem ser relacionados e ter um espaço de tempo relativamente curto entre si. Um exemplo dessa relação pode ser visto em (20).

(20) S1: O discurso de Lula na ONU deu grande ênfase ao fim do protecionismo agrícola.
S2: Depois de Lula, foi a vez do presidente americano George W. Bush discursar na Assembléia Geral da ONU.

Relação *Elaboration* (atemporal) - S2 detalha/refina/elabora algum elemento presente em S1, sendo que S2 não deve repetir informações presentes em S, conforme exemplo (21).

(21) S1: O apresentador foi roubado por duas pessoas que estavam em uma moto logo após sair de um restaurante do bairro.
S2: A polícia de São Paulo afirmou ontem ter detido dois suspeitos de ter participado do roubo do relógio Rolex do apresentador Luciano Huck, da TV Globo, em um semáforo do Itaim Bibi (zona oeste de SP).

c) O atributo "contradição"

Na categoria "contradição" da tipologia de Maziero et al. (2010), há apenas uma relação, no caso, *Contradiction.*

Relação *Contradiction* – essa relação expõe que a sentença S1 e a sentença S2 divergem sobre algum elemento presente em ambas. No exemplo em (22), apresenta-se um par de sentenças ligadas por essa relação. No exemplo, retirado do *corpus* CSTNews, as sentenças se contradizem ao indicarem pessoas diferentes para o destaque de um mesmo evento.

(22) S1: O grande destaque da prova foi Nicolas Oliveira, quarto nadador brasileiro a cair na água.

S2: Thiago, que abriu o revezamento, foi o grande destaque do quarteto nas piscinas do Parque Aquático Maria Lenk.

d) Forma

As relações de apresentação/forma lidam com aspectos secundários da informação. Na subcategoria "apresentação/forma", estão *Attribution, Modality, Citation.* Na outra subcategoria, a de "estilo", estão as relações *Indirect speech* e *Translation.*

Relação *Attribution* - as sentenças S1 e S2 apresentam informação em comum, mas somente S1 atribui essa informação a uma fonte/autoria. Essa relação é apresentada no exemplo (23).

(23) S1: A polícia de São Paulo afirmou ontem ter detido dois suspeitos de ter participado do roubo do relógio Rolex do apresentador Luciano Huck, da TV Globo, em um semáforo do Itaim Bibi (zona oeste de SP).

S2: SÃO PAULO - Um homem suspeito de ter roubado o relógio Rolex do apresentador de televisão Luciano Huck foi detido na quarta-feira, 16, em Taboão da Serra, na Grande São Paulo.

Relação *Modality* - S1 e S2 apresentam informação em comum e em S2 a fonte/autoria da informação é indeterminada/relativizada/amenizada, conforme exemplo em (24).

(24) S1: O CGE (Centro de Gerenciamento de Emergências) da Prefeitura de São Paulo registrava oito pontos de alagamento na cidade, às 9h30 desta segunda-feira.

S2: Até 9h30m foram registrados oito pontos de alagamento, dois deles intransitáveis - na Marginal Pinheiros, na altura da Ponte João Dias, e na Marginal Tietê, no acesso à Rodovia dos Bandeirantes.

Relação *Citation* - S2 cita explicitamente informação proveniente de S1. Não há ocorrências dessa relação no *corpus* em questão.

Relação *Indirect-speech* - S1 e S2 apresentam informação em comum; S2 apresenta essa informação em discurso direto e S1 em discurso indireto. Em (25), apresenta-se um exemplo dessa relação.

(25) S1: Algumas já estão em andamento, outras vão começar a andar agora, outras ainda precisam de licenciamento.

S2: "Algumas (obras) já estão em andamento, outras vão começar a andar agora, outras ainda precisam de licenciamento".

Relação *Translation* - S1 e S2 apresentam informação em comum em línguas diferentes. Essa relação pode ser visualizada no exemplo em (26).

(26) S1: Quinze voluntários da ONG francesa Ação Contra a Fome (ACF) foram assassinados no nordeste do Sri Lanka, informou hoje um porta-voz da organização.

S2: Segundo um representante do grupo *Action Contre la Faim*, os corpos foram encontrados no escritório da organização.

A identificação dos atributos profundos “redundância”, “complemento”, “contradição” e “forma” consistiu na recuperação automática da anotação CST disponível no CSTNews para cada uma das sentenças do *corpus*.

4.1.3 O atributo extralinguístico “fonte”

Tendo em vista a hipótese de que a seleção de um dos textos-fonte de uma coleção como “base” pode ser influenciada por fatores extralinguísticos, como fonte, autoria, etc., optou-se por verificar se a seleção de conteúdo realizada pelos humanos na produção dos sumários do CSTNews revela predileção por alguma das fontes dos textos desse *corpus.* Assim, delimitou-se o atributo “fonte”, que foi classificado como extralinguístico. A sua identificação consistiu, assim como a dos atributos profundos, na recuperação da informação sobre as fontes dos textos já codificada no CSTNews.

4.2 Organização dos dados da caracterização

Os alinhamentos manuais e os valores dos atributos obtidos pela ferramenta automática de descrição foram organizados em arquivos no formato xls (Excel) para facilitar a análise manual da caracterização dos sumários e a possível análise automática dos dados por meio de algoritmos de Aprendizado de Máquina (AM).

Para melhor visualização dos alinhamentos e das caracterizações, os dados gerais obtidos neste trabalho foram divididos em 2 conjuntos, cada um deles ilustrado em uma tabela específica. Um desses conjuntos é formado pelos alinhamentos e atributos superficiais, como ilustram as Tabelas 5 e 6. O outro é composto pelos alinhamentos e pelos atributos profundos e o extralinguístico, como nas Tabelas 7 e 8.

Tomando-se como base a Tabela 5, a primeira coluna, denominada “Sentença”, registra as sentenças dos textos-fonte de uma coleção, identificadas pelo padrão de nomeação S(sentença), D(documento) e C(coleção). Na coluna “Sumário”, registram-se os alinhamentos efetivamente. O valor “sim” indica que a sentença do texto-fonte foi alinhada e o valor “não” indica o contrário.

Nas demais 9 colunas, especificam-se os diversos atributos superficiais (tamanho, palavra-chave, frequência e localização) e suas versões normalizadas. Na primeira linha da Tabela 5, por exemplo, observa-se que a sentença 1 (S1) do documento 1 (D1), pertencente à coleção 1 (C1), está alinhada ao sumário (sim) e possui 9 palavras de classe aberta e 5 palavras-chave. Na sequência, vê-se que a soma da frequência de todas as suas palavras de conteúdo totaliza 40, e está localizada no começo do documento.

Tabela 5 - Caracterização superficial da coleção C1

Alinhamento		**Atributo superficial**						
Sentença	**Sumário**	**Tamanho**	**Tamanho/ Norm**	**Palavra-chave**	**Palavra-chave/ Norm**	**Frequência**	**Frequência/ Norm**	**Localização**
S1_D1_C1	sim	9	0.47	5	0.71	40	0.63	começo
S2_D1_C1	não	13	0.68	4	0.57	43	0.6	meio
S3_D1_C1	sim	11	0.57	1	0.14	25	0.39	meio
S4_D1_C1	não	16	0.84	2	0.28	32	0.50	meio
S5_D1_C1	não	11	0.57	2	0.28	30	0.47	meio
S6_D1_C1	sim	11	0.57	2	0.28	24	0.38	meio
S7_D1_C1	não	13	0.68	4	0.57	33	0.52	meio
S8_D1_C1	não	10	0.52	2	0.28	24	0.38	meio
S9_D1_C1	não	2	0.10	0	0	2	0.03	meio
S10_D1_C1	não	16	0.84	1	0.14	27	0.42	fim
S1_D2_C1	sim	19	1	7	1	63	1	começo
S2_D2_C1	sim	8	0.42	3	0.42	26	0.41	meio
S3_D2_C1	sim	18	0.94	3	0.42	56	0.88	meio
S4_D2_C1	sim	8	0.42	1	0.14	22	0.34	meio
S5_D2_C1	não	9	0.47	2	0.28	33	0.52	meio
S6_D2_C1	sim	5	0.26	1	0.14	14	0.22	meio
S7_D2_C1	sim	16	0.84	5	0.71	57	0.90	fim
S1_D3_C1	sim	18	0.94	7	1	62	0.98	começo
S2_D3_C1	sim	8	0.42	3	0.42	26	0.41	meio
S3_D3_C1	sim	18	0.94	3	0.42	54	0.85	meio
S4_D3_C1	não	9	0.47	2	0.28	33	0.52	meio
S5_D3_C1	sim	5	0.26	1	0.14	14	0.22	meio
S6_D3_C1	sim	16	0.84	5	0.71	57	0.90	meio
S7_D3_C1	sim	8	0.42	1	0.14	22	0.34	fim

Fonte: Elaborado pelo autor.

Tabela 6 - Caracterização superficial da coleção C7

Alinhamento		Atributo superficial						
Sentença	**Sumário**	**Tamanho**	**Tamanho/ Norm**	**Palavra-chave**	**Palavra-chave/ Norm**	**Frequência**	**Frequência/ Norm**	**Localização**
S1_D1_C7	não	12	0.46	3	0.37	24	0.35	começo
S2_D1_C7	sim	20	0.76	8	1	49	0.72	meio
S3_D1_C7	sim	14	0.53	7	0.87	51	0.75	meio
S4_D1_C7	não	18	0.69	3	0.37	34	0.5	meio
S5_D1_C7	sim	10	0.38	4	0.5	33	0.48	meio
S6_D1_C7	sim	9	0.34	1	0.12	21	0.30	meio
S7_D1_C7	não	7	0.26	3	0.37	20	0.29	meio
S8_D1_C7	não	18	0.69	5	0.62	37	0.54	meio
S9_D1_C7	sim	26	1	7	0.87	68	1	meio
S10_D1_C7	sim	7	0.26	1	0.12	17	0.25	meio
S11_D1_C7	não	23	0.88	6	0.75	56	0.82	meio
S12_D1_C7	não	16	0.61	3	0.37	32	0.47	meio
S13_D1_C7	não	14	0.53	3	0.37	31	0.45	meio
S14_D1_C7	não	18	0.69	2	0.25	37	0.54	meio
S15_D1_C7	não	17	0.65	6	0.75	40	0.58	fim
S1_D2_C7	sim	18	0.69	4	0.5	34	0.5	começo
S2_D2_C7	não	17	0.65	2	0.25	31	0.45	meio
S3_D2_C7	sim	15	0.57	6	0.75	43	0.63	meio
S4_D2_C7	sim	12	0.46	7	0.87	49	0.72	meio
S5_D2_C7	não	16	0.61	5	0.62	34	0.5	meio
S6_D2_C7	sim	22	0.84	3	0.37	47	0.69	meio
S7_D2_C7	sim	13	0.5	4	0.5	39	0.57	meio
S8_D2_C7	sim	10	0.38	1	0.12	23	0.33	fim

Fonte: Elaborado pelo autor.

As Tabelas 7 e 8 ilustram a especificação dos atributos profundos e do extralinguístico das sentenças das coleções C1 e C7, respectivamente.

Tabela 7 - Caracterização profunda da coleção C1

Alinhamento		Atributo								
		Profundo								Extralinguístico
Sentença	Sumário	Redundância	Redundância/ Norm	Complemento	Complemento/Norm	Contradição	Contradição/ Norm	Forma	Forma/ Norm	Fonte
S1_D1_C1	sim	2	0.66	4	0.57	0	0	1	1	Folha de S. Paulo
S2_D1_C1	não	2	0.66	7	1	0	0	1	1	Folha de S. Paulo
S3_D1_C1	Sim	1	0.33	0	0	1	1	0	0	Folha de S. Paulo
S4_D1_C1	não	0	0	1	0.14	0	0	0	0	Folha de S. Paulo
S5_D1_C1	não	2	0.66	3	0.42	1	1	0	0	Folha de S. Paulo
S6_D1_C1	Sim	0	0	1	0.14	0	0	0	0	Folha de S. Paulo
S7_D1_C1	não	0	0	1	0.14	0	0	0	0	Folha de S. Paulo
S8_D1_C1	não	0	0	0	0	0	0	0	0	Folha de S. Paulo
S9_D1_C1	não	0	0	0	0	0	0	0	0	Folha de S. Paulo
S10_D1_C1	não	0	0	0	0	0	0	0	0	Folha de S. Paulo
S1_D2_C1	Sim	2	0.66	4	0.57	0	0	0	0	Estadão
S2_D2_C1	Sim	2	0.66	1	0.14	0	0	0	0	Estadão
S3_D2_C1	Sim	3	1	3	0.42	0	0	1	1	Estadão
S4_D2_C1	Sim	1	0.33	2	0.28	0	0	0	0	Estadão
S5_D2_C1	não	1	0.33	3	0.42	0	0	0	0	Estadão
S6_D2_C1	Sim	1	0.33	1	0.14	0	0	0	0	Estadão
S7_D2_C1	Sim	1	0.33	5	0.71	1	1	0	0	Estadão
S1_D3_C1	Sim	2	0.66	2	0.28	0	0	1	1	Jornal do Brasil
S2_D3_C1	Sim	2	0.66	1	0.14	0	0	0	0	Jornal do Brasil
S3_D3_C1	Sim	1	0.33	1	0.14	1	1	0	0	Jornal do Brasil
S4_D3_C1	não	1	0.33	5	0.71	0	0	0	0	Jornal do Brasil
S5_D3_C1	Sim	1	0.33	1	0.14	0	0	0	0	Jornal do Brasil
S6_D3_C1	Sim	2	0.66	3	0.42	0	0	0	0	Jornal do Brasil
S7_D3_C1	Sim	1	0.33	1	0.14	0	0	0	0	Jornal do Brasil

Fonte: Elaborado pelo autor.

