Proceedings of Symposium in Information and Human Language Technology. Natal, RN, Brazil, November 4–7, 2015.

Em Direção à Caracterização da Complementaridade no *Corpus* Multidocumento CSTNews

Jackson Souza[1,3], Ariani Di Felippo[1,2]

[1]Núcleo Interinstitucional de Linguística Computacional – (NILC)
Caixa Postal 668 – 13560-970 São Carlos, SP, Brasil

[2]Departamento de Letras – Universidade Federal de São Carlos (UFSCar)
[3]Programa de Pós-graduação em Linguística (PPGL) – UFSCar
Caixa Postal 676 – 13.565-905– São Carlos, SP, Brasil

{jackcruzsouza, arianidf}@gmail.com

Abstract. *We present a description of the complementarity in CSTNews, multi-document corpus of news texts in Brazilian Portuguese. As a result, we identified a set of formal attributes that characterizes the relations from the Cross-document Structure Theory (CST) model that codify complementarity. Such attributes can be used to automatically detect complementarity.*

Resumo. *Apresenta-se uma investigação da complementaridade no CSTNews, corpus multidocumento de notícias em Português do Brasil. Como resultado, identificou-se um conjunto de atributos explícitos que caracterizam as relações do modelo Cross-document Structure Theory (CST) que codificam a complementaridade, os quais têm potencial para subsidiar a detecção automática desse fenômeno.*

1. Introdução

A análise semântico-discursiva de múltiplos textos que abordam um mesmo assunto tem sido tópico de muitas pesquisas no Processamento Automático das Línguas Naturais (PLN) nos últimos anos. Um exemplo desse tipo de análise é a identificação de relações como as baseadas no modelo/teoria CST (*Cross-document Structure Theory*) [Radev 2000]. No trabalho de Maziero *et al* (2010), há 14 relações CST: *Identity, Elaboration, Equivalence, Contradiction, Summary, Citation, Subsumption, Attribution, Overlap, Modality, Historical background, Indirect speech, Follow-up* e *Translation*. Tais relações codificam diferentes fenômenos multidocumento, a saber: (i) conteúdo, como redundância (p.ex.: *Identity*, *Equivalence*, etc.), complementaridade (p.ex.: *Historical background*) e contradição (*Contradiction*); e (ii) forma, como variação de fonte/autoria (p.ex.: *Citation*) e estilo (p.ex.: *Translation*).

As relações CST são amplamente usadas em aplicações de Sumarização Automática Multidocumento (SAM), as quais comumente buscam gerar uma versão concisa de uma coleção de textos na forma de um sumário coeso e coerente, composto pela justaposição das sentenças mais importantes da coleção, selecionadas na íntegra [Kumar e Salim 2012]. Para tanto, os métodos usuais de SAM ranqueiam as sentenças dos textos-fonte pela redundância de seu conteúdo, codificada pela quantidade de relações CST. Assim, as sentenças com mais relações ocupam o topo do ranque e são selecionadas para compor o sumário até que a taxa de compressão (isto é, tamanho desejado do sumário) seja atingida e desde que não haja redundância ou contradição entre elas. Caso haja alguma relação CST que indica redundância ou contradição entre uma sentença selecionada e a próxima do ranque, a

sentença candidata é descartada. O mesmo não acontece com as sentenças que possuem relações de complementaridade. *Follow-up*, por exemplo, codifica que, em um par de sentenças (S1 e S2), S2 apresenta eventos que sucederam aos de S1. Assim, caso S1 tenha sido selecionada para o sumário, seleciona-se também S2, pois ela expressa informação complementar à de S1.

Para o português, a ferramenta CSTParser [Maziero 2012] (com acurácia de 68,13%) identifica 6 relações de conteúdo de Maziero *et al* (2010)[1] (*Elaboration, Equivalence, Subsumption, Overlap, Historical background* e *Follow-up*) com base na similaridade lexical, posto que as relações CST se estabelecem entre sentenças que possuem algum tipo de sobreposição de conteúdo [Mani 2001].

