CARACTERIZAÇÃO DA COMPLEMENTARIDADE TEMPORAL: SUBSÍDIOS PARA SUMARIZAÇÃO AUTOMÁTICA MULTIDOCUMENTO

Jackson Wilke da Cruz SOUZA[*]
Ariani Di FELIPPO[**]

▪ RESUMO: A complementaridade é um fenômeno multidocumento comumente observado entre notícias que versam sobre um mesmo evento. A partir de um *corpus* em português composto por um conjunto de pares de sentenças manualmente anotadas com as relações da *Cross-Document Structure Theory* (CST) que explicitam a complementaridade temporal (*Historical background* e *Follow-up*), identificou-se um conjunto potencial de atributos linguísticos desse tipo de complementaridade. Por meio de algoritmos de Aprendizado de Máquina, testou-se o potencial dos atributos em distinguir as referidas relações. O classificador simbólico gerado pelo algoritmo JRip obteve o melhor desempenho ao se considerar a precisão e o tamanho reduzido do conjunto de regras. Somente com base em 5 regras, tal classificador identificou *Follow-up* e *Historical background* com precisão aproximada de 80%. Ademais, as regras do classificador indicam que o atributo *ocorrência de expressão temporal na sentença 2* é o mais relevante para a tarefa. Como contribuição, salienta-se que o classificador JRip aqui gerado pode ser utilizado nos analisadores discursivos multidocumento para o português do Brasil que são baseados na CST.

▪ PALAVRAS-CHAVE: Descrição linguística. Complementaridade. CST. Sumarização Multidocumento. Processamento Automático de Língua Natural.

Introdução

O acesso e a disponibilização da informação digital têm crescido muito rapidamente. De acordo com as projeções de Taufer (2013), mais de 40 *zetabytes* de informação *online* serão produzidos em 2020. Várias subáreas do Processamento Automático de Línguas Naturais (PLN) buscam desenvolver aplicações computacionais capazes de lidar com essa vasta quantidade de dados.

Uma dessas subáreas é a Sumarização Automática Multidocumento (SAM), na qual se objetiva automatizar a produção de um sumário a partir de uma coleção de textos-

[*] Universidade Federal de São Carlos (UFSCar) / Núcleo Interinstitucional de Linguística Computacional (NILC), São Carlos - São Paulo - Brasil. jackcruzsouza@gmail.com

[**] Universidade Federal de São Carlos (UFSCar) / Núcleo Interinstitucional de Linguística Computacional (NILC), São Carlos - São Paulo - Brasil. Departamento de Letras. arianidf@gmail.com

fonte, advindos de fontes distintas, que abordam um mesmo assunto (MANI, 2001). A SAM tem visado majoritariamente à produção de sumários extrativos (ou extratos), que são comumente compostos por sentenças copiadas integralmente dos textos-fonte. Tais sumários tendem a ser *informativos*, já que veiculam o conteúdo central da coleção a ponto de substituir a leitura dos textos-fonte, e *genéricos*, ou seja, voltados para uma audiência não específica (KUMAR; SALIM; RAZA, 2012).

Os sumários multidocumento têm sido gerados em 3 etapas: (i) análise, (isto é, interpretação dos textos-fonte na qual se extrai uma representação formal dos mesmos), (ii) transformação (ou seja, etapa principal do processo em que, a partir da representação gerada na análise, o conteúdo dos textos-fonte é condensado em uma representação interna do sumário) e (iii) síntese (isto é, produção do sumário por meio da concatenação das sentenças dos textos-fonte até que se atinja um tamanho pré-determinado para o mesmo) (SPARCK-JONES, 1993; MANI, 2001).

De acordo com a quantidade e o nível de conhecimento linguístico, a SAM pode ser superficial ou profunda (MANI, 2001). A SAM superficial utiliza pouco ou nenhum conhecimento linguístico, tratando os textos-fonte estatisticamente e gerando, por isso, extratos. Os métodos/sistemas superficiais são de baixo custo e apresentam robustez e escalabilidade. Entretanto, há ocorrência frequente de problemas de coerência, coesão e informatividade nos extratos gerados por esse paradigma. A SAM profunda, por sua vez, usa conhecimento codificado em gramáticas, repositórios semânticos e modelos de discurso e, por isso, têm aplicação mais restrita e desenvolvimento caro. Por outro lado, esse paradigma gera extratos com menos problemas linguísticos e também *abstracts*.

Tendo em vista a produção de extratos informativos e genéricos, é preciso selecionar as sentenças mais importantes das coleções, evitando-se que elas sejam redundantes e contraditórias e buscando garantir que elas sejam complementares entre si. A redundância, contradição e complementaridade, aliás, são alguns dos chamados fenômenos multidocumento (os quais resultam da multiplicidade de textos-fonte) e seu tratamento visa garantir informatividade e qualidade linguística aos extratos. Tais fenômenos são ilustrados pelos pares de sentenças (S1 e S2) em (1), (2) e (3).

(1) *Redundância*
S1: A margem de erro é de dois pontos percentuais, para mais ou para menos.
S2: A margem de erro é de 2 pontos porcentuais.

(2) *Complementaridade*
S1: No caso do Japão, a magnitude apontada de 6,8 é considerada “forte”.
S2: Em Niigata, um terremoto em outubro de 2004, também de magnitude 6,8, matou 65 pessoas e deixou mais de 3.000 feridos.

(3) *Contradição*
S1: José Maria Eymael, do PSDC, e Rui Pimenta, do PCO, não chegaram a obter 1% das intenções de voto.
S2: Os candidatos José Maria Eymael (PSDC) e Ruy Pimenta (PCO) não pontuaram.

Entre as sentenças de (1), há uma relação de redundância, já que o conteúdo de ambas é bastante semelhante. Entre as sentenças de (2), estabelece-se uma relação de complementaridade, pois S2 detalha uma informação contida em S1. No caso, S2 fornece informações adicionais não relatadas em S1. Especificamente, o S1 fornece informações históricas sobre o terremoto de magnitude 6,8 que atingiu o Japão em 2004. E, finalmente, observa-se uma relação de contradição entre as sentenças de (3), pois, em S2, os candidatos em questão não pontuaram, ao passo que, em S1, os mesmos candidatos obtiveram pontuação muito baixa, que não chegou a 1%.

Especificamente, a identificação desses fenômenos na fase de análise dos textos-fonte é importante porque: (i) as sentenças mais redundantes na coleção veiculam suas principais informações e, por isso, devem constar do sumário; (ii) as sentenças relevantes e complementares entre si também devem compor o sumário, e (iii) as sentenças redundantes ou contraditórias entre si não devem ser selecionadas para o sumário. Para tanto, a descrição linguística desses fenômenos é essencial, uma vez que fornece as pistas a serem rastreadas pelos métodos de SAM. Neste trabalho, em especial, foca-se na complementaridade (do tipo temporal), posto que a redundância (p.ex.: HATZIVASSILOGLOU et al., 2001; NEWMAN et al., 2004; HENDRICKX et al., 2009; SOUZA, DI-FELIPPO, PARDO, 2013) e a contradição (p.ex.: CONDORAVDI et al., 2003; MARNEFFE; RAFFERTY; MANNING, 2008; MARNEFFE, 2012) são os fenômenos mais estudados da literatura.

Na Seção 2, descrevem-se as relações do modelo CST (*Cross-Document Structure Theory)* (RADEV; JING; BUDZIKOWSKA, 2000) que codificam a complementaridade e os principais trabalhos que focam a identificação automática das relações CST. Na Seção 3, apresentam-se o *corpus* utilizado nesta pesquisa e o recorte dos dados em função do estudo da complementaridade temporal. Na Seção 4, apresentam-se as principais características da complementaridade temporal e tradução das mesmas em atributos computacionalmente aplicáveis à tarefa de detecção automática das relações CST que a recobrem. Na Seção 5, descreve-se o processo de descrição linguística do *corpus* necessário à análise da pertinência dos atributos para a detecção da complementaridade temporal. Por fim, na Seção 6, apresentam-se o resultado das avaliações quanto à pertinência dos atributos na tarefa automática de detecção das relações CST que explicitam o fenômeno da complementaridade temporal e as considerações finais sobre o trabalho.

Trabalhos relacionados

Duas sentenças provenientes de textos distintos que abordam um mesmo assunto podem se relacionar de diferentes formas (MANI, 2001). A análise dos relacionamentos entre tais sentenças (ou seja, a análise multidocumento ou intertextual) tem sido investigada há algumas décadas no âmbito do PLN. Uma aplicação de PLN para a qual a análise intertextual se faz relevante é a Sumarização Automática Multidocumento

(SAM), que visa produzir um único sumário a partir do conteúdo de vários textos-fonte. Como resultado das referidas investigações, identificou-se um conjunto de relações retóricas que se estabelecem entre sentenças de textos semanticamente relacionados. Essas relações baseiam-se no modelo CST (RADEV, 2000).

