Em Busca de Métodos de Detecção da Complementaridade para a Sumarização Automática Multidocumento

Jackson Wilke da Cruz Souza[1,2], **Ariani Di Felippo**[1,2]

[1]Núcleo Interinstitucional de Linguística Computacional – NILC
[2]Departamento de Letras – Universidade Federal de São Carlos – UFSCar
Caixa Postal 676 CEP 13565-905 – São Carlos – SP – Brazil

{jackcruzsouza, arianidf}@gmail.com

Abstract. *We present a proposal for identifying methods for the detection of complementarity between two sentences from topically related documents. Specifically, we intend to specify, based on corpus analysis, linguistic attributes by which the complementarity can be automatically identified in the scenario of Multi-document Summarization.*

Resumo. *Apresenta-se uma proposta para identificar métodos de detecção da complementaridade entre duas sentenças provenientes de textos distintos que abordam mesmo assunto. Objetiva-se especificamente especificar, com base em corpus, atributos linguísticos por meio dos quais a complementaridade possa ser automaticamente identificada.*

1. Introdução

Duas sentenças provenientes de textos distintos que abordam um mesmo assunto podem ser semelhantes ou diferentes de diversas maneiras. O estudo dessas semelhanças e diferenças interdocumentos levou à identificação de relações retóricas como as da *Cross-document Structure Theory* (CST) [Radev 2000].

No Quadro 1, por exemplo, há um par formado por sentenças extraídas de duas notícias jornalísticas que relatam um mesmo acidente aéreo. Entre elas, identifica-se a relação CST de *Subsumption*, já que a sentença 2 apresenta, além do conteúdo da sentença 1, informações adicionais sobre a localização geográfica ("localidade de Bukavu" e "no leste") e temporal ("quinta-feira à tarde") do acidente.

Quadro 1. Exemplo de relação CST.

Subsumption
S1: Ao menos 17 pessoas morreram após a queda de um avião de passageiros na República Democrática do Congo.
S2: Um acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo (RDC), matou 17 pessoas na quinta-feira à tarde, informou nesta sextafeira um porta-voz das Nações Unidas.

Anais da III Jornada de Descrição do Português, páginas 15–23, Fortaleza, CE, Brasil, Outubro 21–23, 2013.

Maziero *et al.* (2010), em especial, propuseram uma tipologia para um conjunto refinado de relações CST que evidencia o fato de que tais relações capturam diferentes fenômenos multidocumento de conteúdo (redundância, complementaridade e contradição) e forma (forma/autoria e estilo). Essa tipologia é ilustrada pela Figura 1.

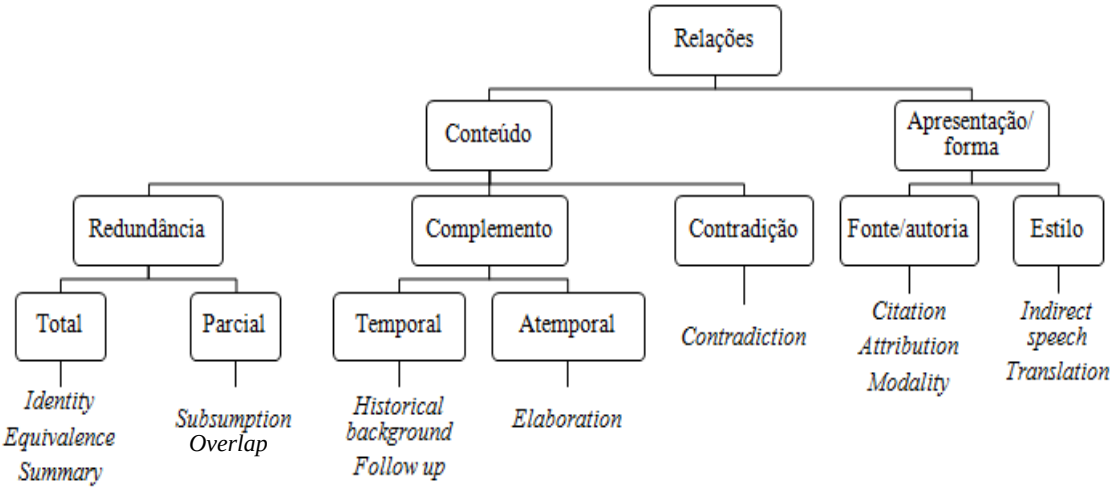

Figura 1. Tipologia das relações CST [Maziero *et al*, 2010].