Neste artigo, apresenta-se a investigação das 3 relações CST de complementaridade (*Historical background, Follow-up* e *Elaboration*) no CSTNews [Cardoso *et al.* 2011], *corpus* multidocumento de textos jornalísticos em português. A descrição linguística manual de uma parcela dos pares do CSTNews anotados com as relações de complementaridade indicou que, além da redundância, comum às relações CST, há certas propriedades específicas que parecem caracterizar as diferentes relações de complementaridade. Tais características, uma vez validadas no restante dos pares com complementaridade do CSTNews, poderão refinar a detecção automática das relações de complemento. Dessa forma, este trabalho produziu uma descrição de um fenômeno textual de natureza semântico-discursiva até então não explorado e gerou subsídios linguísticos para o PLN.

Na Seção 2, apresentam-se a CST e o conjunto de 14 relações de Maziero *et al* (2010), com ênfase às de complementaridade. Na Seção 3, descrevem-se brevemente os trabalhos que focam a detecção automática das relações CST. Na Seção 4, apresentam-se o *corpus* CSTNews, a seleção do *subcorpus* de complementaridade e a delimitação das propriedades a serem descritas manualmente. Na Seção 5, apresenta-se o resultado da descrição dos 135 pares de sentenças do *subcorpus.* Por fim, na Seção 6, apresentam-se algumas considerações finais e trabalhos futuros.

2. As Relações CST e a Complementaridade

A CST é um modelo ou teoria que estabelece um conjunto de relações que permite conectar (em pares) unidades informativas (p.ex.: sentenças) de textos distintos que abordam um mesmo assunto, explicitando similaridades, complementaridades, contradições e variações de estilos de escrita entre elas [Radev 2000]. No trabalho de Maziero *et al.* (2010), o conjunto original de 24 relações foi reduzido a 14, como resultado da anotação manual do *corpus* CSTNews. Além disso, os autores propuseram uma tipologia para as 14 relações (Figura 1)[2].

Figura 1. Tipologia das relações CST de Maziero *et al* (2010).

Relações						
Conteúdo					Forma	
Redundância		Complemento		Contradição	Fonte/Autoria	Estilo
Total	Parcial	Temporal	Atemporal	--	--	--
Identity	*Subsumption*	*Historical*	*Elaboration*	*Contradiction**	*Citation*	*Indirect*

[1] O CSTParser não identifica as relações *Modality* e *Summary* porque o *corpus* CSTNews, usado para seu treino e teste via Aprendizado de Máquina, não possui exemplos suficientes das mesmas.
[2] O símbolo (*) indica que a relação não tem direcionalidade.

		background				*speech**
*Equivalence**	*Overlap**	*Follow-up*			*Attribution*	*Translation*
Summary					*Modality*	

Nessa tipologia, as relações CST foram organizadas em 2 grupos: (i) relações de conteúdo, as quais rotulam os relacionamentos semânticos entre sentenças, e (ii) relações de forma, que rotulam relacionamentos entre sentenças com base na forma. Cada grupo apresenta subdivisões. As relações de conteúdo podem ser classificadas nas categorias "redundância", "complemento" e "contradição".

As relações de complementaridade podem ser de dois tipos: temporal e atemporal. As temporais são *Historical background* e *Follow-up*, as quais estão definidas na Figura 2. Em (1) e (2), ilustram-se *Historical background* e *Follow-up*, respectivamente, com exemplos extraídos do CSTNews, o que é descrito na Seção 4.

Figura 2. Definição das relações CST de complementaridade temporal.

Nome da relação: *Historical background*
Direcionalidade: S1←S2
Restrição: S2 apresenta informações históricas/passadas sobre um elemento presente em S1.
Comentário: 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*
Direcionalidade: S1←S2
Restrição: 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.