A CST estabelece relações para conectar (em pares) unidades informativas (p.ex.: sentenças) de textos distintos que abordam um mesmo assunto. Originalmente, propôs-se um conjunto de 24 relações intertextual (Quadro 1).

Quadro 1 - 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).

Alguns autores têm refinado as relações CST originais, produzindo conjuntos mais compactos (p.ex.: ZHANG; OTTERBACHER; REDEV, 2003; MAZIERO, 2012; MAZIERO; JORGE; PARDO, 2014). Para o português, reduziu-se o conjunto original a 14 relações, as quais foram organizadas em dois grandes grupos (MAZIERO, 2012; MAZIERO; JORGE; PARDO, 2014): (i) relações de conteúdo (*Identity, Elaboration, Equivalence, Contradiction, Summary, Subsumption, Overlap, Historical background* e *Follow-up*) e (ii) relações de forma (*Attribution, Citation, Modality, Indirect speech* e *Translation*). As relações de conteúdo, em particular, explicitam os seguintes fenômenos multidocumento: redundância, complementaridade e contradição.

Diversos trabalhos têm abordado a importância da CST para a tarefa de sumarização automática. O trabalho proposto por Zhang, Blair-Goldensohn e Radev (2002) foi o primeiro a considerar os relacionamentos estruturais intertextuais, codificados pelas relações CST, para gerar um sumário. Especificamente, os autores utilizam o MEAD (RADEV; JING; BUDZIKOWSKA, 2000; RADEV et al., 2003), um sumarizador baseado em *cluster* e *centroide*[1] para ranquear as sentenças dos textos-fonte e produzir um extrato inicial. Em seguida, as sentenças de baixa relevância selecionadas pelo MEAD para compor o sumário são substituídas por sentenças que possuem mais relações CST na coleção, as quais tendem a ser mais informativas.

[1] De uma forma geral, nos métodos baseados em *clusters* e *centroides*, a análise consiste em agrupar as sentenças de dada coleção em *clusters* com base na similaridade lexical. Assim, os *clusters* são formados por sentenças semelhantes entre si, que representam os "tópicos" da coleção. Cada *cluster* é representado por um centroide, ou seja, um conjunto de palavras estatisticamente importantes. Assim, seleciona-se, em cada *cluster,* a sentença que contém o maior número de palavras do centroide. Os conceitos de *cluster* e *centroide* podem ser conferidos em Jurafsky e Martin (2009).

O trabalho de Jorge e Pardo (2010), voltado para o português, é outro em que a CST é aplicada à SAM. Nele, as sentenças são ranqueadas exclusivamente com base na quantidade de relações CST existentes na coleção. Mais recentemente, Cardoso (2014), também para o português, aplicou a CST em combinação com a *Rethorical Structure Theory* (RST) (MANN; THOMPSON, 1987) e com subtópicos textuais para modelar o processo de sumarização. Todos esses trabalhos de sumarização lidam com *corpora* manualmente anotados, sendo que a tarefa de anotação via CST é cara e demorada, já que requer equipe de especialistas altamente treinada e capaz de produzir um montante considerável de dados.

Buscando contornar esse problema, há algumas propostas para a identificação automática das relações CST, as quais se baseiam em técnicas de Aprendizado de Máquina (AM)[2]. Zhang, Otterbacher e Redev (2003) centraram-se na detecção de seis relações CST que ocorrem entre as sentenças de um texto. O classificador[3] desenvolvido conseguiu identificar sentenças entre as quais não há evidências de relações CST e não obteve muito êxito ao apontar as relações.

Miyabe, Takamura e Okomura (2008) investigaram especificamente as relações *Equivalence* e *Transition*. No método dos autores, a identificação automática de *Equivalence* é necessária para a posterior detecção de *Transition* (isto é, relação que ocorre entre sentenças que veiculam o mesmo conteúdo e que diferem quanto a dados numéricos; *Transition* é equivalente à relação CST *Contradiction*).

Em Zahri e Fukumoto (2011), as relações *Identity*, *Paraphrase* (similar à relação CST *Equivalence*), *Subsumption, Overlap* e *Elaboration* são identificadas em uma aplicação de sumarização. Nesse método de SAM, os títulos de uma coleção de notícias são utilizados para extrair as sentenças com as palavras estatisticamente mais salientes dos documentos. Na sequência, os autores identificam as relações retóricas entre as sentenças por meio de um algoritmo de AM. As relações retóricas indicam a complementaridade e a redundância entre as sentenças ranqueadas. Por fim, os autores ranqueiam as sentenças com base na importância relativa das mesmas na coleção por meio do método *PageRank*[4] (ERKAN; RADEV, 2004) e as mais salientes são selecionadas para o extrato. Segundo os autores, a combinação do *PageRank* com relações retóricas contribui, por exemplo, para evitar a geração de extratos com informações redundantes.

---

[2] O Aprendizado de Máquina é uma área da Inteligência Artificial cujo objetivo é o desenvolvimento de técnicas computacionais para o aprendizado bem como a construção de sistemas (ou algoritmos) capazes de adquirir conhecimento de forma automática a partir de exemplos.

[3] No geral, um classificador pode ser definido como um tipo especial de sistema de regras de produção. Ele comumente resulta da aplicação de um algoritmo de Aprendizado de Máquina a um *corpus* composto por um conjunto de instâncias (exemplos) previamente caracterizas e classificadas. O algoritmo aprende as características que definem a classe, traduzindo-as em uma regra que, por conseguinte, é composta pela condição (características que definem uma classe) e pelo rótulo que representa a classe. Uma vez que as condições são satisfeitas por uma instância nova (devidamente caracterizada, mas não classificada), tal instância recebe o rótulo da respectiva classe (MITCHELL, 1997).

[4] Algoritmo usado pelo Google para ranquear as páginas na web em seu motor de busca (<https://en.wikipedia.org/wiki/PageRank>).

Kumar, Salim e Raza (2012) utilizaram atributos *linguísticos* (como similaridade entre verbos de um par de sentenças), *estruturais* (como tamanho de sentenças) e *estatísticos* (como quantidade de palavras em comum entre as sentenças) para identificar automaticamente as relações *Identity, Overlap, Subsumption*, e *Description*. Para avaliar o método, os autores utilizaram o CSTBank (RADEV; OTTERBACHER; ZHANG, 2004), *corpus* multidocumento de textos jornalísticos em inglês em que as sentenças foram manualmente anotadas com as relações CST. Especificamente, Kumar, Salim e Raza (2012) utilizaram 476 pares de sentenças para treinamento e 206 para teste. Os resultados dos testes realizados com 3 algoritmos de AM revelam que, em comparação às outras relações CST, *Identity* é a relação de mais fácil detecção (medida-f > 90%). Os autores salientam que esse resultado pode estar vinculado à alta similaridade lexical que há entre as sentenças conectadas pela relação *Identity*, o que facilita a identificação automática dessa relação.

Dentre os trabalhos que focaram a identificação automática das relações CST em português, estão o de Maziero (2012), Maziero, Jorge e Prado (2014) e Souza, Di-Felippo e Pardo (2012, 2013).

Em Maziero (2012) e Maziero, Jorge e Prado (2014), tem-se o desenvolvimento do analisador discursivo CSTParser, que detecta as relações CST com base nos atributos linguísticos de similaridade sentencial mais utilizados na literatura: (i) diferença de tamanho (em número de palavras) entre S1 e a S2, (ii) porcentagem de palavras em comum entre S1 e S2 presentes em S1, (iii) porcentagem de palavras em comum entre S1 e S2 presentes em S2, (iv) posição de S1 no texto (início, meio ou fim), (v) número de palavras na maior *string* (isto é, cadeia de caracteres) entre S1 e S2, (vi) diferença no número de substantivos entre S1 e S2, (vii) diferença no número de advérbios entre S1 e S2, (viii) diferença no número de adjetivos entre S1 e S2, (ix) diferença entre o número de verbos entre S1 e S2, (x) diferença entre o número de nomes próprios entre S1 e S2, (xi) diferença no número de numerais entre S1 e S2 e (xii) ocorrência de sinônimos. O CSTParser obteve precisão geral de 68,13%, que é a média da precisão obtida por um classificador para as relações *Overlap, Subsumption, Elaboration, Equivalence, Historical background* e *Follow-up*, e por regras pontuais para as relações *Identity, Contradiction* (explícita), *Attribution, Indirect Speech* e *Translation*[5]. Segundo os autores, essa precisão é considerada adequada, tendo em vista a subjetividade inerente à tarefa de identificação das relações multidocumento.