Ao permitir a conexão de diferentes documentos por meio de rótulos que evidenciam os fenômenos multidocumento, a CST tem subsidiado inúmeros métodos de Sumarização Automática Multidocumento (SAM) [Radev e McKeown, 1998; Zhang *et al.*, 2002; Afantenos *et al*., 2004, 2008; Castro Jorge, 2010; Castro Jorge e Pardo, 2010], etc.

Para a utilização da CST na SAM, é preciso identificar automaticamente as relações previstas pelo modelo, como enfatizam, por exemplo, Maziero (2011), Kumar *et al.* (2012a) e Kumar *et al.* (2012b). Atualmente, essa identificação é feita exclusivamente com base em um conjunto de atributos linguísticos que capturam similaridade de conteúdo (ou redundância) entre duas sentenças, como a sobreposição de palavras idênticas. Para o português, Maziero (2011) desenvolveu um analisador semântico-discursivo que, com base nos atributos mais difundidos da literatura, distingue corretamente a complementaridade da redundância em 64% dos casos e distingue as relações temporais da única relação atemporal, *Elaboration*, em 60% dos casos.

Buscando melhorar a identificação automática da complementaridade, propõem-se dois objetivos: (i) refinar o conjunto de atributos que capturam a similaridade entre duas sentenças e (ii) descrever e analisar a complementaridade com base em *corpus* para identificar atributos linguísticos específicos que caracterizam esse fenômeno.

Tais objetivos pautam-se nas seguintes hipóteses: (i) atributos linguísticos subjacentes a métodos de detecção da redundância são pertinentes para a identificação da complementaridade, já que há certa sobreposição de conteúdo entre 2 sentenças em relação complementar; (ii) a complementaridade se manifesta na superfície linguística de tal forma que essa manifestação, uma vez "traduzida" para atributos, subsidia métodos de detecção automática desse fenômeno; (iii) atributos linguísticos que

caracterizam a redundância e a complementaridade capturam os diferentes tipos de complemento (temporais e atemporais); (iv) atributos linguísticos que caracterizam a redundância e a complementaridade capturam as diferentes relações CST que expressam complemento (*Historical background, Follow-up* e *Elaboration*), (v) o refinamento dos atributos que capturam a redundância e a especificação de atributos específicos da complementaridade permitem a detecção automática das relações CST de complemento em português com maior precisão.

Para apresentar a proposta, divide-se este artigo em 5 Seções. Na Seção 2, apresenta-se uma revisão sobre a complementaridade e as estratégias automáticas para identificar esse fenômeno. Na Seção 3, apresenta-se o *corpus* a ser utilizado. Na Seção 4, apresentam-se as etapas metodológicas para a realização da pesquisa. Na Seção 5, algumas considerações finais são feitas.

2. O fenômeno da complementaridade e sua identificação na SAM

Na tipologia de Maziero *et al.* (2010), observa-se que as relações CST estão organizadas em dois grandes grupos: relações de conteúdo e relações de forma, sendo que cada grupo apresenta subdivisões. As relações de conteúdo podem ser classificadas nas categorias "redundância", "complemento" e "contradição".

As relações da categoria "complemento", em especial, podem ser temporais ou atemporais. As relações de complementaridade temporal podem ser de dois tipos diferentes, *Historical background* e *Follow-up*. Dado um par de sentenças, S1 e S2, a relação entre elas é *Historical background* quando S2 apresenta informações históricas e/ou passadas sobre algum elemento presente em S1. A relação *Follow-up*, por sua vez, é identificada quando S2 apresenta acontecimentos/ eventos que sucederam os acontecimentos/ eventos presentes em S1; os acontecimentos em S1 e em S2 devem ser relacionados e ter um espaço de tempo relativamente curto entre si. No Quadro 1, ilustram-se os tipos de complementaridade temporal.

