Violações linguísticas em referências a entidades do tipo "pessoa" em extratos automáticos multidocumento

Luana Fonseca Cristini[1,2] and Ariani Di-Felippo[1,3]

[1] Interinstitutional Center for Computational Linguistics (NILC), São Carlos/SP, Brazil
[2] College of Letters and Sciences (FCL), São Paulo State University (UNESP)
Rodovia Araraquara-Jaú Km 1, Araraquara, 14800-901, Brazil
[3] Language and Literature Department (DL), Federal University of São Carlos (UFSCar)
Rodovia Washington Luís, km 235 - SP 310, São Carlos, 13565-905, Brazil
{luanafcristini;arianidf}@gmail.com

Abstract. *We present the typification of linguistic violations' in references to people occurred in multi-document summaries generated by the RSumm and GistSumm summarizers based on the CSTNews corpus. This task allows us to evaluate the impact of the rewrite rules on the overall quality of automatic summaries and to develop automatic methods of violations detection.*

Resumo. *Descreve-se a tipificação de violações linguísticas em referências a "pessoa" ocorridas em sumários multidocumento gerados pelos sistemas RSumm e GistSumm a partir do corpus CSTNews. Essa tarefa permite avaliar o impacto da reescrita de referências na qualidade dos sumários automáticos e na criação de métodos automáticos de detecção das violações.*

1. Introdução

Os sumarizadores automáticos multidocumento comumente geram um sumário a partir de uma coleção de notícias que tratam de um mesmo assunto [Mani 2001]. Nos métodos extrativos, os sumários (extratos) são compostos pela justaposição das sentenças mais centrais da coleção extraídas integralmente dos textos-fonte, gerando vários problemas de coesão e coerência [Nenkova e McKeown, 2011].

Alguns deles ocorrem no nível das entidades nomeadas: (i) primeira menção sem explicação (1M-EXP), (ii) menção subsequente com explicação (nM+EXP), (iii) acrônimo sem explicação (ACR-EXP), (iv) sintagma nominal (SN) definido com referência a menção anterior (SNdef-REF), (vi) SN indefinido com referência a menção anterior (SNind+REF), (v) pronome sem antecedente (PRO-ANT) e (iv) pronome com antecedente enganoso (PRO-ENG) [Kaspersson *et al.* 2012, Friedrich *et al.* 2014 e Dias 2016].

Neste artigo, apresenta-se a tipificação das violações linguísticas específicas das referências ou menções a entidades do tipo "pessoa" que ocorrem em extratos automáticos multidocumento em português. Tal tarefa pode contribuir para: (i) avaliação do impacto da reescrita de menções na informatividade e na qualidade linguística dos extratos, posto que a reescrita tem se mostrado uma alternativa de pós-edição bastante viável [Nenkova e McKeown 2003a,b e Siddharthan *et al.* 2011] e (ii) na criação de métodos automáticos de detecção das violações em menções desse tipo.

2. Tipificação das Violações nos Extratos Automáticos

Neste trabalho, utilizaram-se dois sumarizadores: GistSumm [Pardo 2005] e RSumm RSumm [Ribaldo *et al* 2012, 2016]. O GistSumm [Pardo 2005] é um sistema superficial que seleciona as sentenças para compor um extrato pela frequência das palavras das sentenças na coleção e similaridade lexical entre a sentença que possui as palavras mais frequentes (*gist sentence*) e as demais da coleção. O RSumm [Ribaldo *et al* 2012, 2016] é um sistema híbrido, pois une medidas estatísticas aplicadas a uma modelagem em grafo dos textos-fonte e informação de subtópicos. Devido à sofisticação do método, o RSumm gera extratos com maior informatividade e qualidade linguística que o GistSumm. Assim, tais sumarizadores foram escolhidos devido ao objetivo de se observar o impacto da reescrita na qualidade linguística e informatividade em extratos gerados por sistemas de desempenho bem diferentes.

Cada um dos sistemas gerou 1 extrato com aproximadamente 100 palavras para cada um dos 50 *clusters* do CSTNews [Cardoso *et al.* 2011]. Tais *clusters* são compostos por 2 ou 3 notícias em português sobre mesmo assunto, provenientes de diferentes fontes jornalísticas, e estão englobam notícias de diferentes domínios: esporte (10 *clusters*), mundo (14 *clusters*), dinheiro (1 *clusters*), política (10 *clusters*), ciência (1 *clusters*) e "cotidiano" (14 *clusters*).