Souza, Di-Felippo e Pardo (2012, 2013), por sua vez, focalizaram a detecção automática das relações CST de redundância (*Identity*, *Equivalence*, *Summary*, *Subsumption* e *Overlap*) e os tipos de redundância (isto é, total, parcial ou nula) que elas codificam (MAZIERO, 2012). Para o estudo formal do comportamento linguístico da redundância, utilizou-se o CSTNews (CARDOSO et al., 2011), *corpus* multidocumento cujos textos jornalísticos estão anotados em nível sentencial com as relações CST. Além

[5] As relações *Summary, Modality* e *Citation* não foram consideras no método de Maziero (2012) devido à baixa frequência no *corpus* utilizado para treinamento.

da localização das sentenças em seus respectivos textos-fonte, exploraram-se vários tipos de atributos, a saber: (i) sobreposição de palavra[6], (ii) sobreposição de nome e verbo, (iii) sobreposição de padrão morfossintático (p.ex.: nome+preposição+nome), (iv) sobreposição de sujeito (isto é, ocorrência de sujeitos idênticos), (v) sobreposição de verbo principal, (vi) sobreposição de objeto (direto ou indireto), (vii) sobreposição de etiqueta morfossintática e (viii) ocorrência de sinônimos. Utilizando os algoritmos PART (WITTEN; FRANK, 1998) e J48 (QUINLAN, 1993), Souza, Di-Felippo e Pardo (2013) constataram que o classificador que se baseia em todos os atributos determina corretamente os tipos de redundância (total, parcial e nula) com 97,7% de precisão, e as relações CST de redundância, com 62,2%. O segundo melhor classificador utiliza apenas um atributo (sobreposição de nome) e obtém de 91,1% de precisão para os tipos e 60% para as relações CST.

Com base nessa breve revisão da literatura, observa-se que alguns trabalhos focam na detecção automática das relações de CST, em especial as relações de redundância. Ademais, as relações que traduzem a complementaridade (isto é, *Follow-up, Historical background* e *Elaboration*) têm sido identificadas com base em atributos característicos da redundância, como os investigados por Souza, Di-Felippo e Pardo (2012, 2013), e não a partir de atributos específicos desse fenômeno.

O *corpus* CSTNews e a complementaridade

Para a caracterização do fenômeno da complementaridade, selecionou-se o CSTNews (CARDOSO et al., 2011), *corpus* multidocumento de referência para as pesquisas sobre SAM em português. O CSTNews está organizado em 50 *clusters* (ou coleções), totalizando 140 textos, 2.088 sentenças e 47.240 palavras. Os *clusters* estão organizados em categorias, cujos rótulos indicam a seção do jornal da qual os textos-fonte foram compilados, a saber: *mundo* (14), *política* (11), *cotidiano* (13), *ciência* (1), *dinheiro* (1) e *esporte* (10). Os textos foram coletados dos jornais *online Folha de São Paulo, Estadão, O Globo, Jornal do Brasil* e *Gazeta do Povo*.

Cada *cluster* é constituído por: (i) 2 ou 3 textos-fonte, (ii) sumários monodocumento, (iii) 6 *abstracts* e 6 extratos manuais multidocumento, (iv) 1 extrato automático multidocumento, (v) relacionamento entre os texto-fonte via CST, (vi) anotação de expressões temporais nos textos-fonte, (vii) etiquetação morfossintática e sintática dos textos-fonte, (viii) anotação semântica de substantivos e verbos via WordNet de Princeton[7] (FELLBAUM, 1998), (ix) anotação de aspectos informacionais (p.ex.: *o quê*,

[6] Em Souza, Di-Felippo e Pardo (2012, 2013), a sobreposição lexical foi calculada pela medida estatística *word overlap,* que calcula o número de palavras em comum entre duas sentenças. Ademais, consideraram-se variações dessa medida em função das classes dos nomes e verbos, denominadas *noun overlap* e *verb overlap.*

[7] Base léxico-conceitual desenvolvida para o inglês norte-americano em que as unidades lexicais (palavras ou expressões) estão divididas em quatro classes: nome, verbo, adjetivo e advérbio. Em cada classe, as unidades estão organizadas em *synsets* (do inglês, *synonym sets*), ou seja, em conjuntos de formas sinônimas ou quase-sinônimas

onde, etc.) em um dos sumários manuais multidocumento, (x) anotação discursiva dos textos-fonte via RST[8] (MANN; THOMPSON, 1987), e (xi) anotação de subtópicos dos textos-fonte. A anotação CST, em especial, foi feita por 4 linguistas computacionais durante o período de 3 meses. Essa anotação foi feita com o auxílio do editor CSTTool (ALEIXO; PARDO, 2008).

Com base na tipologia de relações CST de Maziero (2012), fez-se um recorte no CSTNews, que consistiu em selecionar, por meio de sua interface *online*[9] de consulta, os pares de sentenças anotadas com *Follow-up, Historical background* e *Elaboration*. De acordo com Maziero (2012), as relações CST de conteúdo *Historical background* e *Follow-up* capturam dois diferentes tipos de complementaridade temporal, os quais estão ilustrados no Quadro 2 por (i) e (ii), respectivamente.

Quadro 2 – Exemplos de complementaridade temporal.

COMPLEMENTARIDADE TEMPORAL	SENTENÇAS
(i) S2 apresenta informações históricas/ passadas sobre algum elemento presente em S1 (S1←S2); o elemento explorado em S2 deve ser o foco de S2 (*Historical background*)	S1: Um **acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo (RDC)**, matou 17 pessoas na quinta-feira, informou nesta sexta-feira um porta-voz das Nações Unidas. S2: **Acidentes aéreos** são frequentes **no Congo**, onde 51 companhias privadas operam com aviões antigos principalmente fabricados na antiga União Soviética.
(ii) S2 apresenta acontecimentos/ eventos que sucederam os acontecimentos/ eventos presentes em S1 (S1←S2); os acontecimentos em S1 e em S2 devem ser relacionados e ter um espaço de tempo relativamente curto entre si (*Follow-up*)	S1: **A pista auxiliar de Congonhas abriu** às 6h, apenas para decolagens. S2: **Congonhas só abriu** para pousos às 8h50.

Fonte: Elaboração própria.

(p.ex.: {car, auto, automobile, machine, motorcar}), os quais representam conceitos lexicalizados por suas unidades constituintes. Os *synsets* estão conectados por meio das relações de antonímia, hiponímia/hiperonímia, meronímia/ holonímia, acarretamento e causa. Mais informações sobre a WordNet de Princeton podem ser consultadas no endereço: https://wordnet.princeton.edu/

[8] O objetivo da RST é analisar um texto quanto a sua coerência. Para tanto, verifica se as unidades mínimas de discurso (*Elementary Discourse Units* - EDUs), que desempenham uma função para que o objetivo do texto seja atingido, estão conectadas entre si. Cada EDU é classificada em núcleo (informação principal) ou satélite (informação adicional). Quando coerente, um texto tem suas unidades conectadas entre si por meio de relações retóricas (também chamadas de relações de coerência ou discursivas), representadas na forma de árvore. Caso uma relação realize a conexão entre um núcleo e um satélite, tem-se uma relação mononuclear; caso conecte somente núcleos, tem-se uma relação multinuclear.

[9] Disponível em: <http://nilc.icmc.usp.br/CSTNews/>.

A complementaridade do tipo (i) é ilustrada por um par de sentenças provenientes de textos distintos que relatam *um acidente aéreo no Congo*. Especificamente, as sentenças do par apresentam conteúdo comum (em negrito), sendo que S2 veicula informação histórica (sublinhada) sobre esse conteúdo, que é a *ocorrência frequente de acidentes aéreos no Congo (por causa do uso de aviões antigos)*. De acordo com a tipologia de Maziero (2012), esse tipo de complementaridade é codificado pela relação CST *Historical background*, caracterizando-se como temporal; no caso, a complementaridade está relacionada ao aspecto frequente ou habitual do evento pontual descrito em S1.

A complementaridade do tipo (ii) é ilustrada por um par de sentenças de textos-fonte distintos que focam os *atrasos e cancelamentos no aeroporto de Congonhas devido ao mau tempo*. As sentenças estão em complementaridade temporal porque S1 e S2 apresentam informação comum (*abertura das pistas do aeroporto de Congonhas*), sendo que S2 veicula um acontecimento que sucedeu ao evento descrito em S1 após um intervalo curto de tempo (no caso, *o horário de abertura da pista (principal) para pouso*). Segundo a tipologia de Maziero, Jorge e Prado (2014), essa complementaridade temporal é explicitada pela relação CST *Follow-up*.

A relação *Elaboration*, ao contrário das ilustradas no Quadro 2, não é de natureza temporal. O exemplo do Quadro 3 ilustra a complementaridade atemporal.