Quadro 1. Complementaridade temporal

Complementaridade temporal	Exemplos
***Historical background*:** S2 apresenta informações históricas/ passadas sobre algum elemento presente em S1 (S1← S2)	S1: Um **acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo (RDC)**, matou 17 pessoas na quinta-feira à tarde, informou nesta sexta-feira um porta-voz das Nações Unidas.
	S2: **Acidentes aéreos** <u>são frequentes</u> **no Congo**, onde 51 companhias privadas operam com aviões antigos principalmente fabricados na antiga União Soviética.
***Follow-up*:** S2 apresenta acontecimentos/ eventos que sucederam os acontecimentos/ eventos presentes em S1 (S1← S2)	S1: **A pista auxiliar de Congonhas abriu** às 6h, apenas para decolagens.
	S2: **Congonhas só abriu** <u>para pousos, às 8h50</u>.

A complementaridade temporal codificada pela relação CST *Historical background* é ilustrada no Quadro 1 por um par de sentenças provenientes de textos que relatam "um acidente aéreo em Bukavu, no Congo". Cada sentença é originária de um texto distinto da coleção. As sentenças do par estabelecem relação de *Historical background* porque S1 e S2 apresentam conteúdo comum (isto é, acidente aéreo no Congo), sendo que S2 apresenta uma informação adicional (histórica) sobre esse conteúdo que, no caso, diz respeito à "ocorrência frequente de acidentes aéreos no Congo (por causa do uso de aviões velhos)". O conteúdo em comum entre as sentenças dos exemplos está negritado e o trecho de S2 que indica a informação suplementar está sublinhado.

A complementaridade temporal codificada pela relação CST *Follow-up* é ilustrada no Quadro 1 por um par de sentenças advindas de uma coleção cujos textos relatam "atrasos e cancelamentos no aeroporto de Congonhas devido ao mau tempo". As sentenças estão em relação de *Follow-up* porque S1 e S2 apresentam informação comum (isto é, abertura das pistas do aeroporto de Congonhas), sendo que S2 apresenta um acontecimento que sucedeu ao evento descrito em S1 após um intervalo curto de tempo. No caso, S2 fornece "o horário de abertura da pista (principal) para pouso", que foi posterior ao evento de "abertura da pista auxiliar para decolagem" veiculado por S1. A relação de *Follow-up* entre o evento focalizado em S2 e o evento descrito em S1 envolve a ocorrência de "expressões temporais" que, segundo Baptista *et al.* (2008), são do tipo "tempo_calendário" e subtipo "data" ("6h" e "8h50").

As relações de complementaridade atemporal, não envolvem conteúdo que indica a localização no tempo (anterior ou posterior) de um acontecimento/fato em relação a outro. Essa complementaridade, codificada pela relação CST *Elaboration*, estabelece-se quando, dado um par de sentenças S1 e S2, S2 detalha/refina/elabora algum elemento presente em S1, sendo que S2 não deve repetir informações presentes em S1. Além disso, o elemento elaborado em S2 deve ser o foco de S2. Os exemplos do Quadro 2, retirado do CSTNews, ilustram esse tipo de relação de conteúdo atemporal.

Quadro 2. Complementaridade atemporal

Complementaridade atemporal	**Exemplos**
S1 e S2, S2 detalha/refina/elabora algum elemento presente em S1, sendo que S2 não deve repetir informações presentes em S1 (S1← S2)	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.
	S1: As vítimas do acidente foram 14 passageiros e três **membros da tripulação**.
	S2: Segundo fontes aeroportuárias, os **membros da tripulação** eram de nacionalidade russa.

No primeiro par de sentenças do Quadro 2, sendo cada uma delas proveniente de um texto distinto que comunica a "reforma da pista principal do aeroporto de Congonhas", observa-se que S1 e S2 possuem conteúdo comum ("cronograma da obra"), sendo que

S2 fornece uma informação adicional sobre esse conteúdo. A informação adicional em relação a S1 é o foco de S2 e consiste em “a razão pela qual o cronograma da obra não foi divulgado” (“dependente de estudos finais que estão sendo realizados pela Infraero”).

No segundo par de sentenças do Quadro 2, S1 e S2 também possuem conteúdo comum (“membros da tripulação”), sendo que S2 fornece uma informação adicional sobre os “membros da tripulação”. A informação adicional em relação a S1, que é o foco de S2, diz respeito à “nacionalidade dos membros da tripulação”, como pode ser visto no trecho sublinhado de S2 (“eram de nacionalidade russa”).