As menções problemáticas nos extratos gerados pelos GistSumm e RSumm foram manualmente identificadas e tipificadas com uma anotação no seguinte formato: <e TYPE=(Error Type)>(Text Passage)</e>. Para preencher Error Type, os anotadores dispunham das 7 etiquetas de Dias (2016) (isto é, 1M-EXP, nM+EXP, ACR-EXP, SNdef-REF, SNind+REF, PRO-ANT e PRO_ENG) e de outras 6 etiquetas secundárias que foram propostas para especificar as violações que tipicamente ocorrem em primeiras menções e menções subsequentes a pessoas.

Tendo em vista que as entidades do tipo pessoa tendem a ser introduzidas nas notícias por uma menção com núcleo *full name* e um *pre-modifier* [Di-Felippo 2016], propuseram-se as 3 etiquetas secundárias *–FullName, –PreMod* e *–FullName/–PreMod* para explicitar especificamente o(s) elemento(s) ausente(s), que deveria(m) compor a estrutura prevista para as primeiras menções. Para as menções subsequentes, que tendem a ter somente um *first name* ou *noun* como núcleo [Di-Felippo 2016], as 3 etiquetas secundárias *+PreMod, +PostMod* e *+PreMod/+PostMod* foram propostas para explicitar especificamente o(s) elemento(s) que não deveria(m) estar presentes na estrutura desse tipo de menção.

Na Tabela 1, apresentam-se as combinações de etiquetas genéricas e secundárias empregadas na anotação dos 100 extratos automáticos.

Tabela 1. Etiquetas para anotação de violações em referências a "pessoa".

Violação	**Etiqueta**
Acrônimo sem explicação	ACR-EXP
SN definido sem referência a menção anterior	SNdef-REF
Primeira menção sem "explicação" (nome completo)	1M-EXP [-FullName]
Primeira menção sem "explicação" (pré-modificador)	1M-EXP [-PreMod]
Primeira menção sem "explicação" (nome completo e pré-mod)	1M-EXP [-FullName/-PreMod]
Menção subsequente com "explicação" (pré-modificador)	nM+EXP[+PreMod]
Menção subsequente com "explicação" (pós-modificador)	nM+EXP[+PostMod]
Menção subsequente com "explicação" (pré- e pós-modificador)	nM+EXP [+PreMod/+PostMod]

Assm, a primeira violação ocorrida no extrato automático da Figura 2, por exemplo, é <e TYPE=1M-EXP[-PreMod/-FullName]>Cahe</e>. Nessa anotação, explicita-se que a "primeira menção" (1M) (Cahe) "não possui explicação" (-EXP) adequada para a identificação do referente, a qual, no caso, refere-se à "ausência de pré-modificador"[1] e núcleo do tipo *full name* (-PreMod/-FullName).

Nas Tabela 2 e 3, apresentam-se, respectivamente, os resultados da anotação das violações nos extratos gerados pelo GistSumm e RSumm. Nessas tabelas, as violações estão organizadas por extrato/*cluster*.

Tabela 2. Distribuição das violações nos extratos do GistSumm.

Extrato/Cluster	Violação	Qt
C05	nM+EXP [+PreMod/+PostMod]	1
	nM+EXP [+PreMod]	1
	1M-EXP [-PreMod/-FullName]	2
	1M-EXP [-PreMod]	2
C07	1M-EXP [-PreMod/-FullName]	1
C08	1M-EXP [-PreMod/-FullName]	1
C14	nM+EXP [+PreMod]	1
C17	1M-EXP [-PreMod/-FullName]	1
C18	1M-EXP [-PreMod/-FullName]	1
C19	1M-EXP [-PreMod/-FullName]	2
C21	nM+EXP [+PreMod]	1
C24	1M-EXP [-PreMod]	1
C25	1M-EXP [-PreMod/-FullName]	1
C27	1M-EXP [-PreMod]	1
	1M-EXP [-PreMod/-FullName]	2
C28	1M-EXP [-PreMod/-FullName]	1
C31	nM+EXP [+PreMod]	1
C33	1M-EXP [-PreMod/-FullName]	1
C35	1M-EXP [-PreMod/-FullName]	1
C38	1M-EXP [-PreMod]	4
C40	1M-EXP [-PreMod/-FullName]	2
	nM+EXP [+PostMod]	1
C41	1M-EXP [-PreMod]	5
C42	SNdef-REF [-PreMod/-FullName]	1
	1M-EXP [-PreMod/-FullName]	1
C44	1M-EXP [-PreMod/-FullName]	2
C45	1M-EXP [-FullName]	1
C48	SNdef-REF [-PreMod/-FullName]	1
	1M-EXP [-PreMod/-FullName]	2
	nM+EXP [+PreMod]	1
C50	1M-EXP [-PreMod/-FullName]	1
Total		**45**