(1) S1: O acidente ocorreu no delta do Nilo, ao norte de Cairo, no Egito.
S2: A maior tragédia ferroviária da história do Egito ocorreu em fevereiro de 2002, após o incêndio de um trem que cobria o trajeto entre Cairo e Luxor (sul), lotado de passageiros, e que deixou 376 mortos, segundo números oficiais.

(2) S1: Às 9 horas, a cidade tinha 113 km de lentidão, sendo que a média para o horário é de 82 km, segundo a Companhia de Engenharia de Tráfego (CET).
S2: O estado de atenção na cidade foi suspenso às 9h25.

Em (1), tem-se informação sobre um acidente ferroviário no Cairo (Egito). A relação que há entre S1 e S2 é *Historical background*, já que S2 apresenta um fato histórico ("A maior tragédia ferroviária da história do Egito ocorreu em fevereiro de 2002") relativo ao tópico principal veiculado pela S1 ("O acidente ocorreu no delta do Nilo, ao norte de Cairo, no Egito"). Em (2), o par foi anotado com a relação *Follow-up*, veiculando informações sobre um congestionamento no trânsito da cidade de São Paulo. O evento principal de S2, ou seja, "a suspensão do estado de atenção", ocorreu após ("às 9 horas") o evento veiculado por S1, isto é, "a lentidão registrada às 9 horas".

A relação de complementaridade atemporal é *Elaboration*, definida na Figura 3.

Figura 3. Definição da relação CST de complementaridade atemporal.

Nome da relação: *Elaboration*
Direcionalidade: S1←S2

Restrição: S2 detalha/refina/elabora algum elemento presente em S1, sendo que S2 não deve repetir informações presentes em S1.
Comentário: 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*.

Com base na definição, *Elaboration* não envolve localização no tempo de um acontecimento em relação a outro. As sentenças em (3) ilustram *Elaboration*.

(3) S1: Naquele horário, segundo a CET (Companhia de Engenharia de Tráfego), havia 110 km de congestionamento em toda a cidade enquanto a média para o horário era de 76 km.
S2: Na Avenida dos Bandeirantes, no sentido Marginal do Pinheiros, havia 4,2 km de lentidão, do Viaduto Arapuã até a Rua Daijiro Matsuda.

Em (3), o par de sentenças veicula informação sobre um congestionamento na cidade de São Paulo. O tópico principal de S1, que é a extensão do congestionamento ("110 km"), é detalhado pelo conteúdo de S2. No caso, S2 apresenta a informação de que 4,2 km de lentidão (dos 110 km) foi registrado em um local específico ("Avenida dos Bandeirantes"). Na próxima seção, descrevem-se brevemente os trabalhos que focam a identificação automática das relações CST, inclusive as de complementaridade.

3. Identificação Automática das Relações CST

Há vários métodos de detecção das relações CST. Para o inglês, destacam-se os de Zhang *et al.* (2003), Zhang e Radev (2005) e Kumar *et al.* (2012). Para o português, há o método que fundamenta o CSTParser de Maziero (2012).

Os métodos de Zhang *et al.* (2003) e Zhang e Radev (2005), a identificação das relações CST é feita em 2 etapas. Na primeira, verifica-se se as sentenças de um par possuem similaridade entre si, calculada pela medida estatística *word overlap*. Se sim, a segunda etapa consiste em determinar a relação CST entre elas. Para tanto, os métodos determinam: (i) quantidade de palavras idênticas (atributo lexical), (ii) quantidade de etiquetas morfossintáticas idênticas (atributo sintático), e (ii) distância semântica entre os núcleos de sintagmas nominais e verbais (atributo semântico). Quanto às relações de complementaridade, os métodos identificam somente *Follow-up* e *Elaboration*, com medida-f[3] de 0,35 e 0,18, respectivamente. Tais valores são baixos, o que pode ser justificado pelo tamanho reduzido do *corpus* utilizado para treinamento e teste do método. A medida-f mais baixa obtida *Elaboration* pode ser explicada pela natureza da própria relação, já que é mais genérica (ou menos marcada) que *Follow-up* e, por isso, mais difícil de se detectar.