Tabela 8 - Caracterização profunda da coleção C7

Alinhamento	Atributo									
	Profundo									Extralinguístico
Sentença	Sumário	Redundância	Redundância/ Norm	Complemento	Complemento/Norm	Contradição	Contradição/ Norm	Forma	Forma/ Norm	Fonte
S1_D1_C7	não	0	0	2	0.4	0	0	0	0	Estadão
S2_D1_C7	sim	3	1	2	0.4	0	0	0	0	Estadão
S3_D1_C7	sim	1	0.33	1	0.2	0	0	0	0	Estadão
S4_D1_C7	não	1	0.33	0	0	0	0	0	0	Estadão
S5_D1_C7	sim	1	0.33	2	0.4	0	0	0	0	Estadão
S6_D1_C7	sim	1	0.33	0	0	0	0	0	0	Estadão
S7_D1_C7	não	0	0	0	0	0	0	0	0	Estadão
S8_D1_C7	não	0	0	0	0	0	0	0	0	Estadão
S9_D1_C7	sim	1	0.33	1	0.2	0	0	0	0	Estadão
S10_D1_C7	sim	1	0.33	1	0.2	0	0	0	0	Estadão
S11_D1_C7	não	0	0	2	0.4	0	0	0	0	Estadão
S12_D1_C7	não	0	0	1	0.2	0	0	0	0	Estadão
S13_D1_C7	não	0	0	0	0	0	0	0	0	Estadão
S14_D1_C7	não	0	0	2	0.4	0	0	0	0	Estadão
S15_D1_C7	não	0	0	1	0.2	0	0	0	0	Estadão
S1_D2_C7	sim	1	0.33	0	0	0	0	0	0	O Globo
S2_D2_C7	não	1	0.33	0	0	0	0	0	0	O Globo
S3_D2_C7	sim	1	0.33	5	1	0	0	0	0	O Globo
S4_D2_C7	sim	2	0.66	4	0.8	0	0	0	0	O Globo
S5_D2_C7	não	1	0.33	1	0.2	0	0	0	0	O Globo
S6_D2_C7	sim	1	0.33	0	0	0	0	0	0	O Globo
S7_D2_C7	sim	1	0.33	4	0.8	0	0	0	0	O Globo
S8_D2_C7	sim	1	0.33	1	0.2	0	0	0	0	O Globo
S1_D1_C7	não	0	0	2	0.4	0	0	0	0	O Globo

Fonte: Elaborado pelo autor.

5 IDENTIFICAÇÃO E FORMALIZAÇÃO DAS ESTRATÉGIAS DE SHM

Para a identificação de possíveis estratégias de SHM, especialmente que revelam em quais atributos selecionados da literatura os humanos frequentemente se baseiam para identificar o conteúdo dos textos-fonte de uma coleção a ser levado ao sumário, fez-se primeiramente uma análise manual dos resultados da caracterização das sentenças alinhadas.

Posteriormente, os resultados da caracterização foram submetidos a um ambiente de AM, por meio do qual uma análise automática foi realizada.

Para as análises manual e automática dos atributos superficiais, apenas os valores normalizados foram considerados.

Na sequência, descreve-se a análise manual dos resultados da caracterização automática das sentenças dos textos-fonte.

5.1 Análise manual

5.1.1 Os atributos superficiais

Para a análise dos atributos “tamanho”, “palavra-chave” e “frequência”, os valores dos mesmos foram divididos em 5 intervalos, sendo que x representa o atributo: (i) $0 \leq x \leq 0{,}2$; (ii) $0{,}2 < x \leq 0{,}4$; (iii) $0{,}4 < x \leq 0{,}6$; (iv) $0{,}6 < x \leq 0{,}8$ e (v) $0{,}8 < x \leq 1$. Os resultados da análise manual desses 3 atributos estão ilustrados nos Gráficos 2, 3 e 4, respectivamente.

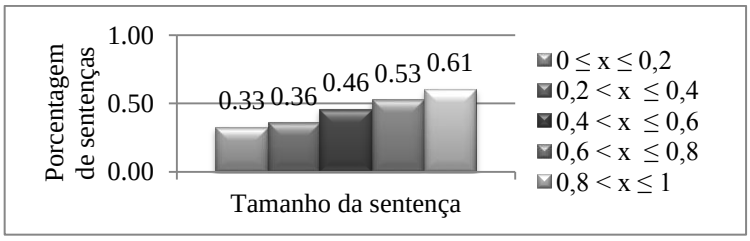
No Gráfico 2, vê-se que de todas as sentenças mais extensas (entre 0,8 e 1) dos textos-fonte, 61% foram alinhadas às sentenças dos sumários.

No Gráfico 3, vê-se que de todas as sentenças que continham um grande número de palavras-chave (entre 0,6 e 0,8), 70% foram alinhadas.

No Gráfico 4, observa-se que de todas as sentenças que continham o maior número de palavras mais frequentes (entre 0,8 e 1), 68% foram alinhadas.

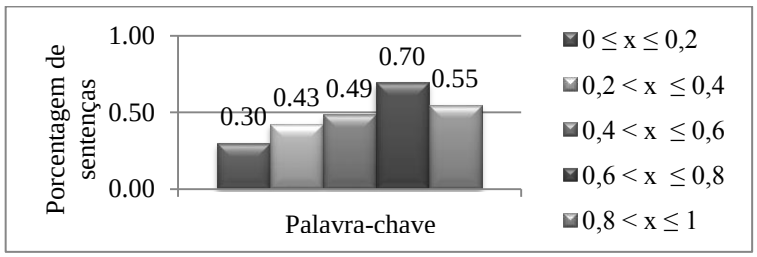
Quanto ao atributo “localização”, cujos resultados da análise manual estão ilustrados no Gráfico 5, vê-se que de todas as sentenças que estavam localizadas na parte inicial dos textos-fonte, principalmente na primeira sentença, 89% foram alinhadas.

Gráfico 2 - Atributo “tamanho”.



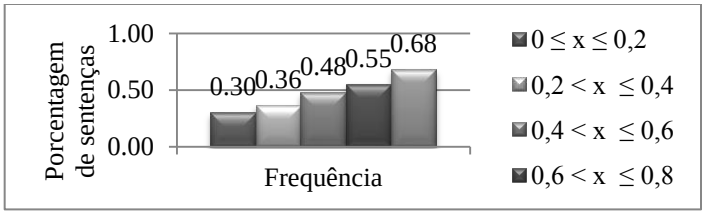
Fonte: Elaborado pelo autor.

Gráfico 3 - Atributo “palavra-chave”.



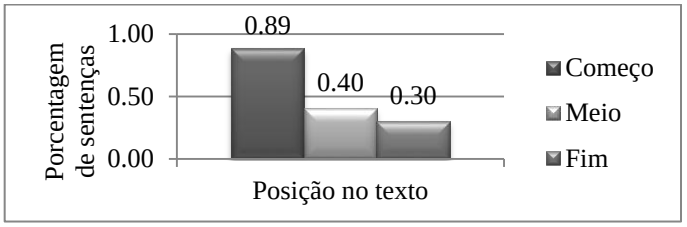
Fonte: Elaborado pelo autor.

Gráfico 4 - Atributo “frequência”.



Fonte: Elaborado pelo autor.

Gráfico 5 - Atributo “localização”.



Fonte: Elaborado pelo autor.

5.1.2 Os atributos profundos

Para a análise dos atributos profundos “redundância”, “complemento”, “contradição” e “forma”, apenas os valores absolutos (não normalizados) foram considerados, pois se

constatou que não há muita variação na quantidade de relações CST no *corpus*. Os resultados da análise manual desses 4 atributos estão ilustrados nos Gráficos 6, 7, 8 e 9, respectivamente.

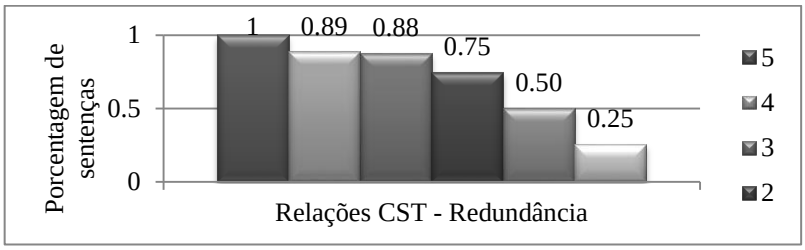
O Gráfico 6 ilustra a quantificação das sentenças alinhadas aos sumários que possuem entre 0 e 5 relações CST de redundância. No geral, observa-se que todas, ou seja, 100% das sentenças com 5 relações CST de redundância foram alinhadas.

No Gráfico 7, ilustram-se os resultados da análise do atributo "complemento". Nele, vê-se que as sentenças alinhadas possuem em média várias relações CST de complemento.

O Gráfico 8 registra os resultados da análise do atributo "contradição". Nele, observa-se que todas, isto é, 100% das sentenças com 3 relações CST de contradição foram alinhadas.

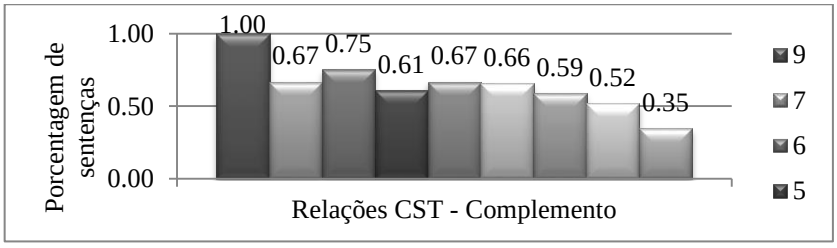
Por fim, no Gráfico 9, registram-se os resultados da análise manual do atributo "forma". Nele, destaca-se que todas as sentenças com 4 e 3 relações de forma foram alinhadas.

Gráfico 6 - Atributo "redundância".



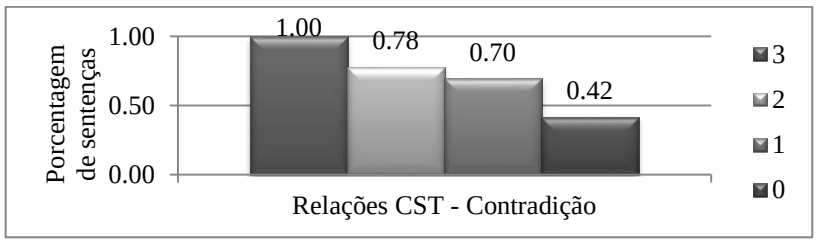
Fonte: Elaborado pelo autor.

Gráfico 7 - Atributo "complemento".



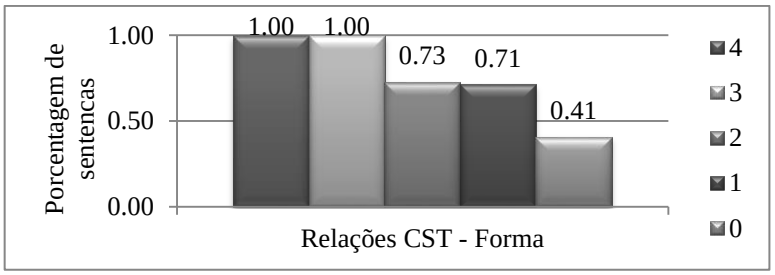
Fonte: Elaborado pelo autor.

Gráfico 8 - Atributo "contradição".



Fonte: Elaborado pelo autor.

Gráfico 9 - Atributo “forma”.

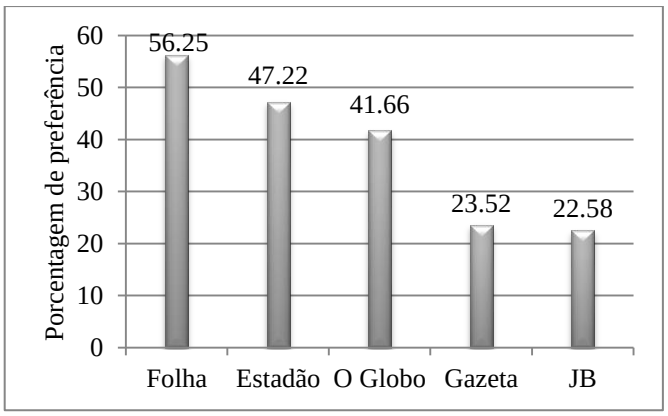


5.1.3 O atributo extralinguístico “fonte”

A análise do atributo “fonte” teve como objetivo verificar se os humanos apresentam alguma predileção por um ou mais veículos jornalísticos dos quais os textos-fonte do CSTNews foram compilados, já que a fonte é um possível atributo utilizado pelos humanos para selecionar o texto-fonte “base”.

Os dados resultantes da análise manual do atributo “fonte” normalizado estão sistematizados no Gráfico 10. Com base nas estatísticas apresentadas, observa-se que 56% das sentenças alinhadas são provenientes da fonte Folha de São Paulo. Nesse caso, pode-se dizer que os sumarizadores humanos, ao produzirem os sumários multidocumento do CSTNews, selecionaram na maioria das vezes os textos provenientes da Folha de São Paulo como “base”. A razão para essa predileção é difícil de ser apontada. Os humanos podem ter selecionado essa fonte por causa, por exemplo, do seu prestígio.

Gráfico 10 - O atributo “fonte”.



Fonte: Elaborado pelo autor.

A análise manual dos resultados das caracterizações dos sumários forneceram algumas pistas sobre os atributos mais relevantes na seleção de conteúdo. Por exemplo, os atributos

“redundância” e “localização” se destacaram na análise manual por apresentarem altíssima porcentagem em alguns de seus valores. No caso, o atributo “redundância” demonstrou que quanto mais relações dessa natureza uma sentença contém, maior a probabilidade de seu conteúdo ser selecionado, já que das sentenças que possuíam 5 relações de redundância, todas foram alinhadas ao sumário de sua coleção. Isso ocorre pelo fato de haver mais de um texto-fonte sob processamento e, por essa razão, as informações mais relevantes são repetidas em todos os textos.

O atributo “localização”, por sua vez, apresentou alta porcentagem (89%) de sentenças alinhadas provenientes do “começo” dos textos-fonte em detrimento das outras posições. Como já é sabido, as informações localizadas no início de textos do gênero jornalístico expressam o fato principal de uma notícia e, por isso, são selecionadas para compor o sumário. Tal fato, comprovado empiricamente em sumários provenientes de apenas 1 texto-fonte também ocorre em sumários advindos de múltiplos textos.

Na sequência, os mesmos dados foram analisados de forma automática, resultando em regras explícitas de seleção de conteúdo que evidenciam estratégias de SHM.

5.2 Geração de regras automáticas

A análise automática dos dados gerados pela caracterização dos sumários em função apenas dos atributos superficiais e profundos consistiu na geração de regras capazes de indicar os atributos que caracterizam o conteúdo das sentenças que foram alinhadas, revelando, assim, estratégias de seleção de conteúdo.