[1] Os pré-modificadores são nomes ou adjetivos que, antepostos ao núcleo da menção, informam o leitor sobre afiliação, cargo ou função exercido pela entidade "pessoa". Os pós-modificadores, pospostos ao núcleo, podem ser do tipo oposto, sintagma preposicional, sintagma adjetival ou oração relativa (ou ainda observações parentéticas).

Tabela 3. Distribuição das violações nos extratos do RSumm.

Extrato/Cluster	Violação	Qt
C02	1M-EXP [-FullName]	2
	1M-EXP [-PreMod]	1
C07	1M-EXP [-PreMod/-FullName]	1
C08	1M-EXP [-FullName]	2
	1M-EXP [-PreMod/-FullName]	1
C19	1M-EXP [-PreMod/-FullName]	2
	nM+EXP [+PostMod]	1
C21	nM+EXP [+PreMod]	1
C24	SNdef-REF [-PreMod/-FullName]	1
C25	1M-EXP [-PreMod/-FullName]	2
	1M-EXP [-PreMod]	1
C33	nM+EXP [+PreMod]	1
	SNdef-REF [-FullName]	1
C34	nM+EXP [+PreMod]	1
C35	1M-EXP [-PreMod/-FullName]	1
	nM+EXP [+PreMod/+PostMod]	1
C36	ACR-EXP	1
C38	1M-EXP [-PreMod]	4
C43	1M-EXP [-PreMod/-FullName]	1
	1M-EXP [-PreMod]	1
	nM+EXP [+PreMod]	1
C44	1M-EXP [-PreMod/-FullName]	2
C45	1M-EXP [-PreMod/-FullName]	1
C47	nM+EXP [+PreMod]	1
C48	1M-EXP [-PreMod/-FullName]	5
C50	1M-EXP [-PreMod/-FullName]	1
Total		**38**

Nas Tabela 4 e 5, apresentam-se as violações organizadas pelo tipo de menção.

Tabela 4. Distribuição das violações nas 1as menções por sistema.

Primeira menção		
Violação	*GistSumm*	*RSumm*
1M-EXP [-PreMod/-FullName]	22	17
1M-EXP [-PreMod]	13	7
1M-EXP [-FullName]	1	4
SNdef-REF [-PreMod/-FullName]	2	1
SNdef-REF [-FullName]	-	1
ACR-EXP	-	1
Total	**38**	**31**

Tabela 5. Distribuição das violações nas menções subsequentes por sistema.

Menção subsequente		
Violação	*GistSumm*	*RSumm*
nM+EXP [+PreMod]	5	5
nM+EXP [+PreMod/+PostMod]	1	1
nM+EXP [+PostMod]	1	1
Total	**7**	**7**

De acordo com as Tabelas 2 e 3, identificaram-se 45 menções problemáticas nos extratos gerados pelo GisSumm e 38 nos do RSumm. Embora haja mais casos nos extratos do GistSum em termos absolutos, a quantidade média de violações é praticamente a mesma para os dois sistemas. Ambos têm média aproximada de 2 violações por extrato, já que as 45 do GistSumm se distribuem em 23 extratos e as 38 do RSumm, em 18 extratos.

Com base nas Tabela 4 e 5, as violações nas 1as menções são mais frequentes que nas subsequentes. Tais violações resultam do fato de que a 1ª menção do extrato automático ocorreu como menção subsequente no texto-fonte do qual foi extraída, não sendo suficientemente informativa. O tipo mais frequente de violação é de 1ª menção sem núcleo *full name* e sem *pre-modifier* (1M-EXP[-PreMod/-FullName]), ilustrado na Figura 2, por exemplo, pela 1ª menção excessivamente curta a "Cahe". O segundo tipo mais frequente é de 1ª menção sem *pre-modifier* (1M-EXP[-PreMod).