Quadro 3 – Exemplo de complementaridade atemporal.

Complementaridade atemporal	Sentenças
S2 detalha/refina/elabora algum elemento presente em S1 (S1 ←S2), sendo que S2 não deve repetir informações presentes em S1; o elemento elaborado em S2 deve ser o foco de S2 (*Elaboration*)	S1: Apesar da definição, o **cronograma da obra** não foi divulgado. S2: O **cronograma da obra** depende de estudos finais que estão sendo realizados pela Infraero.

Fonte: Elaboração própria.

As sentenças do Quadro 3 são provenientes de textos-fonte distintos sobre a *reforma da pista principal do aeroporto de Congonhas*. Observa-se que S1 e S2 possuem sobreposição de conteúdo (*cronograma da obra*), sendo que S2 fornece uma informação adicional. No caso, a informação adicional relativa a S1 é o foco de S2 e consiste em a razão pela qual o cronograma da obra não foi divulgado (*[o cronograma da obra] depende de estudos finais que estão sendo realizados pela Infraero*).

Na Tabela 1, descrevem-se os dados quantitativos do *corpus* CSTNews no que diz respeito à complementaridade e suas respectivas relações CST.

Tabela 1 – A complementaridade no *corpus* CSTNews.

Complementaridade	Relação CST	Qt. de pares	Total
Atemporal	*Elaboration*	343	343
Temporal	*Follow-up*	293	370
	Historical background	77	
			713

Fonte: Elaboração própria.

Na Tabela 1, observa-se que a complementaridade atemporal, explicitada pela relação *Elaboration*, ocorre em 343 pares de sentenças. Já a complementaridade temporal ocorre em 370 pares de sentenças, sendo 293 casos de *Follow-up* e 77 casos de *Historical background*. No total, o CSTNews é composto por 713 pares de sentenças com complementaridade.

Até o momento, analisaram-se manualmente 45 pares de cada relação CST de complementaridade. O total de 90 pares compõem o que aqui se denominou *subcorpus*. A referida análise permitiu identificar algumas características da complementaridade temporal, expressa especificamente pelas relações *Historical background* e *Follow-up.* Na próxima seção, apresentam-se essas características linguísticas e a conversão das mesmas em atributos detectáveis pelas máquinas.

Proposição de atributos computacionalmente tratáveis

As sentenças provenientes de notícias distintas sobre um mesmo evento que estão em relação de complementaridade são relativamente similares, o que pode ser observado nos pares que aqui ilustram o fenômeno. Essa similaridade está presente na própria definição das relações CST, que, segundo Zhang e Radev (2005), sempre se estabelecem entre sentenças semanticamente relacionadas, em menor ou maior grau. Quanto ao grau de similaridade, as 3 categorias nas quais as relações CST de conteúdo estão organizadas na tipologia de Maziero, Jorge e Prado (2014) podem ser assim ranqueadas: redundância > complementaridade > contradição. Dessa forma, a complementaridade se estabelece entre sentenças com grau intermediário de similaridade. Não se sabe, no entanto, se as diferentes complementaridades temporais, expressas pelas relações *Historical background* e *Follow-up*, apresentam níveis distintos de similaridade. Assim, a redundância é uma característica a ser investigada quanto ao seu potencial distintivo. Para tanto, buscaram-se na literatura os atributos ou *features* que capturam com mais eficácia a similaridade na tarefa de detecção automática da redundância.

Segundo Hatzivassiloglou et al. (2001), Newman et al. (2004) e Souza, Di-Felippo e Pardo (2012, 2013), os três atributos por meio dos quais a similaridade entre sentenças é detectada automaticamente com maior eficácia são: (i) sobreposição de nome, (ii) proximidade posicional no texto-fonte e (iii) sobreposição de subtópico. A sobreposição

de nome se mostra eficiente porque as palavras da classe nominal ocorrem em maior quantidade nos textos. Esse atributo é comumente especificado pela adaptação da medida estatística *overlap*, que recebe o nome *noun overlap* (Nol). Em (4), tem-se a equação utilizada para calcular o atributo Nol. Da aplicação da medida Nol, obtém-se um resultado entre 0 e 1, sendo que valores mais próximos de 1 indicam maior similaridade entre as sentenças do par e valores mais próximos de 0 indicam menor similaridade.

$$(4)\quad \text{Nol}(S1, S2) = \frac{\#\text{ Nomes em comum (S1+S2)}}{\#\text{ Nomes (S1)} + \#\text{ Nomes (S2)}}$$

Quanto a (ii), Souza, Di-Felippo e Pardo (2012, 2013) salientam que a proximidade posicional evidencia similaridade entre sentenças devido à estrutura (típica) de *pirâmide invertida* dos textos jornalísticos. Nessa estrutura, as informações são organizadas de forma decrescente de relevância em: (i) *lead*, isto é, informação principal veiculada no(s) primeiro(s) parágrafo(s) do texto, (ii) corpo do texto, ou seja, detalhes sobre a informação principal fornecidos nos parágrafos intermediários, e (iii) o encerramento do texto (LAGE, 2002).

Assim, quanto menor a distância entre as posições ocupadas pelas sentenças nos textos-fonte, mais conteúdo em comum elas possuem, e vice-versa, ou seja, quanto mais distantes, menos conteúdo em comum. Diante disso, Souza, Di-Felippo e Pardo (2012) propuseram o atributo *distância*, cuja equação está descrita em (5). Por exemplo, dado um par de sentenças S1 e S2, sendo S1 a 6ª sentença do Texto 1 de um *cluster* e S2 a 4ª do Texto 2 do mesmo *cluster*, verificou-se que a distância entre elas é igual a 2 (isto é, duas posições). Tendo em vista a variação de tamanho dos textos do *subcorpus*, a distância entre as sentenças é normalizada, dividindo-se a distância simples entre as sentenças pela maior distância observada entre duas sentenças no conjunto de pares do *subcorpus*. O cálculo representado na equação (5) gera um valor entre 0 e 1, sendo que, quanto mais próximo de 0, menor é a distância entre elas e maior é a similaridade.

$$(5)\quad \text{Distância}(S1, S2) = \frac{\#\text{ Distância entre S1 e S2}}{\#\text{ Maior distância no } \textit{subcorpus}}$$

Quanto a (iii), isto é, *sobreposição de subtópico*, salienta-se que, na estrutura de *pirâmide invertida*, o tópico principal é o *lead* e os detalhes sobre o *lead* são os subtópicos, que podem se ligar direta ou indiretamente ao tópico de acordo a progressão temática (KOCH, 2009). Devido a essa estrutura, observou-se que a proximidade de posição indica redundância, ou seja, conteúdo similar. No entanto, a *pirâmide invertida* é apenas uma diretriz de escrita jornalística e não uma regra e, por isso, as notícias podem apresentar estruturas relativamente diferentes. Por conseguinte, a redundância nem sempre é captura pela proximidade posicional. Dessa forma, a sobreposição

de subtópico textual é uma estratégia relevante para capturar a redundância porque independe da posição ocupada pelas sentenças. Ademais, por ser de natureza semântica, ela é mais informativa que a sobreposição de nomes, pautada na forma das expressões linguísticas. Especificamente, para capturar a similaridade com base nos subtópicos, propõe-se o atributo SubT, que, expresso por valores binários (sim/não), indica se as sentenças veiculam ou não o mesmo subtópico.

Além da redundância ou similaridade, a complementaridade nos pares rotulados por *Historical background* e *Follow-up* se caracteriza por vezes pela ocorrência de marcas temporais, que podem ser advérbios simples ou expressões/locuções.

No par em (6a), que ilustra *Historical background,* nota-se a ocorrência de expressões temporais somente em S1, a saber: *na quinta-feira* e *nesta sexta-feira*. Em (6b), tem-se ocorrência de expressão temporal em S1 (*na noite desta quinta-feira* e *desde o último dia 13*) e em S2 (*em 1996*).

(6a) S1: Um acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo (RDC), matou 17 pessoas na quinta-feira, informou nesta sexta-feira um porta-voz das Nações Unidas
S2: Acidentes aéreos são frequentes no Congo, onde 51 companhias privadas operam com aviões antigos principalmente fabricados na antiga União Soviética.
(6b) S1: A TAM confirmou, na noite desta quinta-feira, que o airbus da TAM estava com o reverso do lado direito desligado, desde o último dia 13.
S2: Em 1996, uma falha no reverso foi a causa do acidente com o Fokker-100 da TAM, ocorrido segundos depois da decolagem, também em Congonhas.