Em Kumar *et al.* (2012), investigou-se a identificação de 4 relações CST do conjunto original: *Identity*, *Overlap*, *Subsumption* e *Description*. Apesar de *Description* não compor o conjunto de Maziero *et al.* (2010), ressalta-se que ela pode ser vista como uma especificação de *Elaboration*, já que ocorre quando "S1 descreve uma entidade mencionada em S2". Para identificar as relações mencionadas, utilizaram-se 4 atributos: (i) similaridade lexical, capturada pelas medidas *word overlap* e *cosseno;* (ii) tamanho das sentenças; (iii) similaridade de sintagma nominal, e (iv) similaridade de sintagma verbal. Com base em tais

[3] A medida-f é a média ponderada da precisão (isto é, número de casos corretamente detectados em relação ao número total de casos detectados) e cobertura (isto é, número de casos corretamente detectados em relação à quantidade que deveria ser detectada) [Hirschman e Mani 2003]. Precisão, cobertura e medida-f são medidas comumente utilizadas para determinar o desempenho das aplicações de PLN.

atributos, os autores desenvolveram 3 métodos, sendo que o de melhor desempenho identifica a relação *Description* com medida-f de 0,78.

O método subjacente ao CSTParser de Maziero (2012), desenvolvido para o português, identifica as relações CST de *Elaboration, Equivalence, Subsumption, Overlap, Historical background* e *Follow-up* com base em 11 atributos: (i) diferença de tamanho das sentenças em palavras, (ii) número de palavras em comum, (iii) posição das sentenças em seus respectivos textos-fonte, (iv) número de palavras na maior *substring*, (v) diferença no número de substantivos, (vi) diferença no número de advérbios, (vii) diferença no número de adjetivos, (viii) diferença no número de verbos, (ix) diferença no número de nomes próprios, (x) diferença no número de numerais e (xi) sobreposição de sinônimos. As únicas exceções são *Identity, Contradiction, Indirect Speech, Attribution, Citation* e *Translation*, que são detectadas por regras específicas.

Do exposto, observa-se que a identificação automática das relações CST tem se baseado principalmente em atributos que buscam capturar a similaridade ou redundância entre as sentenças. Isso é justificado, como mencionado, pelo fato de que as relações CST sempre ocorrem entre sentenças que são semanticamente relacionadas [Zhang e Radev 2005]. Além disso, observa-se que, mesmo codificando diferentes tipos de fenômenos multidocumento, os métodos automáticos não se baseiam em características específicas dos mesmos para a identificação das relações, sobretudo as de complementaridade. E isso pode justificar a baixa acurácia dos métodos na detecção das relações que capturam esse fenômeno.

A seguir, apresenta-se o *corpus* utilizado para a descrição das relações CST.