Quanto às menções subsequentes, identificaram-se 7 casos de violações nos extratos do GistSumm e do RSumm. A distribuição dos casos entre os 3 tipos é a mesma nos dois conjuntos de extratos: (i) 5 casos de nM+EXP [+PreMod], (ii) 1 caso de nM+EXP[+PreMod/+PostMod], e (iii) 1 caso de nM+EXP [+PostMod]. A violação mais comum pode ser ilustrada pela menção subsequente "Maradona, 46" da Figura 2, que veicula a idade da entidade "pessoa". Esse tipo de violação resulta do fato de que a menção subsequente de um extrato automático ocorreu como 1ª menção no texto-fonte de origem.

3. Potencialidades da anotação

A anotação aqui descrita permite que se aprofunde o conhecimento sobre os problemas gerados pelos sumarizadores extrativos e que se investigue o impacto da reescrita das menções a pessoas na qualidade linguística e informatividade dos extratos automáticos.

Uma vez que as violações tenham sido anotadas e tipificadas, estas podem ser reescritas de tal maneira que atendam às preferências identificadas em sumários humanos multidocumento. Diz-se isso porque a revisão[2], entendida como um processo de pós-edição dos extratos, é uma estratégia de abstração relativamente mais barata que as demais (p.ex.: compressão sentencial e fusão de informação) reconhecidamente útil para a melhoria dos extratos automáticos [Nenkova e McKeown 2011].

Para tal reescrita das referências, destaca-se que Di-Felippo (2016), com base na descrição manual das cadeias de correferência em 50 sumários humanos (de 100 palavras) produzidos a partir dos *clusters* do CSTNews, identificou preferências quanto à forma e à sequência das menções, o que resultou em um conjunto de regras de reescrita para menções a entidades do tipo "pessoa" (Figura 1).

Tais regras, no entanto, ainda não foram testadas ou avaliadas e, para isso, a produção de versões reescritas de extratos automáticos é necessária. Uma vez que as violações nos extratos gerados pelo GistSumm e RSumm para as 50 coleções do *corpus* CTSNews foram anotadas e que Cristini e Di-Felippo (2018) já haviam anotado as cadeias de correferência de entidades do tipo pessoa nos textos-fonte do CSTNews, pode-se proceder à reescrita das referências como ilustrado na sequência.

[2] A revisão consiste em qualquer modificação realizada nos extratos como eliminação, combinação e/ou substituição de expressões e/ou sentenças.

Regra de reescrita para primeira menção (*discourse-new*) a *person*

1. IF o núcleo da referência não for *full name* THEN:
   (a) Analisar todas as primeiras menções do *input*[3] com o objetivo de identificar *full name*
   (b) IF *full name* for encontrado no *input* THEN:
      i. Reescrever a menção original por *full name*.
   (c) ELSE IF nenhum *full name* for encontrado no *input* THEN:
      i. Não reescrever o núcleo da menção.

2. IF a primeira menção não for acompanhada de *pre-modifier*
   (a) Analisar todas as primeiras menções do *input* com o objetivo de identificar *pre-modifier*
   (b) IF qualquer *pre-modifier* for encontrado no *input* THEN
      i. Inserir o *pre-modifier* mais longo no SN.
   (c) ELSE IF nenhum *pre-modifier* for no *input* THEN:
      i. Analisar todas as primeiras menções do *input* com o objetivo de identificar um pós-modificador do tipo *appositional phrase*
      ii. Selecionar o *appositional phrase* mais longo e incluí-lo no SN da primeira menção.
   (d) ELSE IF nenhum (*pre-* ou *post-*) *modifier* for encontrado no *input* THEN:
      i. Manter o núcleo já reescrito ou a menção original.