Quanto à *Follow-up,* o par em (7a) caracteriza-se pela ocorrência de expressão temporal em ambas as sentenças. No caso, a expressão às 8h50 em S2 evidencia que a informação complementar (isto é, a *abertura de Congonhas para pouso*) sucede a informação principal expressa em S1 (isto é, a *abertura de Congonhas para decolagens às 6h*). No exemplo em (7b), S1 é constituída pela expressão *durante este domingo, dia 6* e S2 possui uma marca temporal do tipo advérbio (*depois*).

(7a) S1: A pista auxiliar de Congonhas abriu às 6h, apenas para decolagens.
S2: Congonhas só abriu para pousos às 8h50. (7b) S1: Durante este domingo, dia 6, foram travadas lutas sangrentas.
S2: A ofensiva israelense foi lançada depois de uma sequência de ataques do Hezbollah no domingo que causou as maiores baixas para Israel nas quatro semanas do conflito.

Assim, para utilizar essas marcas linguísticas de tempo na detecção automática da complementaridade temporal, especificaram-se mais 4 atributos binários: *ocorrência de*

ET em S1 (doravante, ETS1), *ocorrência de ET em S2* (doravante, ETS2), *ocorrência de advérbio em S1* (ADVS1) e *ocorrência de advérbio em S2* (ADVS2).

No Quadro 4, sintetiza-se o conjunto total de 7 atributos.

Quadro 4 – Atributos para a caracterização da complementaridade temporal.

Atributo	Descrição	Abreviatura
Sobreposição de nome	Captura a redundância com base na sobreposição de nomes entre as sentenças de um par.	Nol
Distância	Captura a redundância com base na diferença de posição/ localização que as sentenças de um par ocupam em seus respectivos textos-fonte	Distância
Sobreposição de subtópico	Captura a redundância com base na sobreposição de subtópicos entre as sentenças de um par.	SubT
Ocorrência de expressão temporal na Sentença 1	Captura a complementaridade temporal com base na ocorrência de locução de valor temporal na primeira sentença de um par	ETS1
Ocorrência de expressão temporal na Sentença 2	Captura a complementaridade temporal com base na ocorrência de locução de valor temporal na segunda sentença de um par	ETS2
Ocorrência de advérbio na Sentença 1	Captura a complementaridade temporal com base na ocorrência de advérbio na segunda sentença de um par	ADVS1
Ocorrência de advérbio na Sentença 2	Captura a complementaridade temporal com base na ocorrência de advérbio na segunda sentença de um par	ADVS2

Fonte: Elaboração própria.

Na próxima seção, descreve-se a caracterização linguística dos 90 pares de sentenças para a avaliação dos atributos.

Caracterização linguística do *corpus*

O processo de caracterização do *subcorpus* consistiu na descrição ou explicitação das informações de cada sentença necessárias à especificação dos atributos e na especificação dos atributos de cada par propriamente dita. Especificamente, as características das sentenças foram herdadas de anotações prévias do CSTNews.

Para os atributos Nol, Distância, ADVS1 e ADVS2, o conjunto dos nomes e advérbios de cada sentença e a posição da mesma no seu respectivo texto-fonte

foram herdados da anotação morfossintática (ou *tagging*) do CSTNews realizada pelo analisador sintático (ou *parser*[10]) PALAVRAS (BICK, 2000). Para tanto, a anotação morfossintática das sentenças do *subcorpus* foi manualmente revisada, o que consistiu em excluir da caracterização os casos de ruído[11] (isto é, erros de anotação) e inserir os casos de omissão (isto é, anotações não realizadas pelo *parser*). No Quadro 5, tem-se a anotação[12] simplificada da sentença *Congonhas só abriu para pouso, às 8h50.* Nela, identificou-se que a sentença ocupa a posição 4 (*s4*) no texto-fonte e que é composta por dois *nomes, Congonhas* (*pos*[13]=*np*") e *pousos* (*pos*="*n*"). Para a caracterização, herdou-se especificamente o lema (forma canônica) dos nomes e advérbios, o que é expresso pelos valores da etiqueta "lemma". Do exemplo do Quadro 5, herdaram-se, assim, os lemas *pouso* e *Congonhas*.

Quadro 5 – Exemplo de anotação morfossintática do CSTNews.

```
</s><s id="s4" text="Congonhas só abriu para pousos, às 8h50.">
    <terminals>
        <t id="1" word="Congonhas" lemma="Congonhas" pos="np"/>
        <t id="2" word="só" lemma="só" pos="adv"/>
        <t id="3" word="abriu" lemma="abrir" pos="v-fin" morph="PS 3S IND VFIN"/>
        <t id="4" word="para" lemma="para" pos="prp"/>
        <t id="5" word="pousos" lemma="pouso" pos="n"/>
        <t id="6" word="," lemma="--" pos="pu"/>
        <t id="7" word="a" lemma="a" pos="prp"/>
        <t id="8" word="as" lemma="o" pos="art"/>
        <t id="9" word="8h50" lemma="8h50" pos="n"/>
        <t id="10" word="." lemma="--" pos="pu"/>
    </terminals>
</s>
```

Fonte: Elaboração própria.

Os subtópicos foram recuperados da anotação descrita em Cardoso et al. (2011). Nos Quadros 6 e 7, têm-se os subtópicos que compõem os textos-fonte do caso de *Historical background* do Quadro 2. Na anotação[14], vê-se que a S1 veicula o subtópico

[10] Ferramenta computacional que reconhece a estrutura sintática de uma sentença, atribuindo funções sintáticas aos constituintes reconhecidos (CARROL, 2004).

[11] Esse é o caso, por exemplo, de *8h50,* anotado equivocadamente como *nome*.

[12] Toda sentença (*s*) é anotada com 2 atributos, cujas etiquetas são: *id*, posição da sentença no texto, e *text* (a própria sentença do *corpus*). Os elementos constitutivos de *s* (palavras/expressões e símbolos de pontuação) são chamados *terminals* (ou *tokens*). Cada um deles é descrito por 4 atributos: *id* (posição na sentença)*, word* (ocorrência da palavra/expressão)*, lemma* (forma canônica) e *pos* (classe de palavras).

[13] Do inglês, *part-of-speech*.

[14] Os subtópicos foram explicitados no CSTNews por meio da seguinte notação: <t LABEL= "breve descrição do subtópico"> TOP= "número do tópico na coleção"> (CARDOSO et al*.,* 2011).

1 da coleção, rotulado por *acidente aéreo* e a S2, cujo texto-fonte anotado está no Quadro 7, veicula o subtópico 2 da mesma coleção, representado pelo rótulo *histórico*.

Quadro 6 – Exemplo de anotação de subtópicos do CSTNews (Documento 2 do *Cluster* 1).

S1: Um acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo (RDC), matou 17 pessoas na quinta-feira, informou nesta sexta-feira um porta-voz das Nações Unidas. S2: As vítimas do acidente foram 14 passageiros e três membros da tripulação. S3: Todos morreram quando o avião, prejudicado pelo mau tempo, não conseguiu chegar à pista de aterrissagem e caiu numa floresta a 15 quilômetros do aeroporto de Bukavu. S4: Segundo fontes aeroportuárias, os membros da tripulação eram de nacionalidade russa. S5: O avião explodiu e se incendiou, acrescentou o porta-voz da ONU em Kinshasa, Jean-Tobias Okala. S6: "Não houve sobreviventes", disse Okala. **<t LABEL="acidente aéreo" TOP= "1")>**
S7: O porta-voz informou que o avião, um Soviet Antonov-28 de fabricação ucraniana e propriedade de uma companhia congolesa, a Trasept Congo, também levava uma carga de minerais. **<t LABEL="detalhes do avião" TOP= "3">**

Fonte: Elaboração própria.

Quadro 7 – Exemplo de anotação de subtópicos do CSTNews (Documento 1 do *Cluster* 1).

S1: Ao menos 17 pessoas morreram após a queda de um avião de passageiros na República Democrática do Congo. S2: Segundo uma porta-voz da ONU, o avião, de fabricação russa, estava tentando aterrissar no aeroporto de Bukavu em meio a uma tempestade. S3: A aeronave se chocou com uma montanha e caiu, em chamas, sobre uma floresta a 15 quilômetros de distância da pista do aeroporto. **<t LABEL= "acidente aéreo" TOP= "1">**
S4: Acidentes aéreos são freqüentes no Congo, onde 51 companhias privadas operam com aviões antigos principalmente fabricados na antiga União Soviética. **<t LABEL= "histórico" TOP= "2">**
S5 O avião acidentado, operado pela Air Traset, levava 14 passageiros e três tripulantes. S6: Ele havia saído da cidade mineira de Lugushwa em direção a Bukavu, numa distância de 130 quilômetros. **<t LABEL= "avião acidentado" TOP= "1">**

S7: Aviões são usados extensivamente para transporte na República Democrática do Congo, um vasto país no qual há poucas estradas pavimentadas.
S8: Em março, a União Européia proibiu quase todas as companhias aéreas do Congo de operar na Europa.
S9: Apenas uma manteve a permissão.
S10: Em junho, a Associação Internacional de Transporte Aéreo incluiu o Congo num grupo de vários países africanos que classificou como ““uma vergonha”” para o setor.