Assim, observa-se que as informações adicionais envolvidas na complementaridade são bastante variadas (“motivo/razão” e “nacionalidade”). No que se refere à realização linguística, a informação adicional em ambos os exemplos está expressa por meio de sintagmas verbais compostos por verbo (“depende”/”eram”) e sintagma preposicional (“de estudos finais que estão sendo realizados pela Infraero”/ “de nacionalidade russa”).

Na Sumarização Automática Multidocumento (SAM), subárea do Processamento Automático de Língua Natural (PLN), automatiza-se a produção de *sumários* (ou resumos) coerentes e coesos a partir de coleções de textos, provenientes de fontes distintas, que abordam um mesmo assunto [Mani 2001]. A SAM tem sido motivada pela enorme quantidade de informação disponível na *web* e o pouco tempo que as pessoas têm para assimilá-la.

A SAM monolíngue consiste em gerar um sumário em uma língua *x* a partir de dois ou mais textos-fonte na língua *x*. Na maioria dos trabalhos, os textos-fonte são notícias jornalísticas e os sumários são extratos informativos e genéricos, ou seja, compostos por sentenças extraídas integralmente dos textos-fonte. Idealmente, as sentenças extraídas dos textos-fonte veiculam o conteúdo central da coleção e, por isso, a leitura dos sumários substitui a dos textos-fonte [Mani 2001].

Para a geração de extratos informativos e genéricos a partir de textos jornalísticos, os métodos de SAM precisam realizar duas tarefas centrais: (i) identificar o conteúdo principal da coleção e (ii) identificar a redundância, complementaridade e contradição (os chamados fenômenos multidocumento) entre as sentenças dos diferentes textos-fonte. A identificação desses fenômenos justifica-se pelo fato de que um sumário deve veicular a informação principal da coleção sem redundância e contradição, mas com complementaridade.

Em alguns métodos de SAM, a identificação dos fenômenos multidocumento é feita pela detecção das relações CST. Tendo em vista que a anotação manual das referidas relações é tarefa demorada e cara, tem-se focado a proposição de métodos automáticos. Atualmente, a identificação automática é feita com base em um conjunto de atributos linguísticos que capturam similaridade de conteúdo entre duas sentenças.

Para o português, Maziero (2011) desenvolveu um analisador semântico-discursivo, denominado CSTParser, que, com base nos atributos mais difundidos da literatura, identifica as relações de complementaridade (*Historical-background*, *Follow-up* e *Elaboration*), além das relações *Equivalence*, *Subsumption* e *Overlap*, com 70,51% de precisão [Maziero, Pardo, 2012]. No caso, os atributos linguísticos nos quais Maziero

(2011) se baseia para identificar as relações entre duas sentenças, S1 e S2, são: (i) diferença de tamanho em palavras (S1-S2); (ii) porcentagem de palavras em comum em S1; (iii) porcentagem de palavras em comum em S2; (iv) posição de S1 no texto (0- início, 2- fim, 1- meio); (v) número de palavras na maior substring 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 no número de verbos entre S1 e S2; (x) diferença no número de nomes próprios entre S1 e S2; (xi) diferença no número de numerais entre S1 e S2; (xii) número de sinônimos iguais em S1 e S2.

Souza *et al.* (2012), por sua vez, com base em um conjunto semelhante de atributos obteve 97,7% de precisão na identificação dos diferentes níveis de redundância expressos pelas relações CST (total, parcial e neutro) e 88,8% na identificação das relações de redundância (*Identity*, *Equivalence* e *Summary*, *Subsumption* e *Overlap*). Para tanto, os autores utilizam um conjunto de atributos linguísticos e estatísticos, a saber: (i) sobreposição de padrões morfossintáticos, (ii) sobreposição de verbo principal, (iii) sobreposição de núcleo de sujeito (Suj), (iv) sobreposição de núcleo de objeto/predicativo principal e (v) sobreposição de etiquetas morfossintáticas; (vi) *word overlap* (Wol), (vii) *noun overlap* (Nol) e (viii) *verb overlap* (Vol). Além dos métodos advindos da literatura, especificou-se outro, de natureza estrutural, segundo o qual quanto menor a distância entre as posições que as sentenças ocupam em seus respectivos textos-fonte, maior a redundância entre elas. A esse método, deu-se a denominação de Loc.