Regra de reescrita para menção subsequente (*discourse-old*) a *person*

1. IF o núcleo da referência não for *first name* ou *noun* THEN:
   (a) Analisar todas as menções do *input* com o objetivo de identificar *first name*
   (b) IF *first name* for encontrado no *input* THEN:
      i. Reescrever a menção original por *first name* e remover os *pre-* e *post-modifiers* (a não ser que seja um acrônimo parentético)
   (c) ELSE IF nenhum *first name* for encontrado no *input* THEN:
      i. Analisar todas as menções do *input* com o objetivo de identificar *noun*.
      ii. IF *noun* for encontrado no *input* THEN:
         I. Reescrever a menção original por *noun* e remover os *pre-* e *post-modifiers*
      iii. ELSE IF *noun* não foi encontrado no *input* THEN:
         I. Não reescrever o núcleo da menção.

Figura 1. Algoritmo de reescrita para referências a pessoas [Di-Felippo 2016].

Apenas para ilustrar o processo de reescrita, aplicam-se as regras de Di Felippo às menções problemáticas do extrato da Figura 2. O extrato de C19 apresenta 2 casos de 1M-EXP[-PreMod/-FullName] e 1 caso de nM+EXP[+PostMod].

<e TYPE=1M-EXP[-PreMod/-FullName]>Cahe**</e>** disse ainda que **<e TYPE=1M-EXP[-PreMod/-FullName]>**Maradona**</e>** não voltou a consumir bebidas alcoólicas e que as causas da recaída estão sendo investigadas. Maradona havia recebido alta no último dia 11, mas voltou a ser internado na sexta-feira e os boletins médicos não especificaram o que se passava com o ex-jogador - Cahe descartou pancreatite ou úlcera. Cahe disse ainda que Maradona não voltou a consumir bebidas alcoólicas e que as causas da recaída estão sendo investigadas. **<e TYPE=nM+EXP[+PostMod]>**Maradona, 46,**</e>** desenvolveu um hepatite tóxica por excesso de consumo de álcool, o que já o manteve internado durante 13 dias antes da primeira alta.

Figura 2. Erros anotados no extrato gerado pelo RSumm para C19.

[3] Entende-se *input* como a coleção de textos-fonte a ser sumarizada.

Diante dessas violações, que se referem às entidades "Alfredo Cahe" (E1) e Diego Maradona (E2), recuperam-se de forma manual todas as menções dessas entidades anotadas nos textos-fonte da coleção, juntamente com sua caracterização linguística. Na Figura 3, vê-se que, para a reescrita da 1ª menção à E1, a menção 1 (M1) do texto 1 (D1) (negrito) é a opção mais adequada, pois possui núcleo *full name* ("Alfredo Cahe") e o *pre-modifier* mais longo ("O médico pessoal do argentino Diego Maradona").

Tendo em vista que o pré-modificador da menção reescrita de E1 engloba uma primeira menção à E2 com estrutura *Pre-modifier + Full name* (ou seja, "o argentino Diego Maradona"), a primeira menção original a E2 ("Maradona") passou a ser *uma* menção subsequente. Com núcleo do tipo *last name*, essa menção, agora subsequente, não satisfaz a regra da Figura 1, tendo de ser reescrita por *first name* ou *noun.* Dessas duas opções, observa-se na Figura 3 que somente menções subsequentes com núcleo *noun* ocorrem nos textos-fonte da coleção (em negrito). Em D1, tem-se "o ex-craque" (M2) e "o ex-jogador" (M4) (em negrito). Em D2, por sua vez, ocorrem "o ex-jogador" (M4) e "o ídolo argentino" (M8) (em negrito). Ao descartar "o ídolo argentino" devido à ocorrência de pós-modificação ("argentino"), o que não é desejável em menções subsequentes, restaram duas opções, "o ex-craque" e "o ex-jogador". No caso, selecionou-se "o ex-jogador", pois, embora sendo mais longa, foi considerada mais informativa que "o ex-craque". A menção "o ex-jogador", aliás, também foi utilizada para a reescrita da menção subsequente "Maradona, 46,", cujo problema foi anotado como nM+EXP[+PostMod] devido à presença de um *pós-modificador* ("46").