<t LABEL= “histórico” TOP= “2”>

Fonte: Elaboração própria.

As ETs, por sua vez, foram herdadas da anotação descrita em Menezes-Filho e Pardo (2011). Segundo os autores, as sentenças do CSTNews foram anotadas com base na tipologia de Baptista, Hagège e Mamede (2008), que especificaram 4 tipos de ETs: (i) tempo calendário, (ii) frequência (p.ex.: *Ocorrerá entre os dias 29 e 31 de julho*), (iii) duração (p.ex.: *O Natal é comemorado todo ano*) e (iv) genérico (p.ex.: *Eu gosto do mês de julho*). As ETs que expressam tempo calendário podem ser de 3 subtipos: (i) hora (p.ex.: *Ele chegou às 9h30m*), (ii) data e (iii) intervalo (p.ex.: *Entre junho e julho*). E, por fim, as ETs do subtipo data podem ser: (i) enunciação (p.ex.: *Partiu em março*), (ii) textual (p.ex.: *Um dia após a venda*) e (iii) absoluto (p.ex.: *O acidente ocorreu em fevereiro de 2002*). Dessa forma, o CSTNews fornece não só a anotação de ocorrência, mas também do tipo de ETs, a qual foi considerada. Por exemplo, a S1 do caso de *Follow-up* do Quadro 2 foi anotada como descrito em (8). Nessa anotação, recuperou-se que a ET às 6h é do tipo *tempo calendário* e do subtipo *hora.*

(8) A pista auxiliar de Congonhas abriu **<ET TIPO=”TEMPO_CALEND” SUBTIPO=”HORA”>**às 6h **</ET>**, apenas para decolagens

A descrição do *subcorpus* em função dos dados linguísticos subjacentes aos 7 atributos foi organizada em um único arquivo do tipo *xls,* o qual é aqui ilustrado em dois Quadros (8 e 9) por uma questão de espaço. No Quadro 8, ilustra-se a caracterização dos exemplos do Quadro 2 necessária ao cálculo dos atributos numéricos Nol e Distância.

Quadro 8 – Caracterização das sentenças para o cálculo dos atributos numéricos.

Corpus		Descrição linguística	
Par	Relação CST	Nome	Posição da sentença
1	*Historical background*	acidente, localidade, Bukavu, leste, República Democrática do Congo, RDC, pessoa, porta-voz, Nações Unidas, quinta-feita, sexta-feira	1
		acidente, Congo, companhia, avião, União Soviética	4
2	*Follow-up*	pista, Congonhas, decolagem	6
		Congonhas, pouso	4

Fonte: Elaboração própria.

No Quadro 9, ilustra-se a descrição das informações linguísticas necessárias à especificação dos atributos binários. O símbolo *X* indica que o dado linguístico não ocorreu na sentença.

Quadro 9 – Caracterização das sentenças para o cálculo dos atributos binários.

Corpus		Descrição linguística		
Par	Relação CST	Subtópico	ET	ADV
1	*Historical background*	1	calendário/data/enunciado	X
		2	X	X
2	*Follow-up*	1	calendário/hora	X
		1	calendário/hora	X

Fonte: Elaboração própria.

Uma vez que as sentenças do *subcorpus* foram descritas, procedeu-se à especificação manual dos atributos para cada par sentencial. Os dados resultantes da especificação dos atributos relativos ao conjunto de 90 pares do *subcorpus* foram organizados em um arquivo no formato *xls,* como ilustrado no Quadro 10.

Quadro 10 – Cálculo dos atributos dos pares de sentenças.

Corpus		**Atributos**						
Par	Relação CST	Nol	Distância	SubT	ETS1	ETS2	ADVS1	ADVS2
1	*Historical background*	0,133	3	não	sim	não	não	não
2	*Follow-up*	0,4	2	sim	sim	sim	não	não

Fonte: Elaboração própria.

No Quadro 10, observa-se que, aplicando a medida Nol, o par 1, por exemplo, possui sobreposição de nome com valor 0,125, indicando baixa similaridade. Esse valor resulta do fato de que, entre os 16 nomes distintos presentes na S1 (acidente, localidade, Bukavu, leste, República Democrática do Congo, RDC, pessoa, porta-voz, Nações Unidas, *quinta-feira, sexta-feira*) e os 5 nomes constitutivos da S2 (acidente, Congo, companhia, avião, União Soviética), somente *acidente* ocorre em ambas. Quanto ao atributo Distância, salienta-se que os valores obtidos foram normalizados[15] devido ao tamanho ou extensão diferente dos textos-fonte. Assim, o valor inicial do atributo Distância do par de *Historical background* ilustrado no Quadro 8, 3, é dividido pelo valor mais alto do referido atributo no *subcorpus* (isto é, 34), obtendo-se o valor 0,088 para a distância normalizada do par. Ademais, observa-se ainda com base no Quadro 10 que esse par não possui sobreposição de subtópicos, já que as sentenças S1 e S2 têm, respectivamente, os subtópicos *acidente aéreo* e *histórico*. Observa-se ainda que no referido par de sentenças não há ocorrência de advérbios, mas há ocorrência de expressão temporal em S1 (no caso, *na quinta-feira* e *nesta sexta-feira*).

Diante da especificação dos atributos dos pares do *subcorpus* expandido, procedeu-se à avaliação da pertinência dos mesmos para a distinção das relações CST *Historical background* e *Follow-up*.

Avaliação dos atributos

A pertinência dos atributos foi investigada automaticamente por meio de algoritmos de AM, disponíveis no *Weka* (do inglês, *Waikato Environment for Knowledge Analysis*) (HALL et al., 2009), isto é, *software* de domínio público desenvolvido pela Universidade de Waikato (Nova Zelândia) que contém uma série de algoritmos de AM pertencentes a diferentes paradigmas de Inteligência Artificial.

Neste trabalho, utilizou-se a técnica supervisionada de aprendizado indutivo. Nela, um conjunto de exemplos (ou *corpus*) de treinamento cujas classes (isto é, informação a ser aprendida) são conhecidas é fornecido ao algoritmo de aprendizado. Em geral,

[15] A normalização permite reduzir as chances dos dados se tornarem inconsistentes quando comparados entre si. Como resultado da normalização, obtém-se valores entre 0 e 1.

cada exemplo de treinamento é constituído por um objeto de entrada (isto é, um par de sentenças e seus atributos) e por um objetivo de saída ou classe (isto é, a relação CST do par). O objetivo do algoritmo de indução é construir um classificador que possa determinar corretamente a classe de novos exemplos ainda não rotulados.

Ademais, aplicou-se a técnica *10-fold cross validation*[16], que estima de forma mais realista as taxas de erros de classificação, já que o *subcorpus* é pequeno. Para avaliar os resultados, 3 medidas estatísticas são comumente utilizadas: precisão (do inglês, *precision*), cobertura (do inglês, *recall*) e medida-f (do inglês, *f-measure*) (MITCHELL, 1997). A precisão indica o número de pares de sentenças cujas classes (no caso, as relações CST) foram corretamente identificadas pelo classificador em relação ao total de pares de sentenças. A cobertura indica o número de pares de sentenças corretamente identificados em relação à quantidade que deveria ser identificado. A medida-f é a média ponderada dos valores de precisão e cobertura. Nesta pesquisa, a avaliação dos resultados foi feita com base na precisão (ou acurácia) geral.

Dentre os diferentes paradigmas de AM, isto é, conexionista, matemático (ou probabilístico) e simbólico, focalizaram-se os simbólicos, pois eles descrevem os padrões aprendidos em uma linguagem de fácil compreensão para humanos. No entanto, testaram-se algoritmos dos outros paradigmas para comparação.

Para a tarefa de detecção das relações CST de complementaridade temporal, testou-se o conhecido algoritmo conexionista denominado Multi Layer Perception (MLP), com os configurações *default* do Weka. O algoritmo em questão obteve 82,2% de precisão geral. Dentre os vários do paradigma matemático/probabilístico, testaram-se os algoritmos Naïve-Bayes e SMO. O Naïve-Bayes obteve a mesma precisão alcançada pelo MLP, isto é, 82.2%, enquanto o SMO alcançou 80% de precisão geral.