Maziero (2011) e Souza *et al.* (2012), aliás, utilizaram o mesmo *corpus* em suas pesquisas, o CSTNews, cuja descrição é feita na próxima seção. Souza *et al.* (2012), no entanto, fez um recorte no *corpus*, aprendendo e testando seus métodos em um conjunto menor de dados.

Dessa forma, vê-se que as relações CST, em especial as de complementaridade, são identificadas basicamente em função de atributos que capturam a similaridade entre duas sentenças. Para a identificação dessas relações, não se tem atributos disponíveis que possam subsidiar especificamente a identificação automática da complementaridade, posto que fenômeno multidocumento ainda não foi investigado com sistemática. Com exceção da redundância, fenômeno mais explorado, tem-se conhecimento do trabalho inicial de Mazeiro (2011), a partir do qual algumas regras pontuais foram propostas para a identificação apenas da contradição.

Buscando melhorar a identificação automática da complementaridade, propõem-se dois objetivos: (i) refinar o conjunto de atributos que capturam a similaridade entre duas sentenças e (ii) descrever e analisar a complementaridade com base em *corpus* para identificar atributos linguísticos específicos que caracterizam esse fenômeno. O *corpus* a ser utilizado está descrito na sequência.

3. O *corpus* CSTNews

Para as pesquisas sobre SAM envolvendo o português, construiu-se o *corpus* multidocumento denominado CSTNews [Cardoso *et al.*, 2011], que é composto por 50 coleções de textos. Cada coleção engloba de 2 a 3 textos, cada um deles produzido por uma agência distinta (*Folha de São Paulo*, *Estadão*, *O Globo*, *Gazeta do Povo* e *Jornal do Brasil*), sobre uma mesma notícia. Ao todo, tem-se 14 coleções de textos compilados

da seção “mundo”, 14 coleções de “cotidiano”, 10 de “política”, 10 de “esporte”, 1 coleção de dinheiro e 1 coleção de “ciência”.

Os textos-fonte de cada coleção do CSTNews foram manualmente conectados em nível sentencial por meio de relações semântico-discursivas. Essa conexão ou indexação é tida como uma espécie de anotação de *corpus*. Na Figura 1, ilustra-se o esquema genérico de relacionamento das sentenças entre os textos-fonte, retirado de Maziero (2012).

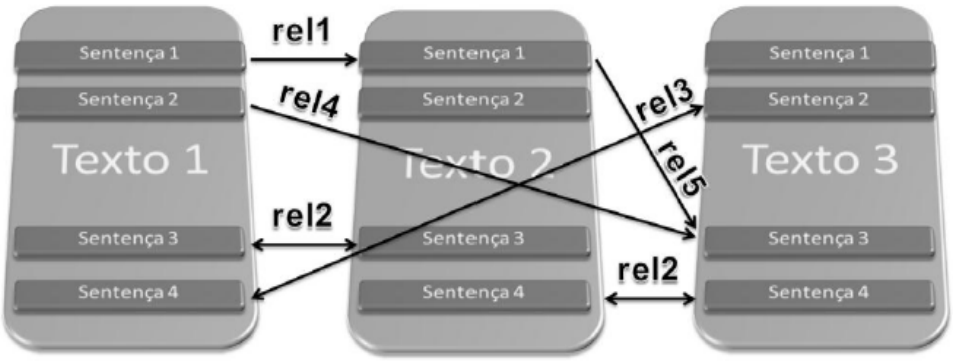

Figura 1: Esquema genérico de análise multidocumento.

No CSTNews, aliás, a complementaridade é o fenômeno mais recorrente, já que 49% delas são do subtipo “complemento”, como indicado no Quadro 3.

Quadro 3. Representatividade dos tipos de relações CST no *corpus* CSTNews

Tipo de relação CST	Representatividade
Complementaridade	49%
Redundância	43%
Fonte/Autoria	4%
Contradição	3%
Estilo	1%

4. Etapas metodológicas

Para alcançar os objetivos traçados, foram especificadas 6 tarefas metodológicas, a saber: (i) recorte e análise de *corpus*, (ii) caracterização de *corpus*, (iii) aplicação de métodos de identificação da complementaridade, (iv) estudo da correlação entre os métodos e os tipos CST, (v) estudo da correlação entre os métodos e as relações CST e (vi) avaliação dos métodos.