Os dados em questão foram submetidos ao ambiente Weka (*Waikato Environment for Knowledge Analysis*) (HALL et al., 2009). O Weka é um ambiente de aprendizado de máquina que disponibiliza um conjunto de algoritmos capazes de analisar automaticamente os dados de entrada e aprender padrões estatisticamente relevantes, os quais são expressos por regras. Para a geração automática de tais regras, optou-se por combinar os atributos superficiais e os profundos.

Os dados foram submetidos a vários algoritmos do Weka, pois cada um deles realiza o aprendizado automático de uma forma específica, gerando diferentes tipos de regras. Os algoritmos testados foram: JRip, J48, PART, Prism e OneR.

Os algoritmos JRip, J48, PART e Prism foram testados de diversas maneiras, variando, no caso, (i) o número de casos/ocorrências dos atributos necessário para validar uma regra e (ii) o número de atributos. Após vários testes, optou-se por considerar os resultados

gerados apenas pelo JRip com base na utilização de (i) todos os atributos superficiais, profundos e o extralinguístico (tamanho, palavra-chave, frequência, localização, redundância, complemento, contradição, forma e fonte) e (ii) número mínimo de casos de alinhamento necessários para validar as regras (ou seja, 2).

No caso, o JRip inferiu ou aprendeu um conjunto formado por 11 regras, quantidade passível de ser manipulada e analisada manualmente, as quais obtiveram 71, 25% de precisão. Obteve-se essa taxa considerando que as regras foram geradas pelo treinamento no CSTNews e teste no próprio *corpus*.

As regras geradas pelo JRip estão no formato de expressões lógicas do tipo *se, então* e, por isso, estão representadas de forma explícita e não ambígua. Além disso, ressalta-se que essas regras evidenciam efetivamente estratégias de seleção de conteúdo na SHM.

No Quadro 19, apresentam-se as 11 regras geradas pelo JRip, as quais devem ser interpretadas de forma sequencial. No quadro, as regras estão seguidas pelo número de casos classificados correta e erroneamente.

Quadro 19 - Regras geradas pelo algoritmo JRip.

Regras	**Acertos/Erros**
1. Se Localização = **começo**, então Sumário=**sim**	(140/16)
2. Senão se Redundância = (0.9-1), então Sumário= **sim**	(81/11)
3. Senão se Redundância = (0.6-0.7), então Sumário= **sim**	(68/12)
4. Senão se Redundância = (0.3-0.4), então Sumário= **sim**	(172/76)
5. Senão se Redundância = (0.7-0.8), então Sumário= **sim**	(46/7)
6. Senão se Redundância = (0.4-0.5), então Sumário= **sim**	(197/88)
7. Senão se Redundância = (0.2-0.3) e Frequência = (0.5-0.6), então Sumário= **sim**	(35/9)
8. Senão se Redundância = (0.1-0.2) e Frequência = (0.4-0.5), então Sumário= **sim**	(10/2)
9. Senão se Redundância = (0.1-0.2) e Tamanho = (0.2-0.3), então Sumário= **sim**	(12/2)
10. Senão se Tamanho (0.1-0.2) e Frequência (0.3-0.4) então, Sumário= **sim**	(14/3)
11. Senão Sumário=**não**	(1305/346)

Fonte: Elaborado pelo autor.

Para ilustrar a interpretação das regras, consideram-se as regras 1, 2 e 3.

Por meio da regra 1, vê-se que o JRip aprendeu que uma sentença do texto-fonte que apresenta o valor “começo” para o atributo “localização” alinha-se ao sumário, indicando que

seu conteúdo foi selecionado para compor o sumário. Com base na regra 1, o algoritmo realizou 140 acertos e 16 erros.

Caso o atributo “localização” apresente um valor diferente (meio ou fim), o algoritmo aplica a regra 2 na sequência. Em outras palavras, se o valor do atributo “localização” for diferente de “começo” e o valor do atributo “redundância” estiver entre 0.9 e 1, então a sentença alinha-se ao sumário. De acordo com essa segunda regra, o JRip identificou 81 casos corretamente e 11 casos foram classificados de forma errada.

Caso o valor do atributo “localização” seja diferente de “começo” e o valor do atributo “redundância” esteja fora do intervalo 0.9-1, o algoritmo aplica a regra 3 na sequência, ou seja, diante de “redundância” com valor entre 0.6-0.7, a sentença alinha-se ao sumário.

No Quadro 20, encontra-se a “matriz de confusão” produzida pelo algoritmo, ou seja, um quadro com os erros e acertos das regras aprendidas pelo algoritmo. Nesse quadro, observa-se que, das 1185 sentenças dos textos-fonte que não foram alinhadas, as regras do JRip classificaram 956 delas corretamente e 229 erroneamente. Ainda, do total de 895 sentenças dos textos-fonte que foram alinhadas aos sumários, o algoritmo identificou 526 delas corretamente e 369 erroneamente. Com base nesse desempenho, conclui-se que o algoritmo classificou corretamente mais casos de não-alinhamento que de alinhamento e isso se justifica pelo fato de que há mais casos de não-alinhamento no *corpus*, a partir dos quais ele aprendeu as regras.

Quadro 20 - Matriz de confusão do algoritmo JRip.

Classe / Teste	Alinhado (895) “Sumário=sim”	Não-alinhado (1185) “Sumário=não”
Alinhado	526	369
Não-alinhado	229	956

Fonte: Elaborado pelo autor.

Ao final, quanto às análises manual e automática, algumas observações são feitas:

a) os atributos mais relevantes das sentenças alinhadas são “localização”, “redundância”, “frequência (das palavras da coleção)” e “tamanho”;
b) os atributos “localização”, “frequência” e “tamanho”, identificados na sumarização monodocumento, também parecem guiar a seleção de conteúdo na SHM;
c) o atributo “redundância”, formulado com base no indício de que as informações mais redundantes da coleção são selecionadas para compor o sumário, parece se confirmar; isso

pode ser notado pelo fato de que 8 das 11 regras do JRip pautam-se nesse atributo, isoladamente ou em conjunto com outros;

d) o atributo “tamanho”, juntamente com “redundância” na regra 9, possui valores entre 0,2 e 0,4 e, juntamente com “frequência” na regra 10, possui valores entre 0,2 e 0,1; isso evidencia que o conteúdo selecionado para compor os sumários é preferencialmente proveniente de sentenças médias ou curtas;
e) a relevância do atributo “localização” justifica-se pelo fato de estar intimamente ligado à macroestrutura dos textos jornalísticos, na qual a informação inicialmente fornecida aos leitores corresponde ao *lead*, ou seja, informação principal;
f) caso nenhuma das regras se aplique, o algoritmo identificou como padrão “o não-alinhamento da sentença”, o que é evidenciado pela última regra do conjunto; essa inferência do algoritmo é resultante do fato de haver muito mais casos de não alinhamento nos dados por ele analisado que de alinhamentos.

A seguir, apresentam-se as diversas formas de avaliação das estratégias de seleção de conteúdo que foram aqui identificadas.

6 AVALIAÇÃO

Diante da complexidade da avaliação extrínseca, optou-se aqui por aplicar apenas a intrínseca. Para tanto, realizaram-se 2 das 3 possibilidades de avaliação intrínseca, mencionadas em 2.6: (i) a verificação da ocorrência das estratégias em outro *corpus*, distinto do utilizado para o aprendizado das estratégias, o qual é comumente denominado "*corpus* de teste", (ii) a comparação da qualidade de sumários produzidos pelas estratégias em questão com a qualidade de sumários produzidos por estratégias diferentes.

O processo de avaliação em (i) foi realizado durante um estágio de 2 meses na Universidade de Oslo (Noruega), orientado pela Profa. Dra. Diana Santos. Tal avaliação consistiu na verificação da ocorrência das estratégias aprendidas no *corpus* de treinamento (CTSNews) em um *corpus* de teste.

A avaliação apresentada em (ii) foi realizada posteriormente e consistiu na comparação da qualidade de sumários gerados pelas estratégias de SHM com a qualidade de sumários gerados com base em outras estratégias/métodos.

Na sequência, apresenta-se a descrição do *corpus* de teste, o qual foi criado especificamente para as avaliações em (i) e (ii).

6.1 Descrição do *corpus* de teste

O *corpus* de teste é composto por 10 coleções de textos jornalísticos, sendo que cada coleção contém: (i) 3 textos sobre um mesmo assunto ou tema compilados de diferentes fontes jornalísticas; (ii) 3 sumários humanos multidocumento, cada um deles produzido por um linguista computacional distinto; (iii) alinhamento sentencial entre textos-fonte e sumários, e (iv) anotação dos textos-fonte segundo a teoria/modelo linguístico-computacional CST.

Objetivando construir um *corpus* de teste com as mesmas propriedades do *corpus de* treinamento, os textos do *corpus* de teste foram compilados de acordo com os critérios adotados para a construção do CSTNews. Tais critérios foram: (i) fontes confiáveis, (ii) diversidade de temas (seções dos jornais) e (iii) notícias do tipo monoevento, ou seja, que abordam um único tema ou assunto.

Assim, os textos foram compilados de algumas das principais fontes jornalísticas *online* do Brasil, as mesmas selecionadas no CSTNews, a saber: Folha de São Paulo, Estadão, Jornal do Brasil, O Globo e Gazeta do Povo. A coleta manual foi feita durante aproximadamente 7 dias e as notícias coletadas foram veiculadas entre janeiro e fevereiro de

2013. Ademais, as notícias foram agrupadas em função dos assuntos que veiculam e os grupos ou coleções foram nomeados de acordo com as "seções" dos jornais das quais as notícias foram compiladas. Assim, o *corpus* é composto por coleções das seguintes categorias: "esporte" (2 coleções), "mundo" (2 coleções), "dinheiro" (1 coleção), "política" (2 coleções), "ciência" (1 coleção) e "cotidiano" (2 coleções).

A produção dos sumários humanos multidocumento do *corpus* de teste também seguiu os critérios adotados na construção dos sumários do CSTNews. Assim, o *corpus* de teste possui sumários construídos manualmente de forma abstrativa (isto é, com reescrita do material dos textos-fonte), segundo uma taxa de compressão de 70%. Essa taxa de compressão significa que os sumários contêm, no máximo, 30% do número de palavras do maior texto-fonte da coleção. Para a produção dos sumários, contou-se com uma equipe de 12 linguistas computacionais (linguistas e informatas) com experiência em Linguística de *Corpus*, os quais compõem o grupo de pesquisa em Sumarização Automática do NILC.

O alinhamento entre as sentenças dos textos-fonte e as sentenças dos 3 sumários foi realizado por 2 anotadores da área de Linguística Computacional e cada anotador ficou responsável por alinhar 5 coleções distintas. O alinhamento seguiu as diretrizes utilizadas na anotação do CSTNews (cf. seção 3.2). No caso, o critério principal para o alinhamento foi a sobreposição de conteúdo. Assim, se uma sentença de um texto-fonte foi alinhada a uma ou mais sentenças do sumário, isso indica que o seu conteúdo foi selecionado para compor o sumário.

Como mencionado, cada coleção do *corpus* de teste possui 3 sumários humanos distintos, cada um deles produzido por um especialista diferente. Tendo em vista o objetivo de se identificar estratégias de seleção de conteúdo, verificaram-se: (i) a concordância entre os sumarizadores humanos quanto à seleção de informação, (ii) a informatividade dos sumários de referência e (iii) a relevância estatística dos atributos. Ressalta-se que as verificações em (i), (ii) e (iii) foram realizadas durante o estágio de pesquisa no exterior, sob a orientação da Prof. Dra. Diana Santos.

a) <u>Investigação da concordância</u>

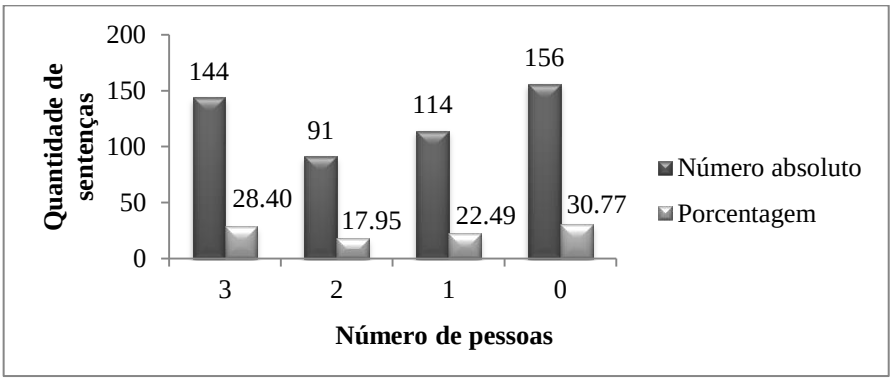
A concordância, em especial, é relevante para verificar o grau de subjetividade de uma tarefa. Assim, quanto mais alta é a concordância entre os humanos, mais bem delimitada é a tarefa e, por isso, computacionalmente viável. No caso, a concordância foi calculada de 3 formas distintas (Conc1, Conc2 e Conc3).

Em Conc1, verificou-se o quanto os humanos concordavam no geral sobre o conteúdo selecionado e sobre o conteúdo não selecionado para compor os sumários. Verificou-se a porcentagem das sentenças dos textos-fonte que foram alinhadas a 3 sumários, 2, 1 e 0, ou seja, que foram selecionadas por 3 humanos, 2, 1 e nenhum anotador.

No Gráfico 11, apresenta-se a quantificação da concordância geral entre os sumarizadores humanos. Nele, verifica-se que, do total de 505 sentenças dos textos-fonte do *corpus* de teste: (i) 144 sentenças (28,4%) foram alinhadas a 3 sumários diferentes, isto é, o conteúdo dessas sentenças foi selecionado pelos 3 sumarizadores humanos; (ii) 91 sentenças (17,95%) foram alinhadas a 2 sumários diferentes, ou seja, o conteúdo dessas sentenças foi selecionado por 2 sumarizadores humanos; (iii) 114 sentenças (22,49%) foram alinhadas a 1 único sumário, isto é, o conteúdo dessas sentenças foi selecionado por apenas 1 dos sumarizadores humanos, e (iv) 156 sentenças (30,77%) não foram alinhadas a nenhum sumário, ou seja, o conteúdo dessas sentenças não foi selecionado para compor os sumários.