Quanto ao paradigma simbólico, selecionaram-se os mesmos algoritmos simbólicos utilizados previamente em investigações correlatas, como nos trabalhos de Maziero (2012) e Souza, Di-Felippo e Pardo (2012, 2013). Assim, dentre os vários algoritmos do paradigma simbólico disponíveis no Weka, utilizaram-se os seguintes: One-R (ou *One Rule*) (HOLTE, 1993), PART (WITTEN; FRANK, 1998), JRip (COHEN, 1995) e J48 (QUINLAN, 1993). O OneR é possivelmente o mais simples, pois ele se baseia na hipótese de que basta utilizar apenas um dos atributos para classificar corretamente os exemplos. Assim, esse algoritmo tem a tarefa de encontrar, durante o treinamento, o atributo que apresenta a menor taxa de erro de classificação. O JRip e o PART são algoritmos baseados em regras de decisão, que analisam o conjunto de instâncias e geram classificadores que se constituem de regras no formato lógico, os quais comumente combinam atributos para capturar mais adequadamente as classes. O J48 disponibiliza os resultados no formato de árvores de decisão. A abordagem de construção da árvore é *top-down* e, por isso, o atributo mais discriminativo (logo, o mais genérico) constitui

[16] O método de validação cruzada, denominado *10-fold cross validation*, consiste em dividir o conjunto total de dados em 10 subconjuntos mutuamente exclusivos do mesmo tamanho. A partir deles, 1 subconjunto é utilizado para teste e os 10-1 restantes são utilizados para treinamento. Esse processo é realizado 10 vezes, alternando-se o subconjunto de teste a cada iteração. Ao final das 10 iterações, a acurácia média é calculada.

o nó inicial da árvore e os nós-filhos são os menos discriminativos em comparação ao(s) anterior(es).

Os algoritmos J48 e PART geraram os maiores conjuntos de regras (13 e 8, respectivamente) com precisão geral bastante similar. No caso, ambos obtiveram aproximadamente 81% de precisão. O One-R, como mencionado, utiliza o atributo mais discriminativo para produzir um único conjunto de regras a partir desse atributo. No caso, esse algoritmo selecionou o atributo ETS2 e surpreendentemente também atingiu precisão aproximada de 80%. O JRip, por sua vez, obteve o menor conjunto de regras com precisão também similar aos demais, isto é, 79%.

Assim, ao combinar alta precisão com menor (e mais simples) conjunto de regras, considerou-se o classificador do JRip como o de melhor desempenho para a tarefa em questão. Na Tabela 2, apresentam-se as regras do JRip, sendo que cada uma delas é seguida pela quantidade de instâncias (pares) que foi correta e incorretamente classificada e a precisão da regra, dada pelo número de instâncias classificadas corretamente sobre todas as instâncias classificadas pela regra.

Tabela 2 – Classificador do JRip para as relações de complementaridade temporal.

Regra	Correta	Incorreta	Precisão
1. Se ETS2= hora, então *Follow-up*	12	0	100%
2. Senão ETS2= não, então *Follow-up*	30	8	78.9%
3. Senão Nol>=0.315 e ETS1=data-absoluto, então *Follow-up*	4	0	100%
4. Senão ETS2=hora_data-enunciacao, então *Follow-up*	3	0	100%
5. Senão *Historical background*	41	4	91,1

Fonte: Elaboração própria.

Nessas regras, observa-se que os atributos ETS2, Nol e ETS1 caracterizam os pares anotados com a relação CST *Follow-up.* Dentre eles, ETS2 é o mais característico, já que 3 das 5 regras se baseiam somente nele. Aliás, a relevância de ETS2 também foi detectada pelo algoritmo de seleção de atributos[17] denominado Info Gain Attribute Eval, também disponível no Weka. Além disso, observa-se que, caso nenhuma das 4 primeiras regras seja aplicável (ou satisfeita), a relação padrão é *Historical background,* o que é indicado pela regra 5.

É interessante ressaltar também a produtividade das regras. Por exemplo, as regras 1 e 2 lidam com mais casos que as regras 3 e 4, o que é natural dado o modo como o processo de AM escolhe os atributos para iniciar o conjunto de regras. Assim, pode-se dizer que o classificador também atingiria bons resultados somente com base nas regras

[17] O objetivo do processo de seleção de atributos é melhorar o desempenho dos algoritmos. Esse processo é importante porque os atributos podem ser relevantes ou irrelevantes e identificá-los pode reduzir o tempo de processamento dos dados e gerar modelos mais simples.

1 e 2 para *Follow-up* e a regra 5 para a relação *Historical background*. Na Tabela 3, tem-se a matriz de confusão gerada pelo JRip quando da aplicação das regras, por meio da qual é possível verificar como o classificar lida especificamente com cada classe (relação CST). Cada coluna da matriz representa as instâncias (pares) preditas para uma classe (relação), enquanto que cada linha representa as instâncias reais de cada classe.

Tabela 3 – Matriz de confusão gerada pelo algoritmo JRip.

Classe / **Teste**	*Follow-up* (45)	*Historical background* (45)
Follow-up	35	10
Historical background	9	36

Fonte: Elaboração própria.

Com base na Tabela 3, observa-se que, do total de 45 pares de *Follow-up,* as regras do JRip classificaram corretamente 35 delas, e que, dos 45 pares de *Historical background,* o mesmo conjunto de regras classificou corretamente 36. Diante desse desempenho, vê-se que o JRip classificou os pares de cada relação de forma muito similar.

Ressalta-se, por um lado, que tais resultados são somente indicativos das características da complementaridade temporal e das relações CST que a codificam, assim como do poder discriminativo dos atributos que aqui foram propostos. Diz-se isso principalmente por conta do tamanho reduzido do *subcorpus,* a partir do qual as características/atributos foram identificados e também testados. Por outro lado, vale dizer que este trabalho é pioneiro no que diz respeito à investigação da complementaridade como fenômeno linguístico que precisa ser tratado na SAM.

Assim, como trabalho futuro, objetiva-se criar um *corpus* de teste distinto do *subcorpus* de 90 pares aqui utilizado. Para tanto, pretende-se recortar outro *subconjunto* de pares de sentenças do CSTNews anotados com as relações *Follow-up* e *Historical background.* Com isso, o *subcorpus* de 90 pares poderá ser usado somente como *corpus* de treinamento e o classificador resultante poderá ser testado em um conjunto de pares distinto do que treinamento.

Ademais, pretende-se estudar se a complementaridade atemporal, codificado pela relação CST *Elaboration* apresenta características linguísticas de podem ser traduzidas em atributos detectáveis pelas máquinas.

Agradecimentos

Agradecemos a CAPES e a FAPESP pelo financiamento, e ao coordenador do Núcleo Interinstitucional de Linguística Computacional (NILC) por ceder espaço e conhecimento para a realização desta pesquisa.

SOUZA, J.; Di FELIPPO, A. Characterization of temporal complementarity: fundamentals for multi-document summarization. **Alfa**, São Paulo, v.62, n.1, p.125-150, 2018.

- *ABSTRACT: Complementarity is a usual multi-document phenomenon that commonly occurs among news texts about the same event. From a set of sentence pairs (in Portuguese) manually annotated with CST (Cross-Document Structure Theory) relations (Historical background and Follow-up) that make explicit the temporal complementary among the sentences, we identified a potential set of linguistic attributes of such complementary. Using Machine Learning algorithms, we evaluate the capacity of the attributes to discriminate between Historical background and Follow-up. JRip learned a small set of rules with high accuracy. Based on a set of 5 rules, the classifier discriminates the CST relations with 80% of accuracy. According to the rules, the occurrence of temporal expression in sentence 2 is the most discriminative feature in the task. As a contribution, the JRip classifier can improve the performance of the CST-discourse parsers for Portuguese.*

- *KEYWORDS: Linguistic description. Complementarity. CST. Multi-document Summarization. Natural Language Processing.*

REFERÊNCIAS

ALEIXO, P.; PARDO, T. A. S. CSTTool: um parser multidocumento automático para o Português do Brasil. In: IV Workshop on MSc Dissertation and PhD Thesis in Artificial Intelligence – WTDIA. 2008. **Proceedings**… Salvador/Brasil.

BAPTISTA, J.; HAGÈGE, C.; MAMEDE, N. Proposta de anotação e normalização de expressões temporais da categoria TEMPO para o HAREM II. In: Encontros do Segundo HAREM, 2008. **Actes**... p.1-24.

BICK, E. **The parsing system "PALAVRAS":** Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. 2000. 412 f. Thesis (PhD) - Aarhus University, Denmark University Press, 2000.

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

CARDOSO, P. C. F.; MAZIERO, E. G.; JORGE, M. L. C.; SENO, E. M. R.; DI FELIPPO, A.; RINO, L. H. M.; NUNES, M. G. V.; PARDO, T. A. S. CSTNews – A discourse-annotated corpus for single and multi-document summarization of news texts in brazilian portuguese. In: 3rd RST Brazilian Meeting, 2011. **Proceedings**… Cuiabá/ Brazil. p.88-105.