- **Tarefa 1:** Recorte e análise do *corpus*: consistiu em um recorte no *corpus* CSTNews, caracterizado pela seleção manual de todos os pares formados por sentenças conectadas pelas relações de complementaridade temporal e atemporal, como consta no Quadro 4.

Quadro 4. Características do *subcorpus* do CSTNews

Tipo de complementaridade	Relação CST	Quantidade de pares
Temporal	*Historical background*	77
	Follow-up	293
Atemporal	*Elaboration*	343

Uma vez recortado, o conjunto de sentenças que compõe o *subcorpus* de complementaridade do CSTNews será analisado com o objetivo de delimitar atributos que possam subsidiar a detecção da complementaridade. Buscar-se-á delimitar um conjunto formado por atributos de detecção da redundância (p.ex.: Souza *et al.* (2012)) e atributos específicos que caracterizam a complementaridade temporal e atemporal. Para a detecção da complementaridade temporal, por exemplo, a ocorrência de expressões temporais parece ser um atributo relevante.

- **Tarefa 2:** Caracterização de *corpus*: consiste na descrição manual de cada uma das sentenças do *subcorpus* em função dos atributos identificados na tarefa anterior.
- **Tarefa 3:** Aplicação de métodos de detecção da complementaridade: consiste na especificação de métodos baseados nos atributos identificados na tarefa anterior e na subsequente aplicação dos mesmos às sentenças caracterizadas.
- **Tarefa 4:** Estudo da correlação entre os métodos e os tipos de relação CST: consiste no estudo da correlação entre os métodos delimitados na tarefa anterior e os tipos de complementaridade (temporais e atemporais). Com base nesse estudo, pretende-se identificar os métodos que expressam mais adequadamente tais diferenças de complemento.
- **Tarefa 5:** Estudo da correlação entre os métodos e as relações CST: consiste o estudo da correlação entre os métodos delimitados na tarefa anterior e as relações CST de *Historical background, Follow-up* e *Elaboration*. Com base nesse estudo, pretende-se identificar os métodos que expressam mais adequadamente essas relações.
- **Tarefa 6:** Avaliação: nessa tarefa, os métodos mais eficientes serão aplicados à parcela do *subcorpus* destinado ao teste. Além de testar os métodos em um conjunto distinto de dados, essa tarefa deve englobar a geração automática de sumários multidocumento com base nesses métodos de detecção da complementaridade e posterior comparação dos mesmos com sumários manuais que compõem originalmente o CSTNews.

4. Considerações Finais

O trabalho está em estágio inicial, concentrando-se nas tarefas de revisão bibliográfica e levantamento e seleção de recursos e ferramentas de SAM. Ao final, espera-se identificar estratégias que possam efetivamente subsidiar a sumarização automática multidocumento, em PB, ao que se refere à identificação automática do fenômeno da complementaridade.

Agradecimentos

Os autores agradecem ao CNPq e à FAPESP pelo apoio financeiro.

Referências Bibliográficas

Baptista, J.; Hagège, C. e Mamede, N. (2008). Identificação, classificação e normalização de expressões temporais do português: A experiência do segundo HAREM e o futuro. In: Mota, C. e Santos, D. (Eds.), Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: o segundo HAREM. Linguateca

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

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

Maziero, E.G. and Pardo, T.A.S. (2012). Automatic Identification of Multi-document Relations. In the (on-line) Proceedings of the PROPOR 2012 PhD and MSc/MA Dissertation Contest, p. 1-8. April 17-20, Coimbra, Portugal.

Radev, D. R. (2000). A common theory of information fusion from multiple text sources, step one: cross-document structure. In the Proceedings of the ACL Sigdial Workshop on Discourse and Dialogue, Hong Kong. p. 74-86.

Souza, J.W.C.; Di-Felippo, A. e Pardo, T.A.S. Investigação do fenômeno da redundância na Sumarização Automática Multidocumento. Série de Relatórios Técnicos do NILC, NILC-TR-12-03, Outubro, 2012. 31p.