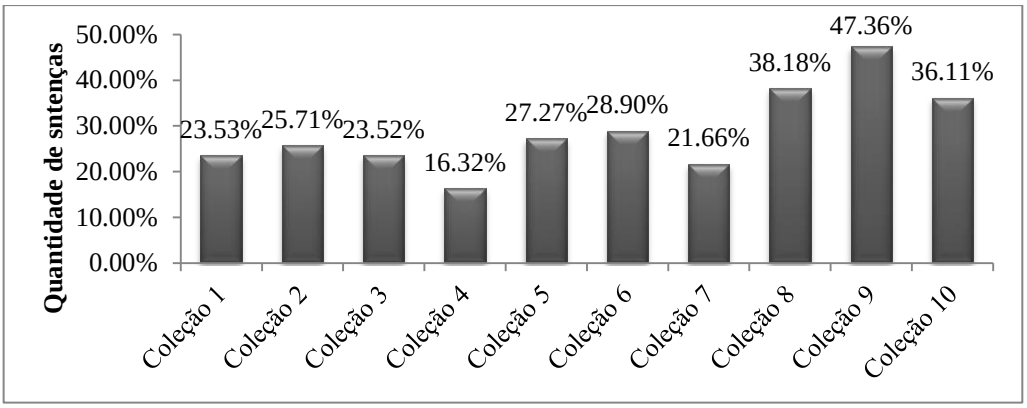
Assim, somando-se a concordância sobre o que foi selecionado pelos 3 especialistas e sobre o que não foi selecionado por nenhum deles, tem-se a porcentagem de 59,17% (=28,4% + 30,77%).

Gráfico 11 - Concordância geral.



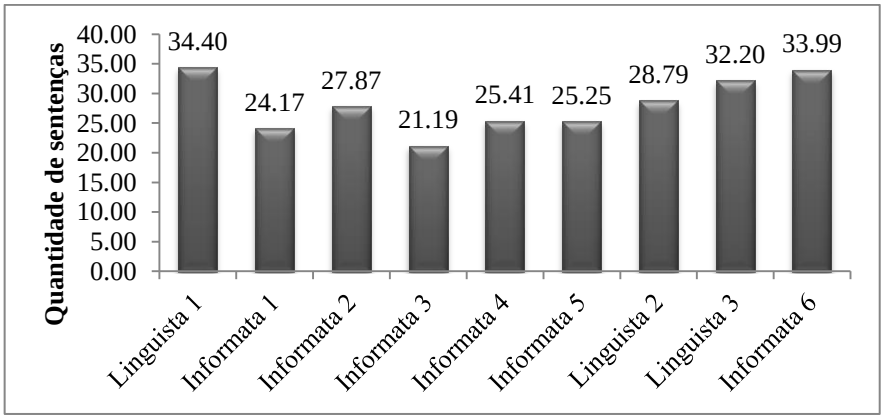
Em Conc2, analisou-se a concordância entre os humanos em cada uma das 10 coleções do *corpus* de teste, com o objetivo de observar se ela era uniforme ou não entre as coleções. No Gráfico 12, observa-se que os humanos menos concordaram quanto ao conteúdo selecionado para a compor o sumário da coleção 4 (16,32%), cuja temática é "fusão de empresas" (categoria "dinheiro"). A concordância mais alta refere-se ao conteúdo selecionado para compor o sumário da coleção 9 (47,36%), que versa sobre o senador Eduardo Suplicy e integra a categoria "política". A média da concordância por coleção foi de 28,86%, com um desvio padrão de 9,19 pontos para mais ou para menos.

Gráfico 12 - Concordância por coleção.



Por fim, em Conc3, calculou-se a concordância por especialista linguista e informata, com o objetivo de verificar se estes concordavam muito ou pouco com os demais que sumarizaram as mesmas coleções. No Gráfico 13, vê-se que a maior porcentagem de concordância entre as pessoas que sumarizaram a mesma coleção (34,4%) aponta para um linguista. Quanto à menor porcentagem (21,19%), as estatísticas indicam que um informata concordou menos vezes com os demais. A média da concordância por pessoa resultou em 28,14%, resultado que demonstra valor semelhante à concordância por coleção, porém, com menor desvio, podendo oscilar 4,61 pontos para mais ou para menos.

Gráfico 13 - Concordância por sumarizador humano.



Com base nos Gráficos 11, 12 e 13, observa-se que a concordância confirma o que está na literatura, ou seja, os humanos não concordam com toda a informação selecionada, mas com apenas a principal. Se os humanos concordam pouco quanto à seleção do conteúdo, isso significa que os critérios nos quais se baseiam variam muito e, portanto, é difícil identificar estratégias que sejam extremamente recorrentes.

b) Investigação da informatividade dos sumários do *corpus* de teste

Além das diferentes concordâncias quanto à seleção de conteúdo, buscou-se analisar o nível de informatividade dos sumários do *corpus* de teste. Essa tarefa objetivou verificar se, no interior de uma coleção, um dos 3 sumários de referência se destacava em função de sua informatividade.

Essa análise foi feita de forma manual por 4 juízes com diferentes níveis de proficiência em português, a saber: 3 falantes nativos do português (2 brasileiros e 1 português) e 1 norueguês com nível avançado de proficiência na língua em questão. Dada uma coleção, cada juiz estabeleceu uma escala para os sumários de referência de acordo com sua informatividade, não levando em conta a leitura dos textos-fonte. A análise foi feita em apenas 3 coleções do *corpus* de teste, escolhidas aleatoriamente: coleção 2 (cotidiano), coleção 4 (dinheiro) e coleção 6 (esporte).

No Quadro 21, apresentam-se as escalas produzidas por cada juiz para as 3 coleções escolhidas. Nesse quadro, o padrão "x > y >z" indica que o sumário x é o melhor da coleção, seguido por y e z. O asterisco (*) indica que o juiz não estabeleceu a escala porque não conseguiu distinguir o nível de informatividade dos sumários.

Com base no Quadro 21, observa-se que os juízes não concordaram totalmente quanto às escalas dos sumários. No entanto, observam-se algumas concordâncias parciais quanto às coleções 4 e 6. Para a coleção 4, o juízes 2 e 4 concordaram totalmente com a escala dos 3 sumários (2>1>3) e os 4 juízes concordaram que o sumário 2 é melhor que o 3. Para a coleção 6, os juízes 1 e 4 concordaram totalmente com a escala dos 3 sumários (1>3>2) e todos os juízes concordaram que o sumário 1 é melhor que o 3. Sobre a coleção 2, os juízes discordaram totalmente quanto à escala dos sumários em função da informatividade.

Quadro 21 - Resultados da avaliação.

Juízes	**Coleção 2**	**Coleção 4**	**Coleção 6**
Juiz1	3 > 2 > 1	1 > **2 > 3**	**1 > 3 > 2**
Juiz 2	1 > 2 > 3	**2 > 1 > 3**	2 > 1 > 3*
Juiz 3	1 > 3 > 2	**2 > 3** > 1	*
Juiz 4	2 > 3 > 1	**2 > 1 > 3**	**1 > 3 > 2**

Fonte: Elaborado pelo autor.

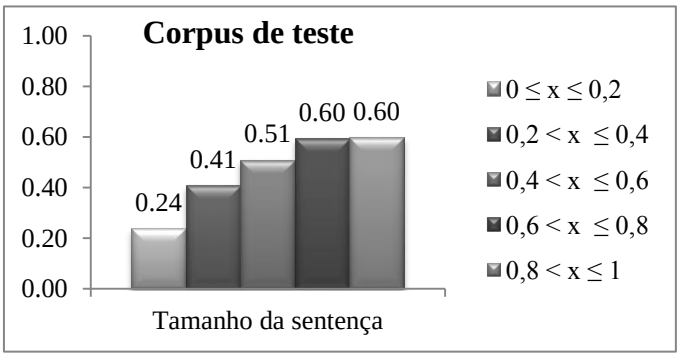
Generalizando os resultados da avaliação, considera-se que, dada uma coleção do *corpus* de teste, os 3 sumários são suficientemente informativos, não havendo um que claramente se destaca frente aos demais.

c) Análise da relevância estatística dos atributos no *corpus* de teste

A caracterização das sentenças alinhadas dos textos-fonte do *corpus* de teste também foi analisada manualmente, buscando verificar se a relevância estatística dos atributos era similar à do *corpus* de treinamento. Comparando-se os Gráficos 14, 15, 16, 17, 18, 19, 20 e 21 aos Gráficos 2, 3, 4, 5, 6, 7, 8 e 9 (cf. seções 5.1.1 e 5.1.2) respectivamente, observa-se que:

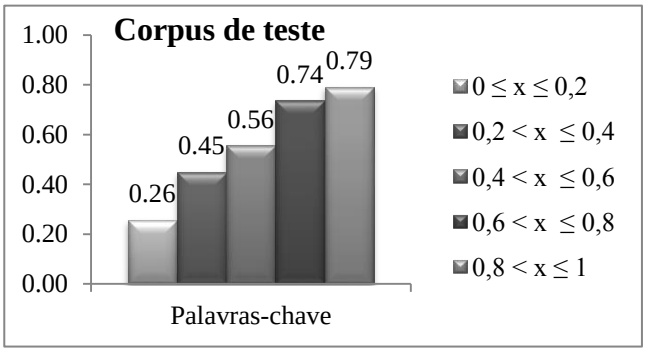
a. O tamanho das sentenças dos textos-fonte alinhadas às sentenças dos sumários em ambos os *corpora* são extensas, o que demonstra a ocorrência das características encontradas manualmente no CSTNews em um novo conjunto de textos.
b. 68% das sentenças do CSTNews que foram alinhadas contêm palavras muito frequentes (entre 0,8 e 1). O fato se replica para o *corpus* de teste, pois das sentenças que possuíam muitas palavras frequentes (entre 0,8 e 1), 75% foram alinhadas aos sumários.
c. O atributo "palavra-chave" de 71% das sentenças alinhadas está entre 0,8 e 1. Seguindo a mesma proporção, o *corpus* de teste apresenta dados semelhantes, isto é, que 79% das sentenças de seus textos-fonte alinhadas aos sumários possuem grande quantidade de palavras-chave, também pertencentes ao intervalo entre 0,8 e 1.
d. Quanto ao atributo "localização", vê-se que 89% das sentenças do CSTNews alinhadas estão localizadas na parte inicial dos textos-fonte. Tal fato se confirma quando se observa os dados do *corpus* de teste, em que 95% de todas as sentenças que estão localizadas no começo do documento foram alinhadas aos sumários.
e. Quanto à redundância, observa-se no *corpus* CSTNews que todas, ou seja, 100% das sentenças com 5 relações CST de redundância (número máximo de relações no *corpus*) foram alinhadas. Similarmente, observa-se no *corpus* de teste que todas as sentenças que possuem 5 ou 7 relações de redundância foram alinhadas aos sumários.
f. Quanto aos resultados da análise do atributo "contradição", observa-se que todas, isto é, 100% das sentenças com 3 relações CST de contradição do *corpus* CSTNews e do *corpus* de teste foram alinhadas.
g. Quanto ao atributo "complemento", vê-se que tanto no *corpus* CSTNews quanto no *corpus* de teste as sentenças alinhadas possuem em média várias relações CST de complemento.
h. Quanto ao atributo "forma", destaca-se que todas as sentenças com 4 e 3 relações de forma do *corpus* CSTNews foram alinhadas e que 93% das sentenças com 3 relações de forma do *corpus* de teste foram alinhadas.

Gráfico 14 - Tamanho da sentença.



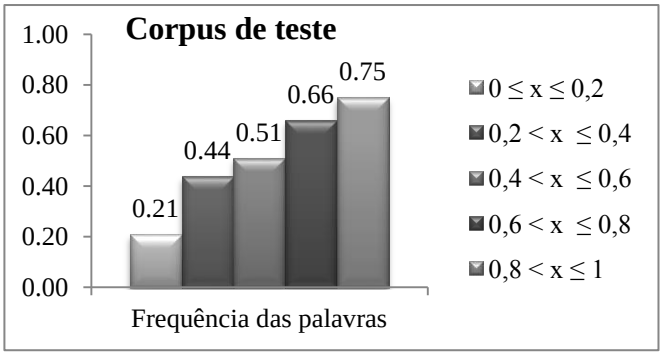
Fonte: Elaborado pelo autor.

Gráfico 15 - Palavra-chave.



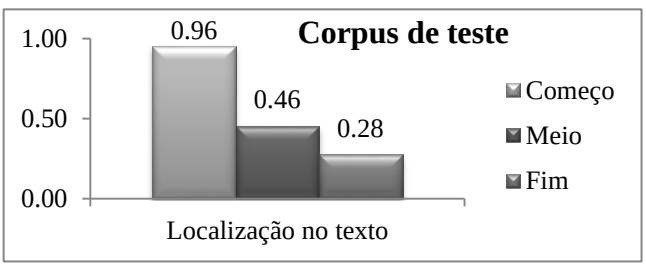
Fonte: Elaborado pelo autor.

Gráfico 16 - Frequência.



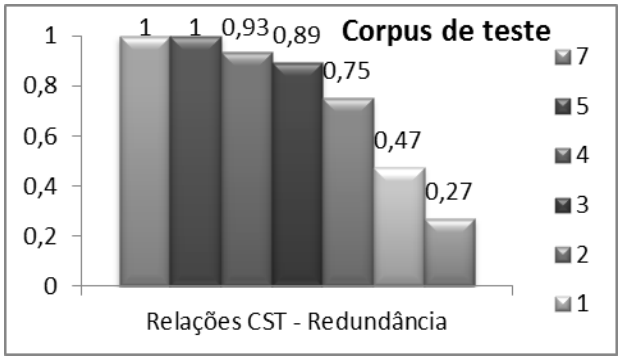
Fonte: Elaborado pelo autor.

Gráfico 17 - Localização.