CARROL, J. Parsing. In: MITKOV, R. (Ed.). The **Oxford handbook of computational linguistics.** Oxford/USA. Ed. Oxford University Express, 2004. p.233-248.

COHEN, W. Fast effective rule induction. In: 12th International Conference on International Conference on Machine Learning, 1995. **Proceedings**... California/USA. p.115-123.

CONDORAVDI, C.; CROUCH, D.; DE PAIVA, V.; STOLLE, R.; BOBROW, D. G. Entailment, intensionality and text understanding. In: HLT-NAACL 2003 workshop on Text meaning, 9., 2003. **Proceedings**... Edmonton/Canada: Association for Computational Linguistics, 2003. p.38-45.

ERKAN, G.; RADEV, D. R. LexPageRank: prestige in multi-document text summarization. In: Empirical Methods in Natural Language, 2004. **Proceedings**… Barcelona/Spain. p.365-371.

FELLBAUM, C. **WordNet:** an electronic lexical database. California: Ed. MIT Press, 1998.

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

HATZIVASSILOGLOU, J. L.; KLAVANS, J. L.; HOLCOMBE, M.; BARZILAY, R.; MCKEOWN, K. Simfinder: a flexible clustering tool for summarization. In: NAACL Workshop on Automatic Summarization, 2001. **Proccedings**… Pittsburgh/USA. p.1-9.

HENDRICKX, I.; DAELEMANS, W.; MARSI, E.; KRAHMER, E. Reducing redundancy in multi-document summarization using lexical semantic similarity. In: 2009 Workshop on Language Generation and Summarisation, 2009. **Proceedings**... Suntec/Singapore: Association for Computational Linguistics, 2009. p.63-66.

HOLTE, R. C. Very simple classification rules perform well on most commonly used datasets. **Machine Learing**, Ed. Springer, v.11, n.1, p.63-90, 1993.

JORGE, M. L. C.; PARDO, T. A. S. Experiments with CST-based Multidocument Summarization. In: ACL Workshop TextGraphs-5: Graph-based Methods for Natural Language Processing, 2010. **Proceedings**… Uppsala/Sweden. p.74-82.

JURAFSKY, D.; MARTIN, J. H. **Speech and language processing:** an introduction to natural language processing, computational linguistics, and speech recognition. v.2. Englewood: Ed. Prentice Hall, 2009.

KOCH, I. G. V. **Introdução à linguística textual:** trajetória e grandes temas. 2. ed. São Paulo: Contexto, 2009.

KUMAR, Y. J.; SALIM, N.; RAZA, B. Cross-document structural relationship identification using supervised machine learning. **Applied Soft Computing**, v.12, n.10, p.3124-3131, out. 2012.

LAGE, N. **Estrutura da notícia**. São Paulo: Ed. Ática, 2002.

MARNEFFE, M-C DE. **What's that supposed to mean? Modeling the pragmatic meaning of utterances**. 2012. 178 f. Thesis (PhD in Linguistics) – Department of Linguistics, Stanford University, Stanford, 2012.

MARNEFFE, M-C DE.; RAFFERTY, A. N.; MANNING, C. D. Finding contradictions in text. In: Annual meeting of the ACL, 46., 2008. **Proceedings**... Columbus/USA, p.1039-1047.

MANI, I. **Automatic Summarization.** Amsterdam/Netherlands: Ed. John Benjamins Publishing Company, 2001.

MANN, W. C.; THOMPSON, S. A. **Rhetorical structure theory**: a theory of text organization. California/USA: Ed. University of Southern California – Information Sciences Institute, 1987. p.87-190.

MAZIERO, E. G. **Identificação automática de relações multidocumento**. 2012. 117 f. Dissertação (Mestrado em Ciências da Computação) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2012.

MAZIERO, E. G.; JORGE, M. L. R. C.; PARDO, T. A. S. Revisiting Cross-document Structure Theory for multi-document discourse parsing. **Information Processing & Management**, v.50, n.2, p.297-314, 2014.

MENEZES FILHO, L. A.; PARDO, T. A. S. Detecção de Expressões Temporais no Contexto de Sumarização Automática. In: 2nd STIL Student Workshop on Information and Human Language Technology, 2011. **Proceedings**… Cuiabá/Brasil. p.1-3.

MITCHELL, T. M. **Machine learning**. v.45. Burr Ridge, IL: McGraw Hill, 1997.

MIYABE, Y.; TAKAMURA, H.; OKUMURA, M. Identifying cross-document relations between sentences In: 3rd International Joint Conference on Natural Language, 2008. **Proceedings**… Hyderabad/India. p.141-148.

NEWMAN, E.; DOMN, W.; STOKES, N.; CARTHY, J.; DUNNION, J. Comparing redundancy removal techniques for multi-document summarization. In: Starting AI researchers' symposium, 2004. **Proceedings**…Valencia/Spain. p.223-228.

QUINLAN, J. **Programs for machine learning.** San Mateo/USA: Ed. Morgan Kaufmann Publishers, 1993.

RADEV, D. A common theory of information fusion from multiple text sources step one: cross-document structure. In: 1st SIGdial workshop on Discourse and dialogue, 10,

2000. **Proceedings**… Hong Kong/China: Association for Computational Linguistics, 2000. p.74-83.

RADEV, D. R.; JING, H.; BUDZIKOWSKA, M. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation and user studies. In: ANLP/NAACL Workshop on Automatic Summarization, 2000. **Proceedings...** Seatle/WA: North American Association for Computational Linguistics, 2000. p.21-29.

RADEV, D. R.; TEUFEL, S.; SAGGION, H.; LAM, W.; BLITZER, J.; QI, H.; CELEBI, A.; LIU, D.; DRABEK, E. Evaluation challenges in large-scale multi-document summarization: the mead project. In: 41st Annual Meeting of the Association for Computational, 2003. **Proceedings**… Sapporo/Japan: Association for Computational Linguistics, 2003. p.375-382.

RADEV, D.; OTTERBACHER, J.; ZHANG, Z. CSTNank: A corpus for the study of Cross-document Structural Relationship. In: **Proceedings fo Fourth International Conference on Language Resources and Evaluation**. Lisboa, 2004.

SOUZA, J. W. C. **Descrição linguística da complementaridade para a Sumarização Automática Multidocumento.** 2015. 105 f. Dissertação (Mestrado em Linguística) – Centro de Educação e Ciências Humanas, Universidade Federal de São Carlos, São Carlos, 2015.

SOUZA, J. W. C.; DI-FELIPPO, A.; PARDO, T. A. S. **Investigação de métodos de identificação de redundância para Sumarização Automática Multidocumento**. Série de Relatórios do NILC (NILC-TR-12): São Carlos/Brasil. 30 f. 2012.

_____. Identificação da redundância na Sumarização Automática Multidocumento: explorando métodos superficiais. In: 3rd Student Workshop on Information and Human Language Technology (TILic), 2013. **Proceedings**… Fortaleza/Brasil. p.1-3.

SPARCK JONES, K. What might be in a summary? **Information Retrieval**. v.93, p.9-26, 1993.

TAUFER, P. Massa de informações digitais pode ser usada em benefício da população. **Jornal da Globo**, 26 dez. 2013. Disponível em: <http://g1.globo.com/jornal-da-globo/noticia/2013/12/massa-de-informacoes-digitais-pode-ser-usada-em-beneficio-da-populacao.html>. Acesso em: 02 fev. 2015.

WITTEN, I. H.; FRANK, E. **Generating accurate rule sets without global optimization.** Working paper series, 1998.

ZAHRI, N. A. H. B.; FUKUMOTO, F. Multi-document summarization using link analysis based on rhetorical relations between sentences. In: 12th International Conference on Computational Linguistics and Intelligent Text Processing, 2., 2011. **Proceedings**… Tokyo/Japan. p.328-338.

ZHANG, Z.; RADEV, D. Combining labeled and unlabeled data for learning cross-document structural relationships. In: **Natural Language Processing** – I JCNLP. Springer, p.32-41, 2005.

ZHANG, Z.; OTTERBACHER, J.; REDEV, D. R. Learning cross-document structural relationships using boosting. In: International Conference on Information and Knowledge Management, 2003. **Proceedings**… Las Vegas/USA. p.124-130.

ZHANG, Z.; BLAIR-GOLDENSOHN, S.; RADEV, D. R. Towards CST-enhanced summarization. In: Innovative applications of artificial intelligence conference, 2002. **Proceedings**… Edmonton/Canada. p.439-446.

Recebido em 30 de dezembro de 2016

Aprovado em 30 de julho de 2017