4. O *Corpus* CSTNews e o *Subcorpus* de Complementaridade

O fenômeno em questão foi investigado com base no CSTNews [Cardoso *et al*. 2011], *corpus* multidocumento de referência em português para a SAM. O CSTNews está organizado em 50 coleções, distribuídas nas categorias "esporte" (10), "mundo" (14), de "dinheiro" (1), "política" (10), "ciência" (1) e "cotidiano" (14). Cada coleção é composta por: (i) 2 ou 3 notícias sobre um mesmo assunto, coletadas de diferentes jornais; (ii) 5 *abstracts* multidocumento manuais e 5 extratos multidocumento manuais; (iii) sumários automáticos multidocumento, (iv) anotações linguísticas diversas, como anotação sintática dos textos-fonte, anotação dos sentidos dos substantivos e verbos nos textos-fonte, anotação de aspectos informacionais de 1 *abstract* multidocumento manual, anotação discursiva de cada texto-fonte, anotação de subtópicos dos texto-fonte e a interconexão dos textos-fonte via CST. Tendo em vista os objetivos deste trabalho, selecionaram-se os pares cujas sentenças haviam sido anotadas com as relações de complementaridade, o que resultou em um total de 713 pares, sendo: (i) 343 pares de *Elaboration*, (ii) 293 pares de *Follow-up* e (iii) 77 pares de *Historical background*. Desse total, delimitou-se um *corpus* de estudo de aproximadamente 20%, resultando em um conjunto composto por 135 pares, sendo: (i) 45 pares anotados com a relação *Elaboration*, (ii) 45 pares de *Historical background* e (iii) 45 pares *Follow-up*.

5. Seleção e Descrição dos Atributos para Caracterização das relações CST

Objetivando identificar as propriedades comuns às 3 relações, partiu-se da afirmação empírica registrada na literatura de que as relações CST ocorrem entre sentenças com certa sobreposição de conteúdo. Assim, para verificar se, de fato, a redundância define as relações de complementaridade, verificou-se a ocorrência de 3 atributos nos 135 pares: (i) similaridade lexical, (ii) localização e (iii) sobreposição de subtópico.

A similaridade lexical foi capturada pela medida *noun overlap* (Nol), bastante eficiente porque os nomes são frequentes na constituição das sentenças e carregam a maior carga semântica das mesmas [Souza *et al*. 2012]. A medida Nol de um par de sentenças (S1 e S2) é calculada pela divisão do número total de nomes idênticos entre as sentenças pela soma do número total de nomes de cada sentença, obtendo um valor entre 0 e 1, sendo que 1 indica redundância total e 0 indica nenhuma redundância.

A localização é outro atributo eficiente para capturar a redundância. Tendo em vista a estrutura típica dos textos jornalísticos ("pirâmide invertida"), em que se tem um tópico ou *lead* (veiculado pela 1ª sentença) e detalhes sobre esse tópico (subtópicos) (expressos pelas demais sentenças) [Lage 2002], Souza *et al*. (2012) observaram que, quanto mais próximas as posições de origem de duas sentenças, maior a sobreposição de conteúdo. O cálculo da localização traduz a distância entre as posições de origem das sentenças por meio de um valor entre 0 e 1: (i) 0 indica que as sentenças ocorrem na mesma posição e, por isso, são totalmente redundantes, e (ii) 1 indica que as posições são muito distintas, havendo, portanto, redundância nula[4].

Optou-se também por verificar a redundância em função de um atributo profundo: a sobreposição de subtópico. Essa sobreposição está relacionada ao atributo localização, pois, se as sentenças com posições idênticas são redundantes, isso significa que tais sentenças veiculam o mesmo conteúdo, que pode ser capturado pelo subtópico. Para verificar a redundância com base em subtópico, utilizou-se a anotação manual de subtópico disponível no CSTNews [Cardoso *et al*. 2012]. Assim, dado um par de sentenças complementares, recuperou-se do CSTNews o subtópico veiculado por cada uma e verificou-se a sobreposição entre eles.

Os 135 pares do *subcorpus* também foram descritos em função de alguns atributos potencialmente relevantes para a caracterização de cada uma das relações CST. Tendo em vista que as relações *Historical Background* e *Follow-up*, segundo Maziero *et al.* (2010), caracterizam-se pela localização da informação complementar no tempo, antes ou depois do evento de referência, os 135 pares foram descritos em função de mecanismos linguísticos por meio dos quais essa localização temporal poderia se manifestar: (i) ocorrência de advérbio de tempo (p. ex.: "hoje" e "amanhã") (em S1 e S2) e (ii) ocorrência de expressões temporais (p.ex.: "em 1996) (em S1 e S2). A verificação da ocorrência de expressões temporais, em especial, foi feita com base na anotação de tais expressões já disponível no CSTNews [Menezes Filho e Pardo 2011].