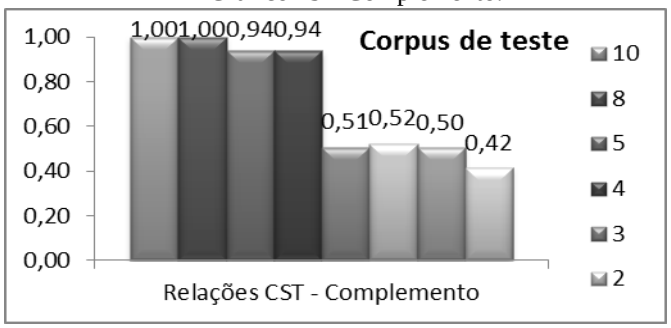
Fonte: Elaborado pelo autor.

Gráfico 18 - Redundância.



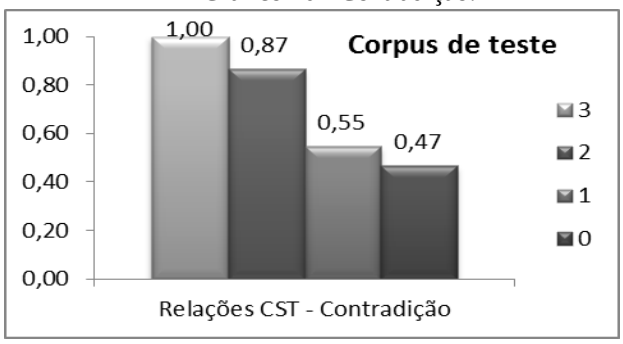
Fonte: Elaborado pelo autor.

Gráfico 19 - Complemento.



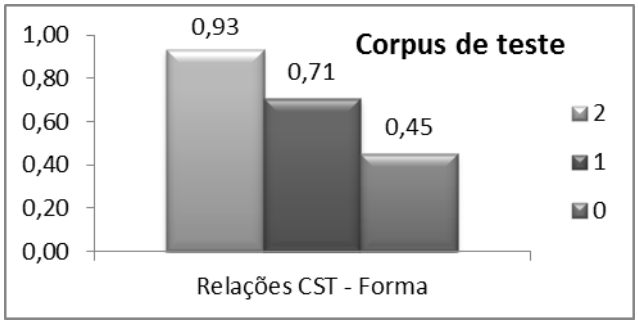
Fonte: Elaborado pelo autor.

Gráfico 20 - Contradição.



Fonte: Elaborado pelo autor.

Gráfico 21 - Forma



Fonte: Elaborado pelo autor.

Diante da comparação estatística ente os *corpora*, constata-se que a relevância dos atributos do CSTNews se replica a um novo conjunto de textos, o que indica que os atributos selecionados são comprovadamente pertinentes ao cenário da sumarização humana multidocumento.

Na sequência, apresenta-se uma das avaliações intrínsecas das estratégias de seleção de conteúdo, a qual consistiu na verificação da ocorrência das estratégias de SHM no *corpus* de teste descrito aqui.

6.2 Avaliação das estratégias em um *corpus* de teste

Essa verificação foi feita por meio do ambiente Weka, seguindo-se especificamente os passos previstos pelo ambiente:

(i) *upload* dos arquivos com a caracterização do *corpus* de treinamento e com a caracterização do *corpus* de teste

(ii) discretização dos dados relativos à caracterização do *corpus* de treinamento e de teste, a fim de que se obtivessem, de maneira uniforme, intervalos entre 0 e 1 para todos os atributos;

(iii) divisão do único arquivo em dois: um contendo os dados já discretizados do *corpus* de treinamento e outro contento os dados discretizados do *corpus* de teste;

(iv) *upload* do arquivo com o *corpus* de treinamento (ou seja, arquivo com a caracterização das sentenças dos textos-fonte do CSTNews), a partir do qual as estratégias são geradas;

(v) seleção do algoritmo JRip;

(vi) seleção da função *supplied test set*, por meio da qual as estratégias aprendidas no *corpus* de treinamento são testadas no *corpus* de teste;

(vii) *upload* do arquivo único com o *corpus* de teste, ou seja, arquivo com a caracterização das sentenças dos textos-fonte do *corpus* teste;

(viii) geração das regras a partir do *corpus* de treinamento e aplicação das mesmas no *corpus* de teste.

Antes de se apresentar os resultados da avaliação, ressalta-se que apenas 1 dos 3 sumários de cada coleção do *corpus* de teste foi utilizado para a avaliação. Essa restrição foi necessária para evitar dados repetidos. No caso, escolheu-se 1 dos 3 sumários em cada coleção de forma

aleatória, já que todos se mostraram suficientemente informativos. A partir da escolha do sumário em uma coleção, realizou-se a caracterização das sentenças dos textos-fonte que foram alinhadas (e não-alinhadas) ao sumário escolhido. Esse procedimento foi feito em cada uma das 10 coleções. Assim, as sentenças dos textos-fonte caracterizadas em função do alinhamento ao sumário escolhido em cada coleção foram reunidas em um único arquivo, o qual é referido no item (vii).

Ao testar as regras aprendidas a partir do CSTNews no *corpus* de teste, o algoritmo JRip obteve uma precisão de 70%. Esse resultado demonstra que as regras aprendidas codificam estratégias realmente recorrentes de seleção de conteúdo. Diz-se isso porque as regras, quando aplicadas a outro *corpus* em português (o de teste), obtiveram precisão geral muito próxima à obtida quando testadas no próprio *corpus* de treinamento (CSTNews), 72%.

Ademais, se se comparar a quantidade de erros para SIM e para NÃO em ambos os testes, verifica-se que as porcentagens também são semelhantes. Tais observações podem ser vistas no Quadro 22.

Quadro 22 - Comparação das taxas de erro e acerto.

Taxas	**Treino e teste no CSTNews**	**Treino no CSTNews e teste no novo *corpus***
Acerto	72%	70%
Erro no SIM	39%	35,4%
Erro no NÃO	18,9%	24,5%

Fonte: Elaborado pelo autor.

Aliás, a quantidade de erros no SIM é superior à quantidade de erros no NÃO em ambos os testes, pois há mais sentenças dos textos-fonte que não foram alinhadas em detrimento das sentenças que foram alinhadas aos sumários. Assim, a precisão do NÃO é maior, já que havia maior quantidade de dados para aprendizagem dessa natureza, isto é, as regras aprendidas se basearam em poucos dados que levavam ao SIM e muitos que resultavam no NÃO.

Vale ressaltar que a taxa de precisão de 72% obtida pelo treino e teste no CSTNews difere da taxa de precisão para as regras obtidas na fase de aprendizagem (71,25%), pois ao se discretizar os dados para este experimento, o algoritmo JRip gerou intervalos diferentes para os atributos.

A seguir, apresenta-se a avaliação da qualidade dos sumários extrativos gerados pela aplicação manual das estratégias de SHM.

6.3 Avaliação da qualidade de sumários automáticos

Posteriormente à avaliação das estratégias de SHM por meio da verificação da ocorrência das mesmas em um *corpus* de teste, realizou-se a avaliação intrínseca de comparação da qualidade de sumários automáticos gerados por métodos diferentes. Para tanto, foi necessária a criação dos extratos baseados em SHM. Esse processo é apresentado na sequência.

6.3.1 Geração de extratos segundo as estratégias aprendidas

Para a avaliação da qualidade, consideraram-se apenas 6 coleções (C1, C2, C3, C4, C5 e C6) do total de 10 que compõem o *corpus* de teste. Para essa avaliação, geraram-se sumários extrativos com base na aplicação manual das regras aprendidas pelo algoritmo de AM a partir da caracterização do CSTNews e teste das mesmas no *corpus* de teste. Tais regras são apresentadas no Quadro 23. Destaca-se que essas regras diferem do conjunto obtido por meio do treinamento e teste exclusivamente no CSTNews (cf. Quadro 19) devido à discretização distinta que o algoritmo JRip realizou.

Quadro 23 - Regras geradas pelo AM após experimento de avaliação

Regras	**Acertos/Erros**
1. Se Localização = **começo**, então Sumário=**sim**	(140.0/16.0)
2. Senão se Redundância = 0.9-1, então Sumário=**sim**	(81.0/11.0)
3. Senão se Redundância = 0.6-0.7, então Sumário=**sim**	(68.0/12.0)
4. Senão se Redundância = 0.4-0.5, então Sumário=**sim**	(197.0/88.0)
5. Senão se Redundância = 0.3-0.4, então Sumário=**sim**	(172.0/76.0)
6. Senão se Redundância = 0.7-0.8, então Sumário=**sim**	(46.0/7.0)
7. Senão se Redundância = 0.2-0.3e Frequência = 0.5-0.6, então Sumário=**sim**	(35.0/9.0)
8. Senão se Frequência = 0.4-0.5 e Complemento = 0.9-1, então Sumário=**sim**	(13.0/4.0)
9. Senão se Redundância = 0.1-0.2 e Tamanho = 0.2-0.3, então Sumário=**sim**	(11.0/2.0)
10. Senão Sumário=**não**	(1317.0/357.0)

Fonte: Elaborado pelo autor.

Devido à complexidade de implementação das regras em um sistema de sumarização multidocumento, optou-se por gerar manualmente extratos a partir das regras do Quadro 23. A

opção por gerar extratos (e não *abstracts*) foi motivada pelo fato de que os sumários de comparação eram extrativos.

Para tanto, realizou-se a sumarização com base nos processos ou etapas previstos na literatura para a sumarização multidocumento. Especificamente, a seleção do conteúdo a compor os sumários extrativos com base nas estratégias de SHM aqui identificadas foi feita por meio do ranqueamento das sentenças dos textos-fonte e remoção da redundância. Para a síntese ou produção dos sumários, as sentenças selecionadas foram justapostas segundo a ordem de ocorrência nos textos-fonte. Ademais, considerou-se a taxa de compressão de 70%.

Para a seleção de conteúdo, partiu-se de um ranque inicial composto somente pelas sentenças categorizadas por SIM pelo AM, ou seja, pelas sentenças que o AM, com base nas regras de SHM, pré-selecionou para compor o sumário. Esse ranque inicial foi refinado em função da precisão das regras por meio das quais cada uma das sentenças do ranque inicial foi categorizada por SIM. Consequentemente, o ranque refinado passou a ter no topo as sentenças categorizadas por SIM pelas regras mais precisas, ou seja, as que apresentavam menor quantidade de erros.

Quando houvesse mais de uma sentença no ranque caracterizada por SIM pela mesma regra (p. ex.: S2_D1_RX e S3_D2_RX), considerava-se a ordem de ocorrência das sentenças nos textos-fonte para selecionar a primeira entre elas a compor o sumário (p. ex.: S2_D1_RX > S3_D2_RX). Caso as sentenças ocorressem na mesma ordem em seus respectivos textos-fonte, considerava-se a ordem de preferência das fontes de divulgação dos textos descrita em 5.1.3 (isto é, Folha>Estadão>O Globo>Gazeta>JB) para selecionar a primeira entre elas a compor o sumário.

Na sequência, verificava-se se havia redundância entre a próxima sentença do ranque e a já selecionada para compor o sumário. Para tanto, verificou-se se havia alguma relação CST de redundância entre as sentenças. Em caso positivo:

a. Se *Overlap*, *Equivalence* ou *Summary*: selecionava-se a menor sentença entre elas;
b. Se *Identity*, selecionava-se qualquer uma das duas sentenças, já que ambas são idênticas;
c. Se *Subsumption*, selecionava-se a sentença que englobava o conteúdo da outra.

Realizava-se a seleção de conteúdo conforme os passos descritos até que o tamanho mais próximo a 30% do maior texto-fonte fosse atingido. Houve um único caso em que a eliminação da redundância excluiu grande quantidade de sentenças categorizadas por SIM a ponto de a seleção não atingir a taxa de compressão, gerando um sumário menor que o tamanho desejado.

Tomando-se como exemplo a coleção C3 do *corpus* de teste, tem-se que o AM pré-selecionou o total de 12 sentenças, as quais foram categorizadas por SIM em função de 4 regras distintas. No Quadro 24, observa-se que as sentenças foram pré-selecionadas especificamente pela aplicação das regras R1, R2, R3 e R5. Para ilustração, ressalta-se que, no Quadro 24, as fontes Estadão, Folha de São Paulo e O Globo são referidas respectivamente por D1, D2 e D3.

Quadro 24 - Sentenças pré-selecionadas pelo AM

Regras	**Sentenças**
R1	S1_D1; S1_D2; S1_D3
R2	S3_D1; S6_D1
R3	S3_D2; S9_D2; S3_D3
R5	S5_D2; S2_D3; S4_D3; S5_D3

Fonte: Elaborado pelo autor.

Com base na precisão das regras, as sentenças do ranque inicial foram reorganizadas. No ranque refinado, 3 sentenças foram categorizadas por SIM com base na regra R1, a mais precisa delas. Tendo em vista que essas 3 sentenças ocorreram na mesma posição em seus respectivos textos-fonte (S1), selecionou-se inicialmente aquela proveniente da fonte Folha de São Paulo (S1_D2). Na sequência, selecionou-se a sentença proveniente do texto publicado pela fonte Estadão, ou seja, S1_D1. Entre elas, verificou-se 1 relação de *Overlap*, resultando na exclusão de S1_D2 e seleção da S1_D1, menor sentença, para compor o sumário. Na sequência, verificou-se se havia redundância entre a próxima sentença do ranque categorizada por SIM pela mesma regra R1 (S1_D3) e a já selecionada para compor o sumário (S1_D1). Como entre elas havia 1 relação de *Overlap*, a sentença S1_D1, que é a de menor tamanho, foi selecionada para compor o sumário. Dentre todas as sentenças categorizadas por SIM pela R1, apenas a sentença S1_D1 foi selecionada, ao final, para o sumário.

Quando as sentenças pré-selecionadas pela R1 se esgotaram, iniciou-se o mesmo processo para as sentenças pré-selecionadas pelas próximas regras (R2, R3 e R5) até que a taxa de compressão fosse atingida.

Tendo em vista a taxa de compressão de 70%, o tamanho desejado para o sumário multidocumento para a coleção C3 era de 101 palavras. Porém, o número total de palavras do extrato foi de 116 palavras, valor mais próximo dos 30% desejados. Ao final, as sentenças efetivamente selecionadas foram justapostas com base na ordem de ocorrência nos textos-

fonte. No Quadro 25, apresenta-se o sumário extrativo resultante da aplicação manual das estratégias de SHM identificadas neste trabalho.