A relação *Elaboration,* segundo a definição de Maziero *et al.* (2010), parece não ser caracterizada pela presença de marcas linguísticas explicitas na superfície textual. De forma exploratória, optou-se por verificar se a ocorrência de marcadores discursivos nas sentenças que compõem os pares pode indicar algo sobre a complementaridade atemporal. Assim, dado um par, verificou-se a ocorrência de tais marcadores em S1 e S2 com base na lista de marcadores discursivos de Mazeiro e Pardo (2010).

Ao final, os 135 pares do *subcorpus* foram descritos em função de 9 atributos: (i) distância, (ii) sobreposição de nome (Nol), (iii) advérbio em S1, (iv) advérbio em S2, (v) expressão temporal em S1, (vi) expressão temporal em S2, (vii) sobreposição de subtópico, (viii) marcador discursivo em S1, (ix) e marcador discursivo em S2. A seguir, apresentam-se os resultados da descrição desses 9 atributos nos 135 pares.

[4] O atributo "distância" de cada par foi normalizado porque os textos têm tamanhos diferentes. A normalização foi feita em função da maior distância entre sentenças do respectivo *cluster* do par.

6. Resultados

Na Tabela 1, tem-se o resultado da descrição dos 135 pares em função dos 9 atributos da seção anterior. A Tabela 1 expressa o número de pares que obtiveram valores iguais ou superiores à média simples de cada atributo. Por exemplo, a distância média dos 45 pares com *Follow-up* foi de 0,14, sendo que 19 dos 45 pares estão acima dessa média.

Tabela 1. Ocorrência dos atributos no *corpus* de estudo.

Atributo	**Tipo/Relação CST**		
	Temporal		**Atemporal**
	Follow-up	***Historical background***	***Elaboration***
Distância	19/45	23/45	20/45
Similaridade lexical (Nol)	22/45	24/45	27/45
Sobreposição de subtópico	21/45	10/45	22/45
Advérbio em S1	0/45	8/45	7/45
Advérbio em S2	6/45	11/45	5/45
Expressão temporal em S1	16/45	22/45	8/45
Expressão temporal em S2	23/45	31/45	17/45
Marcador discursivo em S1	7/45	2/45	8/45
Marcador discursivo em S2	5/45	6/45	3/45

Com base na Tabela 1, tecem-se as seguintes observações as relações CST:

a) Não há distinção entre as relações de complementaridade quanto à redundância capturada pela "similaridade lexical" e "distância", pois os três subconjuntos de 45 pares apresentam comportamento similar no que diz respeito a esses atributos.
b) *Historical background* se distinguir de *Follow-up* e *Elaboration* quanto ao subtópico, pois, a sobreposição de subtópico foi registrado em apenas 10 dos 45 pares com *Historical background* e em quase metade dos casos com *Follow-up* (21/45) e *Elaboration* (22/45). Assim, parece que o evento histórico veiculado pela S2 de um par com *Historical background* se caracteriza como conteúdo (subtópico) distinto do expresso em S1.
c) As relações *Elaboration* e *Historical background* se caracterizam pela baixa ocorrência de advérbios em S1 (8/45 e 7/45, respectivamente). A relação *Follow-up* se caracteriza pela não ocorrência de advérbio em S1 (0/45).
d) As 3 relações CST de complementaridade se caracterizam por apresentarem baixa ocorrência de advérbios em S2. Apesar de a relação *Historical background* possuir frequência um pouco mais alta (11/45), isso não é suficiente para afirmar que esse atributo caracteriza essa relação.
e) *Historical background* se caracterizar pela ocorrência frequente de expressões temporais em S1 (22/45) e *Elaboration* pela baixa ocorrência (8/45).
f) Não há distinção entre *Follow-up* (23/45), *Historical background* (31/45) e *Elaboration* (17/45) quanto à ocorrência de expressões temporais em S2, já que a frequência deste é similar.
g) Não há distinção entre as relações CST de complementaridade quanto à ocorrência de marcadores discursivos em S1 e S2; as 3 relações se caracterizam de forma similar pela ocorrência pouco frequente desses marcadores.