Quadro 25 - Sumário extrativo considerando estratégias de SHM

A queda de uma estrutura que estava sendo montada para um evento no balneário baiano da Costa do Sauípe deixou ao menos 40 operários feridos nesta quarta-feira, informou a Polícia Militar da Bahia. Um operário que chegou a ficar preso na estrutura foi resgatado, mas seu estado de saúde é grave. Os feridos foram encaminhados para três hospitais da rede da Secretaria da Saúde do Estado: Hospital Geral de Camaçari e Hospital Menandro de Faria, ambos na região metropolitana de Salvador e o Hospital Geral do Estado, na capital. O Bradesco, patrocinador do evento onde a tenda montada pelos operários desabou, lamentou o ocorrido e disse que está tomando "todas as providências" para atender as vítimas.

Fonte: Elaborado pelo autor.

Na sequência, apresenta-se a avaliação dos extratos gerados aqui quanto à sua qualidade, comparando-os com sumários produzidos por outro método.

6.3.2 Avaliação da qualidade dos extratos

Como mencionado, a informatividade de um sumário é comumente avaliada pela medida ROUGE. Há outras propriedades textuais, no entanto, que a ROUGE não é capaz de julgar, as quais influenciam a qualidade dos sumários.

De acordo com a DUC, a qualidade de um sumário pode ser avaliada em função das seguintes propriedades: (i) gramaticalidade, que diz respeito à ausência de erros de ortografia, pontuação e sintaxe, (ii) não redundância, que se refere à ausência de informações repetidas, (iii) clareza referencial, que diz respeito à clara identificação dos componentes da superfície textual que fazem remissão a outro(s) elemento(s) do universo textual, (iv) foco, que se refere ao fato de que as informações de uma sentença devem se relacionar com as informações do restante do sumário, e (v) estrutura e coerência, que diz respeito à organização do sumário considerando sua textualidade.

Com o intuito de avaliar a qualidade, os 6 sumários utilizados na avaliação anterior, gerados pela aplicação manual das estratégias/regras de SHM identificadas pelo AM, foram avaliados manualmente quanto às propriedades especificadas pela DUC.

A título de comparação, os sumários gerados por outro método de SAM para as mesmas 6 coleções também foram avaliados em função das mesmas propriedades textuais. No caso, o método automático utilizado para gerar os sumários de comparação foi o de Ribaldo (2012), que se baseia em grafos e redes complexas, como mencionado em 6.3.2.

A avaliação das propriedades relativas à qualidade foi realizada por 10 linguistas computacionais. Para cada um dos 6 sumários selecionados para a avaliação, os juízes pontuaram cada uma das 5 propriedades textuais por meio de um formulário *online*. Para todas as propriedades, os juízes dispunham de uma escala de 1 a 5 pontos, a qual está descrita no Quadro 26.

Quadro 26 – Pontuações e níveis estipulados para a avaliação da qualidade

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

Os resultados da avaliação manual da gramaticalidade, não redundância, clareza referencial, foco e estrutura/coerência são apresentados nas Tabelas 9, 10, 11, 12 e 13, respectivamente. Em cada tabela, as estratégias de SHM identificadas neste trabalho são referenciadas por "método 1" e o método de Ribaldo (2012), por "método 2".

O valor de cada célula das tabelas indica a quantidade de avaliações que cada nível recebeu (cf. Quadro 26) conforme a propriedade em questão, para os dois métodos. Para tanto, apresentam-se os valores de duas formas: (i) absoluto, e (ii) porcentagem. Para cada propriedade, calculou-se a média ponderada das avaliações, isto é, o nível "péssimo" recebeu peso 1, o nível "ruim" recebeu peso 2, o nível "regular" recebeu peso 3, o nível "bom" recebeu peso 4 e o nível "excelente recebeu peso 5. Portanto, quanto mais próxima de 5 a média for, melhor o resultado e quanto mais próximo de 1, pior.

Na Tabela 9, por exemplo, observa-se que a gramaticalidade dos 6 sumários gerados pelo método 1: (i) não recebeu as pontuações 1 ("péssimo") e 2 ("ruim"), (ii) recebeu 3 vezes a pontuação 3 ("regular"), isto é, 5% do total; (iii) recebeu 18 vezes a pontuação 4 ("bom"), ou seja, 30% do total, (iv) recebeu 39 vezes a pontuação 5 ("excelente"), equivalente a 65% do total. Com isso, a gramaticalidade teve média ponderada de 4,7, revelando que os juízes a consideram de nível "excelente" na média, já que a pontuação 4,7 está mais próxima de 5.

Ainda quanto à Tabela 9, observa-se que a gramaticalidade dos sumários gerados pelo método 2: (i) não recebeu as pontuações 1 ("péssimo") e 2 ("ruim"), (ii) recebeu 7 vezes a pontuação 3 ("regular"), ou seja, 11,6% do total, (iii) recebeu 22 vezes a pontuação 4 ("bom"), isto é, 36,6% do total, e (iv) recebeu 31 vezes a pontuação 5 ("excelente"), o que

equivale a 51,6%. Assim, essa propriedade textual teve média ponderada de 4,4, indicando que os juízes a consideraram de nível "bom" (mais próximo de 4).

Tabela 9 – Avaliação manual da "gramaticalidade"

	Gramaticalidade										
	Péssimo (1)		Ruim (2)		Regular (3)		Bom (4)		Excelente (5)		**Média**
Método 1	0	**0%**	0	**0%**	3	**5%**	18	**30%**	39	**65%**	4,7 (**excelente**)
Método 2	0	**0%**	0	**0%**	7	**11,6%**	22	**36,6%**	31	**51,6%**	4,4 (**bom**)

Fonte: Elaborado pelo autor.

Na média, a gramaticalidade dos sumários gerados pelo método 1 foi identificada como "excelente", posto que ela recebeu a pontuação média de 4,7. A gramaticalidade dos sumários gerados pelo método 2 foi definida em média como "bom", pois essa propriedade recebeu a pontuação média de 4,4. Tal fato significa que o método 1, ou seja, a que consistiu na aplicação manual das regras de SHM, gera sumários extrativos com poucos problemas de ortografia, pontuação e sintaxe. Entretanto, a gramaticalidade poderia ser desconsiderada na avaliação, pois os problemas presentes nos extratos são integralmente advindos dos textos-fonte, já que nenhum dos métodos gera abstratos.

Tabela 10 – Avaliação manual da "não redundância"

	Não redundância										
	Péssimo (1)		Ruim (2)		Regular (3)		Bom (4)		Excelente (5)		**Média**
Método 1	0	**0%**	0	**0%**	2	**3,3%**	15	**25%**	43	**71,6%**	4,7 (**excelente**)
Método 2	0	**0%**	2	**3,3%**	17	**28,3%**	17	**28,3%**	24	**40%**	4,1 (**bom**)

Fonte: Elaborado pelo autor.

O mesmo foi observado para a propriedade "não redundância". No caso, a "não redundância" dos sumários gerados pelo método 1 recebeu a pontuação média de 4,7, ou seja, "excelente". Os sumários gerados pelo método 2, por sua vez, receberam a pontuação média de 4,1 ("bom"). Essa diferença pode ser justificada pelo fato de que o método 1 engloba um processo de eliminação da redundância baseado na CST. O método 2, por sua vez, também incorpora um processo de remoção de redundância, porém, ele se baseia em uma estratégia superficial que compara a similaridade dos itens lexicais entre duas sentenças. Castro Jorge (2010), aliás,

comprova que métodos que se baseiam nas relações CST para tratar a redundância dos sumários melhoram significativamente os resultados que se referem a essa propriedade.

Tabela 11 – Avaliação manual da "clareza referencial"

	Clareza Referencial										
	Péssimo (1)		Ruim (2)		Regular (3)		Bom (4)		Excelente (5)		**Média**
Método 1	0	**0%**	0	**0%**	9	**15%**	20	**33,3%**	31	**51,6%**	4,4 (**bom**)
Método 2	0	**0%**	2	**3,3%**	5	**8,3%**	26	**43,3%**	27	**45%**	4,3 (**bom**)

Fonte: Elaborado pelo autor.

Quanto à propriedade "clareza referencial", os sumários gerados pelo método 1 e pelo método 2 receberam pontuações médias bastante semelhantes, 4,4 e 4,3, respectivamente. Essas pontuações indicam que os sumários apresentam "bom" nível de "clareza referencial". No caso, essas pontuações justificam-se pelo fato de que nenhum dos métodos realiza qualquer processo de resolução de correferência.

Tabela 12 – Avaliação manual do "foco"

	Foco										
	Péssimo (1)		Ruim (2)		Regular (3)		Bom (4)		Excelente (5)		**Média**
Método 1	0	**0%**	0	**0%**	3	**5%**	24	**40%**	33	**55%**	4,5 (**excelente**)
Método 2	1	**1,6%**	4	**6,6%**	11	**18,3%**	22	**36,6%**	22	**36,6%**	4 (**bom**)

Fonte: Elaborado pelo autor.

Com base na Tabela 12, observa-se que o "foco" dos sumários gerados pelo método 1 recebeu a pontuação média de 4,5, sendo, portanto, considerado de nível "excelente" por apresentar sentenças altamente relacionadas. O "foco" dos sumários gerados pelo método 2, por sua vez, recebeu a pontuação média de 4, sendo considerado, por conseguinte, de nível "bom"; nesse caso, os sumários gerados pelo método 2 apresentam maior quantidade de sentenças com conteúdo disperso ou pouco relacionado.

Tabela 13 – Avaliação manual da “estrutura e coerência”

	Estrutura e Coerência										
	Péssimo (1)		Ruim (2)		Regular (3)		Bom (4)		Excelente (5)		**Média**
Método 1	0	**0%**	0	**0%**	7	**11,6%**	33	**55%**	20	**33,3%**	4,2 (**bom**)
Método 2	0	**0%**	6	**10%**	19	**31,6%**	23	**38,3%**	12	**20%**	3,7 (**bom**)

Fonte: Elaborado pelo autor.

Com base na Tabela 13, observa-se que a propriedade “estrutura e coerência” dos sumários gerados pelos métodos 1 e 2 receberam as pontuações médias de 4,2 e 3,7, respectivamente. Nesse caso, a estrutura e a coerência foram consideradas de nível “bom” nos sumários gerados pelos 2 métodos. Por se tratar de sumários extrativos, isto é, compostos por sentenças extraídas na íntegra dos textos-fonte, a ausência de reescrita pode ter prejudicado a estrutura dos sumários, os quais, por isso, apresentam sentenças desconexas umas com as outras.

No geral, os sumários gerados pela aplicação manual das estratégias de SHM apresentam melhor qualidade que os sumários gerados pelo método de Ribaldo (2012).

7 CONSIDERAÇÕES FINAIS

Diante dos experimentos realizados, discutem-se neste capítulo algumas contribuições e limitações deste trabalho de mestrado. Ademais, propõem-se alguns trabalhos futuros relacionados a esta pesquisa.

7.1 Contribuições

Neste trabalho, realizou-se a primeira pesquisa sistemática sobre a SHM. Dessa pesquisa, destacam-se algumas contribuições à comunidade do PLN.

Uma dessas contribuições diz respeito ao alinhamento das sentenças dos sumários manuais às sentenças dos textos-fonte do *corpus* de referência CSTNews. Esse alinhamento consistiu em uma anotação incorporada ao CSTNews, enriquecendo-o para pesquisas linguístico-computacionais.

Ademais, destaca-se a tipificação dos alinhamentos que, embora não tenha sido efetivamente utilizada neste trabalho, poderá ser utilizada em trabalhos futuros. Ambas as anotações estarão em breve disponíveis nas páginas dos projetos SUSTENO e SUCINTO.

Além das contribuições relativas ao alinhamento e sua tipificação, destaca-se o levantamento de características de um objeto textual até então não explorado em função de atributos linguísticos de diferentes níveis: (i) superficial, (ii) profundo, e (iii) extralinguístico. Fazendo uso de tais atributos, este trabalho gerou uma vasta caracterização de cada sentença dos textos-fonte do *corpus* CSTNews, caracterizando indiretamente as sentenças dos sumários humanos que foram alinhadas. Essa caracterização, aliás, gerou conhecimento específico para a comunidade linguística.

Para a comunidade linguística em específico,

A identificação de estratégias de SHM é, com certeza, a principal contribuição deste trabalho. Tais estratégias traduzem-se em regras formais aprendidas por meio de um algoritmo de AM, que obtiveram 71,25% de precisão. Quando avaliadas em um *corpus* distinto, composto por 10 coleções de textos-fonte, elas obtiveram 70% de precisão, revelando sua pertinência. A qualidade dos sumários gerados pela aplicação manual das regras de SHM foi superior à qualidade de sumários gerados com base em outras estratégias de sumarização multidocumento.

A avaliação das estratégias, aliás, gerou um *corpus* de teste composto por 10 coleções de textos jornalísticos. Construído segundo as diretrizes usadas na construção do CSTNews,

esse *corpus* de teste pode ser visto como outra contribuição deste trabalho, já que consiste em outro recurso linguístico disponível para os pesquisadores do PLN.

7.2 Limitações

Apesar das contribuições oferecidas por este trabalho de mestrado à comunidade do PLN, identificaram-se algumas limitações, as quais interferiram no andamento desta pesquisa e nos resultados obtidos.

Uma dessas limitações incidiu no fato de as regras geradas pelo AM, maior contribuição deste trabalho, terem sido testadas em um conjunto relativamente pequeno de textos. Com relação à verificação da ocorrência das estratégias aprendidas no *corpus* de treinamento em outro *corpus*, ressalta-se que tais estratégias foram testadas em um *corpus* com apenas 10 coleções de textos multidocumento, dada a complexidade e subjetividade da tarefa de criação e anotação de *corpus* e a inexistência de um *corpus* que satisfizesse os requisitos básicos para o teste.

Quanto à avaliação da qualidade dos extratos, ressalta-se a necessidade de criação de extratos para a posterior comparação com a qualidade de extratos gerados por outros métodos. Para tanto, utilizaram-se apenas 6 coleções do *corpus* de teste, devido à ausência de um sistema que simulasse o esforço humano.