7. Considerações finais e trabalhos futuros

Com base no estudo preliminar ora descrito, identificaram-se propriedades ou atributos comuns às 3 relações CST de complementaridade e específicos a cada uma delas. Na sequência, pretende-se validar os atributos de maior relevância no restante do *subcorpus* ou em uma parcela dele. Ademais, pretende-se submeter *subcorpus,* devidamente descrito em função dos atributos de destaque, a algoritmos de Aprendizagem de Máquinas, cujas regras aprendidas poderão subsidiar a detecção automática da complementaridade.

Referências

Cardoso, P.C.F. *et al.* (2012) "Anotação de subtópicos do córpus multidocumento CSTNews". Série de Relatórios Técnicos do ICMC, Universidade de São Paulo, n. 389. NILC-TR-12-07. São Carlos-SP, Junho, 18p.

Cardoso, P.C.F. *et al.* (2011) "CSTNews - A discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese". In: Proceedings of the 3rd RST Brazilian Meeting, pp. 88-105. Cuiabá/MT, Brasil.

Hirschman, L.; Mani, I. (2003). "Evaluation". In: Mitkov, R. (ed.). Handbook of Computational Linguistics, Oxford University Press, pp. 415-429.

Kumar, Y.J.; Salim, N. (2012) Automatic multi-document summarization approaches. *Journal of Computer. Science*, 8, p. 133-140.

Kumar,Y.J.; Salim,N.; Raza, B. (2012) Cross-document structural relationship identification using supervised machine learning. *Applied Soft Computing*, 12, p.3124-31.

Lage, N. Estrutura da notícia. Ática, 1987.

Mani, I. (2001). "Automatic Summarization". John Benjamins Publishing Co., Amsterdam.

Maziero, E.G. (2012) "Identificação automática de relações multidocumento". Tese de Mestrado. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos-SP.

Maziero, E.G.; Jorge, M.L.C.; Pardo, T.A.S. (2010) "Identifying multi-document relations". In: Proceedings of the 7th International Workshop on Natural Language Processing and Cognitive Science. Madeira/Portugal, 2010, p.60-69.

Mazeiro, E.G.; Pardo. T.A.S. (2010) "DiZer 2.0: a Web Interface for Discourse Parsing". In: Extended Activities Proceedings of the 9th PROPOR. Porto Alegre/RS, Brazil.

Menezes Filho, L.A. Pardo, T.A.S. (2011) "Detecção de Expressões Temporais no Contexto de Sumarização Automática". In: Proceedings of the 2nd STIL Student Workshop on Information and Human Language Technology, pp. 1-3. 24 a 25 de Outubro, Cuiabá/MT, Brasil.

Radev, D.R. (2000) "A common theory of information fusion from multiple text sources step one: cross-document structure". In: Proceedings of the 1st SIGdial Workshop on Discourse and Dialogue, v. 10, p. 74-83.

Souza, J.W.C.; Di-Felippo, A.; Pardo, T.A.S. (2012) "Investigação de métodos de identificação de redundância para Sumarização Automática Multidocumento". Série de Relatórios do NILC. NILC-TR-12. São Carlos-SP. Outubro, 30p.

Zhang, Z.; Otterbacher, J.; Radev, D (2003). Learning cross-document structural relationships using boosting. In the Proceedings of the 12th CIKM, New Orleans.

Zhang, Z. Radev, D.R. (2005) "Combining labeled and unlabeled data for learning cross-document structural relationships". In: Natural Language Processing – I JCNLP 2004. Springer. p. 32-41.