Outra limitação decorre da dificuldade de implementação das estratégias em um método de SAM. Essa tarefa implica em caracterizar automaticamente as sentenças dos textos-fonte em função de um conjunto extenso e complexo de atributos linguísticos, o que pode dificultar o processo, já que a identificação automática de alguns desses atributos pode gerar muitos erros e lentidão. Ademais, a identificação das relações CST nos textos-fonte dependeria de um *parser* discursivo, considerando que o *parser* que se tem atualmente para o português (MAZIERO; PARDO, 2011) apresenta baixa precisão para alguns tipos de relações.

7.3 Trabalhos futuros

Tendo em vista a exploração mais profunda da SHM, sugerem-se os seguintes trabalhos futuros:

- Avaliar os sumários extrativos gerados com base na aplicação manual das regras aprendidas pelo algoritmo de AM quanto à sua informatividade, isto é, identificar o quanto de informação relevante dos textos-fonte o sumário automático incorpora.
- Refinar a avaliação, gerando sumários extrativos para as outras 4 coleções do *corpus* de teste. Por meio de tal refinamento, torna-se mais viável avaliar os extratos quanto à informatividade, já que as medidas contidas no pacote ROUGE se baseiam na generalização de grande quantidade de textos sob processamento.
- Integrar as estratégias de SHM identificadas neste trabalho em um método de SAM, isto é, implementar as estratégias em um sistema de SAM para o português.
- Investigar os fenômenos textuais multidocumento, visando eliminar principalmente os problemas relacionados à contradição, redundância e complementariedade. Tal investigação poderá melhorar a qualidade e informatividade dos sumários.
- Caracterizar os sumários humanos multidocumento quanto a outros atributos linguísticos que possam ser relevantes para a seleção de conteúdo na SHM. Por exemplo, uma possibilidade é verificar se os sumários apresentam uma quantidade média de substantivos, por exemplo, de tal forma que essa quantidade possa ser um requisito para a seleção de conteúdo na SAM.

REFERÊNCIAS BIBLIOGRÁFICAS

AFANTENOS, S.D.; DOURA, I.; KAPELLOU, E.; KARKALETSIS, V. Exploiting Cross-Document Relations for Multi-document Evolving Summarization. In: VOUROS, G. A., PANAYIOTOPOULOS, T. (Eds.). Methods and applications of Artificial Intelligence/ Helenic Conference on AI, 3, 2004, Samos, Greece. **Proeedings…** Samos, 2004. p. 410-419.

_____; KARKALETSIS, V.; STAMATOPOULOS, P.; HALATSIS, C. Using synchronic and diachronic relations for summarizing multiple documents describing evolving events. **Journal of Intelligent Information Systems**, Vol. 30, N. 3, pp. 183-226, 2008.

AKABANE, A. T.; 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.

ALEIXO, P.; PARDO, T.A.S. CSTNews: um córpus de textos jornalísticos anotados segundo a Teoria Discursiva Multidocumento CST (Cross-document Structure Theory). **Technical Report**, Universidade de São Paulo, n. 326. São Carlos-SP, 12p, 2008.

ALUÍSIO, S.M.; ALMEIDA, G.M.B. **O que é e como se constrói um corpus? Lições aprendidas na compilação de vários corpora para pesquisa lingüística**. Calidoscópio (UNISINOS). Vol. 4, n. 3. p. 155-177, set/dez 2006.

BARBOSA, J.P. **Trabalhando com os gêneros do discurso**: relatar: notícia. São Paulo: FTD, 2001.

BARZILAY, R.; MCKEOWN, K.; ELHADAD, M. Information fusion in the context of multi-document summarization. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS ON COMPUTATIONAL LINGUISTICS, 37, 1999, Maryland. **Proceedings…** Maryland, 1999, p. 550-557.

BARZILAY, R., ELHADAD, N. Sentence Alignment for Monolingual Comparable Corpora. In: CONFERENCE ON EMPIRICAL METHODS FOR NATURAL LANGUAGE. **Proceedings...** Sapporo, Japan, 2003, p. 25-32.

BAXENDALE, P. Machine-made index for technical literature - an experiment. **IBM Journal of Research Development**, Vol. 2, N. 4, pp. 354-361, 1958.

BERBER SARDINHA, T. **Linguística de Corpus**. Manole. Barueri, SP. 2004.

BERBER SARDINHA, T. O banco de palavras-chave como instrumento de identificação de palavras-chave exclusivas no programa *WordSmithTools*. **The Especilist**, São Paulo, v.27, n.1, p. 1-19, out. 2006.

CARBONELL, J.; GOLDSTEIN, J. The Use of MMR, Diversity-Based Reranking For Reordering Documents And Producing Summaries. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 21., 1998, Melbourne. **Proceedings…** Melbourne, 1998. p. 335-336.

CARDOSO, P.C.F.; MAZIERO, E.G.; CASTRO JORGE, M.L.R.; SENO, E.M.R.; DI-FELIPPO, A.; RINO, L.H.M.; NUNES, M.G.V.; PARDO, T.A.S. A CSTNews - A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese. In: RST BRAZILIAN MEETING, 3., 2011a, Cuiabá. **Proceedings…** Cuiabá: UFMT, 2011a. p. 88-105.

CARLETTA, J. **Assessing Agreement on Classification Tasks**: The Kappa Statistic. Computational Linguistics, v. 22, n. 2, pp. 249-254, 1996.

CASTRO JORGE, M.L.; PARDO, T.A.S. Experiments with CST-based Multidocument Summarization. In: ACL WORKSHOP TEXTGRAPHS, 5, 2010, Uppsala, Sweden. **Proceedings…** Uppsala, 2010. p. 74-82, 2010.

_____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: UFMT, 2011. p. 75-87.

_____; AGOSTINI, V.; PARDO, T.A.S. Multi-document Summarization Using Complex and Rich Features. In: ENCONTRO NACIONAL DE INTELIGÊNCIA ARTIFICIAL, 8., 2011, Natal. **Anais…** Natal: UFRN, 2011. p. 1-12.

CELIKYILMAZ, A.; HAKKANI-TUR, D. Discovery of Topically Coherent Sentences for Extractive Summarization. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 49., 2011, Portland. **Proceedings…** Portland, 2011. p. 491-499.

CLARKE, J.; LAPATA, M. Discourse Constraints for Document Compression**.** **Computational Linguistics**, v. 36, N. 3, pp. 411-441, 2010.

CLOUGH, P.; GAIZAUSKAS, G.; PIAO, S. S. L.; WILKS, Y. 'METER': MEasuringTExt Reuse. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (ACL), 40., Philadelphia. **Proceedings…** Philadelphia, 2002, p. 152-159.

CREMMINS, E.T. **The art of abstracting**. Arlington, Virginia: Information Resources Press, 1996.

DAUMÉ III, H., MARCU, D. A Phrase-Based HMM Approach to Document/Abstract Alignment. In: EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING (EMNLP). **Proceedings…** Barcelona, Spain, 2004.

_____. Induction of Word and Phrase Alignments for Automatic Document Summarization. Computational Linguistics, v. 31, n. 4, p. 505-530, 2005.

DIAS-DA-SILVA, B. C. **A face tecnológica dos estudos da linguagem: o processamento automático das línguas naturais**. Araraquara, 1996. 272p. Tese (Doutorado em Letras) - Faculdade de Ciências e Letras, Universidade Estadual Paulista, Araraquara, 1996.

_____. O estudo linguístico-computacional da linguagem. **Letras de Hoje**, Porto Alegre, v. 41, n. 2, p. 103-138, 2006.

DOLZ, J.; SCHNEUWLY, B. **Gêneros orais e escritos na escola**. Campinas, SP: Mercado de Letras, 2004. 278 p. (Tradução e organização: Roxane Rojo; Glaís Sales Cordeiro).

DONAWAY, R. L., DRUMMEY, K. W., MATHER, L. A. A comparison of rankings produced by summarization evaluation measures. In: ANLP/NAACL Workshop on Automatic Summarization, 2000, Stroudsburg, PA. **Proceedings…** Stroudsburg, 2000.

EDMUNDSON, H. P. New Methods in automatic extracting. **Journal of the ACM**, Vol. 16, pp. 264-285, 1969.

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

FELLBAUM, C. (editor). **WordNet: An Electronic Lexical Database**. The MIT Press, Cambridge, MA, 1998.

GUPTA, V; LEHAL, G. S. A Survey of Text Summarization Extractive Techniques. **Journal of Emerging Technologies in Web Intelligence**, Oulu, v. 2, n. 3, p. 258-268, 2010.

HAGHIGHI, A.; VANDERWENDE, L. Exploring content models for multi-document summarization. In: HUMAN LANGUAGE TECHNOLOGIES: THE 2009 ANNUAL CONFERENCE OF NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 10., 2009, Boulder, Colorado. **Proceedings...** Boulder: University of Colorado, 2009. p. 362-370.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1, 2009.

HALTEREN, H.; TEUFEL, S. Examining the consensus between human summaries: initial experiments with factoid analysis. In: HLTNAACL DUC Workshop, 2003, Edmonton. **Proceedings…** Edmonton, 2003.

HASLER, L. From extracts to abstracts: human summary production operations for Computer-Aided Summarisation. In: RANLP WORKSHOP ON COMPUTER-AIDED LANGUAGE PROCESSING, 2007, Borovets, Bulgaria. **Proceedings….** Borovets, 2007, p.11-18.

HAYES-ROTH, F. Expert systems. In: SHAPIRO, E. (Ed.). **Encyclopedia of artificial intelligence**. New York, Wiley, 1990, p. 287-298.

HATZIVASSILOGLOU, V., KLAVANS, J. L., ESKIN. E. 'Detecting Text Similarity over Short Passages': Exploring Linguistic Feature Combinations via Machine Learning. In: EMPIRICAL METHODS FOR NATURAL LANGUAGE PROCESSING, 1999. **Proceedings…** Maryland, 1999, p. 203–212.

HIRAO, T., SUZUKI, J., ISOZAKI, H., MAEDA, E. Dependency-based Sentence Alignment for Multiple Document Summarization. In: International conference on Computational Linguistics (COLING), 2004. **Proceedings…** Switzarland, 2004, p. 446-452.

JAIDKA, K.; CHRISTOPHER S. G. KHOO, JIN-CHEON NA. Imitating human literature review writing: an approach to multi-document summarization. In: ICADL, 2011, Beijing, China. **Proceedings…** Beijing, 2010, p. 116–119.

JING, H., BARZILAY, R., MCKEOWN, K., ANDELHADAD, M. Summarization Evaluation Methods: Experiments and Analysis. In: WORKING NOTES OF THE WORKSHOP ON INTELLIGENT TEXT SUMMARIZATION, 1998. **Proceedings…**1998. California: American Association for Artificial Intelligence Spring Symposium Series, 1998. p. 60-68.

JING, H.; MCKEOWN, K.R. The decomposition of human-written summary sentence. In: INTERNATIONAL ACM SIGIR, 22., 1999, New York. **Proceedings…** New York, 1999. p. 129-136.

JOHNSON , R. E. Recall of prose as a function of the structural importance of the linguistic units. **Journal of Verbal Learning and Verbal Behavior**, 9, p.12-20, 1970.

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

KUMAR, Y. J; SALIM, N. Automatic Multi Document Summarization Approches. J. Comput. Sci., 8, 2012, p. 133-140.

KUPIEC, P.; PEDERSEN, J.; CHEN, F. A Trainable Document Summarizer. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 18., 1995, New York, NY. **Proceedings...** New York, 1995, p. 68-73.

LAGE, N. **Estrutura da Notícia**. 5ª ed. São Paulo: Ática, 2002.

_____. **A reportagem**: teoria e técnica de entrevista e pesquisa jornalística. Rio de Janeiro: Record, 2004.

LEECH, G. Introducing corpus annotation. In: R. GARSIDE et al. (org.). Corpus Annotation – Linguistic Information from Computer Text Corpora. London and New York: Longman, 1997a.

LIDDY, E. D. The discourse-level structure of empirical abstracts: an exploratory study. **Information Processing & Management**, 1991, p. 55-81.

LHUN, H. P. The automatic creation of literature abstracts. **IBM Journal of Research,** Riverton, v. 2, n. 2, p. 159-165, 1958.

LIN, C.; HOVY, E.H. From Single to Multi-document Summarization: A Prototype System and its Evaluation. In: ANNIVERSARY MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (ACL-02), 40, 2002, Philadelphia, Pennsylvania. **Proceedings…** Philadelphia, 2002. p. 7-12.

LIN, C.; HOVY, E.H. Automatic Evaluation of Summaries Using N-gram Cooccurrence Statistics. In: LANGUAGE TECHNOLOGY CONFERENCE, 2003, Edmonton, Canada. **Proceedings…** Edmonton, 2003.

LOUIS, A.; JOSHI, A.; NENKOVA, A. Discourse indicators for content selection in summarization. In: ANNUAL MEETING OF THE SPECIAL INTEREST GROUP ON DISCOURSE AND DIALOGUE, 11, 2010, Tokyo. **Proceedings…** Tokyo, 2010. p. 147–156.

LOUIS, A; NENKOVA, A. Automatically Assessing Machine Summary Content Without a Gold Standard. **Computational Linguistics**, Cambridge, MA, USA. v. 39. p. 267-300, 2013.

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

_____; GATES, B.; BLOEDORN, E. Improving summaries by revising them. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 37., 1999, New Brunswick, New Jersey. **Proceedings…** New Brunswick, 1999. p. 558-565.

_____; MAYBURY, M.T. **Advances in automatic text summarization**. Cambridge, MA: The MIT Press, 1999.

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

MARCU, D. From discourse structure to the text. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS/EACL SUMMARIZATION WORKSHOP, 1997. **Proceedings...** 1997a, p. 82-88.

_____. Discourse trees are good indicators of importance in text. In: MANI, I., MAYBURY, M. **Advances in Automatic Text Summarization**, Cambridge, MA: The MIT Press, 1999. p. 123-136.

_____. **The Theory and Practice of Discourse Parsing and Summarization**. The MIT Press. Cambridge, Massachusetts, 2000.

MAZIERO, E. G.; JORGE, M. L. C.; PARDO, T. A. S. Identifying Multidocument Relations. In: INTERNATIONAL WORKSHOP ON NATURAL LANGUAGE PROCESSING AND COGNITIVE SCIENCE, 7., 2010, Funchal, Madeira. **Proceedings...** Funchal, 2010. p. 60-69.

_____; 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: UFMT, 2011. p. 1-10.

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

_____; PASSONNEAU, R.; ELSON, D.; NENKOVA, A.; HIRSCHBERG, J. Do Summaries Help? A Task-Based Evaluation of Multi-Document Summarization. In: ANNUAL INTERNATIONAL ACM-SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 28., 2005, Salvador, Bahia. **Proceedings…** Salvador**,** 2005. p. 210-217.

NENKOVA, A., PASSONNEAU, R. Evaluating content selection in summarization: The pyramid method. In: HLT/NAACL, 2004, Boston. **Proceedings…** Boston, 2004.

NENKOVA, A Discourse Factors in Multi-Document Summarization. In: ANNUAL AAAI/SIGART DOCTORAL CONSORTIUM, 10., 2005a, Pittsburgh. **Proceedings…** Pittsburgh, 2005a. p. 1654-1655.

_____. Automatic text summarization of newswire: Lessons learned from the document understanding conference. In: NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE, 20., 2005b, Pittsburgh. **Proceedings…** Pittsburgh, 2005b.

_____. **Understanding the process of multi-document summarization**: content selection, rewrite and evaluation. PhD Thesis, Columbia University, January 2006.

O'DONNELL, M. Variable-Length On-Line Document Generation. In: EUROPEAN WORKSHOP ON NATURAL LANGUAGE GENERATION, 6., 1997a, Duisburg. **Proceedings…** Duisburg, 1997a.

OKUMURA, M., FUKUSIMA, T., NANBA, H. 'Text Summarization Challenge 2' - Text Summarization Evaluation at NTCIR Workshop 3. HLT-NAACL 2003 WORKSHOP: TEXT SUMMARIZATION (DUC03), 2003, p. 49-56.

OTTERBACHER, J.; RADEV, D.R.; LUO, A. Revisions that improve cohesion in multi-document summaries: a preliminary study. In: WORKSHOP ON AUTOMATIC SUMMARIZATION, 2002, Philadelphia. **Proceedings…** Pennsylvania, 2002, p. 27-36.

PARDO, T. A. S; RINO, L.H.M. DMSumm: Review and Assessment. In: RANCHHOD, E.; MAMEDE, N. J. (Eds.), **Advances in Natural Language Processing**. London: Springer-Verlag/Germany, 2002. p. 263-273.

PARDO, T.A.S.; RINO, L.H.M.; NUNES, M.G.V. GistSumm: A Summarization Tool Based on a New Extractive Method. In N.J. MAMEDE, J. BAPTISTA, I. TRANCOSO, M.G.V. NUNES (EDS.), Workshop on Computational Processing of the Portuguese Language - Written and Spoken – PROPOR (Lecture Notes in Artificial Intelligence 2721), 6., 2003, Faro. **Proceedings...** Faro, 2003, p. 210-218.

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

PITLER, E.; LOUIS, A.; NENKOVA, A. Automatic evaluation of linguistic quality in Multi-Document Summarization. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 48., 2010, Uppsala, Sweden. **Proceedings…** Uppsala, 2010, p. 544–554.

RADEV, D.; MCKEOWN, K. R. Generating natural language summaries from multiple on-line sources. **Journal Computational Linguistics - Special issue on natural language generation**, Cambridge, v. 24, p. 470–500, 1998.

RADEV, D. **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. 74-86.

_____; JING, H.; BUDZIKOWSKA, M. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation and user studies. In: ANLP/NAACL WORKSHOP, 2000. Seattle, Washington. **Proceedings…** Seattle, 2000. p. 21-29.

_____; OTTERBACHER, J.;ZHANG, Z. CSTBank: a corpus for the study of cross-document structural relationships. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 4, 2004, Lisbon. **Proceedings…** Lisbon, 2004.

RATH, G. J.; RESNICK, A. SAVAGE, R. The formation of abstracts by the selection of sentences: Part 1: sentence selection by man and machines. **American Documentation**, v. 2, n 12, p. 139-208, 1961.

RATNAPARKHI, A. A maximum entropy part-of-speech tagger. In: EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING CONFERENCE, 1, 1996. **Proceedings**… Philadelphia, 1996. p.133-142.

RIBALDO, R.; PARDO, T.A.S.; RINO, L.H.M. Sumarização Automática Multidocumento com Mapas de Relacionamento. In: BRAZILIAN SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY, 8., 2011, Cuiabá. **Proceedings…** Cuiabá: UFMT, 2011. p. 1-3.

RIBALDO, R.; AKABANE, A.T.; RINO, L.H.M.; PARDO, T.A.S. Graph-based Methods for Multi-document Summarization: Exploring Relationship Maps, Complex Networks and Discourse Information. In: 10TH INTERNATIONAL CONFERENCE ON COMPUTATIONAL PROCESSING OF PORTUGUESE (LNAI 7243), 10, 2012, Coimbra. **Proceedings…** Coimbra, 2012. , p. 260-271.

RINO, L.H.M.; PARDO, T.A.S.; SILLA Jr., C.N.; KAESTNER, C.A., POMBO, M. A Comparison of Automatic Summarization Systems for Brazilian Portuguese Texts. In: BRAZILIAN SYMPOSIUM ON ARTIFICIAL INTELLIGENCE – SBIA (LECTURE NOTES IN ARTIFICIAL INTELLIGENCE 3171), 17., 2004, São Luís. **Proceedings…** São Luís, 2004, p. 235-244.

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

SAGGION, H.; RADEV, D.; TEUFEL, S.; LAM, W. Meta-evaluation of summaries in a crosslingual environment using content-based metrics. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS (COLING'02), 2002, Taipei, Taiwan. **Proceedings…** Taipei, 2002.

SALTON, G.; SINGHAL, A.; MITRA, M.; BUCKLEY C. Automatic text structuring and summarization. **Information Processing & Management**, v. 33, n. 2, p. 193-207, 1997.

SCHIFFMAN, B.; NENKOVA, A.; MCKEOWN, K. Experiments in multidocument summarization. In: HUMAN LANGUAGE TECHNOLOGY CONFERENCE, 2, 2002. San Diego. **Proceedings…** San Diego, 2002. p. 52-58.

SHILDER, F.; KONDADADI, R. FastSum: fast and accurate query-based multi-document summarization. In: Association for Computational Linguistics, 2008. Ohio. **Proceedings…** Ohio, 2008, p.205-208.

SPARCK JONES, K. Discourse modeling for Automatic Summarisation. **Tech. Report No. 290**. University of Cambridge. UK, February, 1993.

_____, GALLIERS, J. R. **Evaluating Natural Language Processing Systems:** An Analysis and Review. Cambridge: Springer, 1996.

_____; WILLET, P. **Readings in information retrieval**. São Francisco: Morgan Kaufmann, 1997.

_____. Automatic Summarizing: factors and directions. In: MANI, I; MAYBURY, M. (Ed.). **Advances in automatic text summarization**. The MIT Press, 1999, p. 1-12.

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.

TEIXEIRA, H. M. L. **O clipping de mídia impressa numa abordagem interdisciplinar sob os prismas da ciência da informação e da comunicação social; o jornal de recortes da Assembleia Legislativa de Minas Gerais**. Belo Horizonte, 2001, 3–5 p. Dissertação (Mestrado) — Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte – MG, 2001.

TEUFEL, S.; Moens, M. Summarizing scientific articles: experiments with relevance and rhetorical status. **Computational Linguistics**, v. 28, n. 4, p. 409–445, 1997.

UZÊDA, V.R.; PARDO, T.A.S.; NUNES, M.G.V. A comprehensive comparative evaluation of RST-based summarization methods. **ACM Transactions on Speech and Language Processing**, n. 6, v. 4, 2010. p. 1-20.

VOUTILAINEN, A. Part-of-speech tagging. In: MITKOV, R. (Ed.). **The Oxford handbook of computational linguistics**. Oxford, New York: Oxford University Express, 2004, cap. 11, p. 219-232.

WAN, X.; YANG, J. Improved affinity graph based multi-document summarization. In: HUMAN LANGUAGE TECHNOLOGY CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ACL, 7, 2006, New York. **Proceedings…** New York, 2006. p. 181-184.

WAN, X. An Exploration of Document Impact on Graph-Based Multi-Document Summarization. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 2008, Honolulu, Hawaii. **Proceedings…** Honolulu, 2008. p. 755-762.

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

ZHANG, Z.; BLAIR-GOLDENSOHN, S.; RADEV, D.R. Towards CST-enhanced summarization. In: AAAI CONFERENCE, 2002. Edmonton, Alberta. **Proceedings…** Edmonton, 2002. p. 439-445.

_____; OTTERBACHER, J.; RADEV, D.R. Learning Cross-document Structural Relationships using Boosting. In: ACM INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, 12, 2003, New Orleans, Louisiana. **Proceedings…** New Orleans, 2003. p. 124-130.

APÊNDICE A – Sumários extrativos baseados nas estratégias de SHM

Coleção 1

Cientistas anunciaram pela primeira vez a descoberta de bactérias vivendo sob o gelo antártico, no escuro e sob baixíssimas temperaturas. O achado pode dar pistas de como a vida pode prosperar em outros planetas e luas. Mas John C. Priscu, da Universidade do Estado de Montana, nos EUA, líder da expedição científica, declarou ao jornal “The New York Times” que a descoberta “transforma a maneira de se ver o continente”. Depois de perfurar cerca de 800 metros gelo adentro no Lago Whillians, a equipe conseguiu recuperar amostras de água e sedimentos que claramente mostravam sinais de vida. Ainda é preciso realizar mais estudos, incluindo uma análise de DNA, para determinar que tipo de bactéria foi encontrada e como ela vive.

Coleção 2

Após quase quatro meses, o horário de verão terminará à 0h do próximo domingo (17).
Os relógios devem ser atrasados em uma hora nas regiões Sul, Sudeste e Centro-Oeste e no estado do Tocantins. O horário de verão é aplicado no Brasil desde o início da década de 30 e começou em 1985 a ser adotado sem interrupções. Desde 2008, um decreto presidencial estabelece datas fixas para o início e término do horário de verão. De acordo com o decreto, a mudança no horário ocorre, todos os anos, no terceiro domingo de outubro e termina no terceiro domingo de fevereiro.

Coleção 3

A queda de uma estrutura que estava sendo montada para um evento no balneário baiano da Costa do Sauípe deixou ao menos 40 operários feridos nesta quarta-feira, informou a Polícia Militar da Bahia. Um operário que chegou a ficar preso na estrutura foi resgatado, mas seu estado de saúde é grave. Os feridos foram encaminhadas para três hospitais da rede da Secretaria da Saúde do Estado: Hospital Geral de Camaçari e Hospital Menandro de Faria, ambos na região metropolitana de Salvador e o Hospital Geral do Estado, na capital. O Bradesco, patrocinador do evento onde a tenda montada pelos operários desabou, lamentou o ocorrido e disse que está tomando "todas as providências" para atender as vítimas.

Coleção 4

Os conselhos de administração do grupo da American Airlines e da US Airways aprovaram a fusão entre as duas empresas após reuniões na noite de quarta-feira. O negócio foi confirmado em anúncio conjunto nesta quinta-feira. A transação cria a maior companhia aérea do mundo e deve ajudar a American a deixar o processo de recuperação judicial em que se encontra desde 2011. O novo grupo deverá manter a marca American e terá um valor de mercado próximo a US$ 11 bilhões. Fusão pode pôr fim à onda de consolidação no setor.

Coleção 5

Oito anos depois de sua única participação num torneio no Brasil, Rafael Nadal resolveu incluir o País em seu calendário. No Brasil Open, Nadal vai tentar conquistar o bicampeonato, já que foi campeão do torneio em 2005, ano em que conquistou mais dez títulos, incluindo Roland Garros. Além de Nadal, estarão em São Paulo o também espanhol Nicolas Almagro (11º do ranking), o argentino Juan Monaco (12º) e o suíço Stanislas Wawrinka (17º). A presença de Nadal em São Paulo foi confirmada por Luis Felipe Tavares, presidente da empresa de marketing esportivo Koch Tavares, organizadora da competição de nível ATP 250. A participação do espanhol no principal torneio de tênis do país pode marcar sua volta às quadras. Ele está fora do circuito desde julho do ano passado por causa de uma grave lesão no joelho esquerdo e foi obrigado a desistir de competições, como os Jogos Olímpicos de Londres e o US Open. Sem entrar em quadra desde 28 de junho, quando foi eliminado precocemente na segunda rodada de Wimbledon, o tenista número 4 do mundo resolveu dar prioridade às competições em quadra de saibro em seu retorno ao circuito. A compra pode ser feita no site da Tickets For Fun e nas bilheterias do Credicard Hall.

Coleção 6

A equipe do Ravens de Baltimore venceu na noite deste domingo, em Nova Orleans, o 49ers de San Francisco e conquistou seu segundo título do Super Bowl, a Liga Nacional de Futebol Americano (NFL). A 47ª edição do Super Bowl, no Superdome, entrou para a história do futebol americano, atingindo telespectadores em mais de 180 países. O time de Baltimore, fundado em 1996, já havia conquistado o Super Bowl na temporada de 2000. Já o time de San Francisco conheceu seu primeiro revés em uma final da NFL. Até o momento, os 49ers possuem cinco títulos. No primeiro quarto, o Ravens abriu o placar com um touchdown de Anquan Boldin. Logo depois, os 49ers tiveram que se contentar com um field goal. Já no segundo quarto, a equipe de Baltimore abriu grande vantagem, resultando em 21 a 6. O

intervalo foi conduzido por Beyoncé e a seguir no terceiro quarto, o jogador do Ravens, Jacoby Jones, fez mais um touchdown, marcando 28 a 6. E com 13 minutos e 22 segundos no relógio, parte da iluminação do estádio acabou, paralisando o jogo por 34 minutos. Após o retorno da partida, o San Francisco reagiu e Michael Crabtree marcou um towchdown no fim do terceiro quarto. Dois minutos depois, Akers marcou um field goal, diminuindo o placar para 28 a 23. No quarto final, o time de Baltimore finalmente reage e faz um field goal, mas o 49ers diminui o placar para 31 a 29. Ao final, o resultado terminou em 34 a 31 para o Ravens, campeão do Super Bowl.