Entidade/Menção/Doc	Texto da Menção	Headedness	Definiteness	PreMod	PostMod
E1_M1_D1	**O médico pessoal do argentino Diego Maradona, Alfredo Cahe,**	FullName	DefArt	Any	None
E1_M2_D1	Cahe	LastName	None	None	None
E1_M3_D1	Cahe	LastName	None	None	None
E1_M4_D1	Cahe	LastName	None	None	None
E1_M1_D2	seu médico pessoal, Alfredo Cahe	FullName	Possessive	Any	None
E1_M2_D2	o médico	Noun	DefArt	None	None
E1_M3_D2	Cahe	LastName	None	None	None
E2_M1_D1	o argentino Diego Maradona	FullName	DefArt	Any	None
E2_M2_D1	**o ex-craque**	Noun	DefArt	None	None
E2_M3_D1	Maradona	LastName	None	None	None
E2_M4_D1	**o ex-jogador**	Noun	DefArt	None	None
E2_M5_D1	Maradona	LastName	None	None	None
E2_M6_D1	Maradona, 46	LastName	None	None	Other
E2_M7_D1	Maradona	LastName	None	None	None
E2_M1_D2	Maradona	LastName	None	None	None
E2_M2_D2	ele	Pronoun	None	None	None
E2_M3_D2	ele	Pronoun	None	None	None
E2_M4_D2	**o ex-jogador**	Noun	DefArt	None	None
E2_M5_D2	Maradona	LastName	None	None	None
E2_M6_D2	ele	Pronoun	None	None	None
E2_M7_D2	Maradona	LastName	None	None	None
E2_M8_D2	**o ídolo argentino**	Noun	DefArt	None	AdjP

Figura 3. Cadeias de correferência dos textos-fonte de C19.

Ao final da aplicação das regras, tem-se a versão reescrita do extrato (Figura 4), a qual não apresenta mais os problemas destacados na Figura 2.

O médico pessoal do argentino Diego Maradona, Alfredo Cahe, disse ainda que **o ex-jogador** não voltou a consumir bebidas alcoólicas e que as causas da recaída estão sendo investigadas. Maradona havia recebido alta no último dia 11, mas voltou a ser internado na sexta-feira e os boletins médicos não especificaram o que se passava com o ex-jogador --Cahe descartou pancreatite ou úlcera. Cahe disse ainda que Maradona não voltou a consumir bebidas alcoólicas e que as causas da recaída estão sendo investigadas. **O ex-jogador** desenvolveu uma hepatite tóxica por excesso de consumo de álcool, o que já o manteve internado durante 13 dias antes da primeira alta.

Figura 4. Versão reescrita do extrato de C19 gerado pelo RSumm.

3. Considerações finais

Uma vez que se tenha gerado as versões reescritas dos extratos gerados pelo GistSumm e RSumm que apresentam violações em menções a pessoas, pretende-se avaliar o impacto da reescrita na qualidade linguística e na informatividade dos extratos automáticos. A avaliação da qualidade poderá ser feita de duas formas distintas. Uma das avaliações pode consistir na análise dos extratos automáticos (original e versão reescrita) por meio do julgamento humano quanto aos 5 parâmetros proposta na *Document Understanding Conference* (DUC) de 2005 (DANG, 2005): (i) gramaticalidade, (ii) não-redundância, (iii) clareza referencial, (iv) foco (temático), e (v) estrutura/coerência. Na outra avaliação, pretende-se aplicar o mesmo procedimento realizado por Siddharthan *et al*. (2011), que consistiu no julgamento das versões reescritas em comparação às suas versões originais. No caso, um extrato automático original e a sua respectiva versão com as referências reescritas são submetidos à avaliação de um humano. Quanto à avaliação do impacto das reescritas na informatividade dos extratos automáticos multidocumento, poder-se-á utilizar o traducional pacote de medida ROUGE [Lin 2004], que calcula a informatividade pela coocorrência de n-gramas entre sumários automáticos e humanos (ou de referência) e a expressa pelas medidas "precisão", "cobertura" e "medida-f".

Agradecimento. À FAPESP, pelo suporte financeiro (Proc. Nº 2017/15344-8)

Referências bibliográficas

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. (2011). CSTNews - A discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese. In the Proceedings of the 3rd RST Brazilian Meeting, Cuiabá/MT, Brazil.

Cristrini, L.F.; Di-Felippo, A. (2018) Source Texts Annotation for Rewriting References to People in Automatic Multi-Document Extracts. In the Proceedings of the PROPOR Student Research Workshop (Tilic), pp. 1-5. September, 24. Canela, RS/Brazil.

Dang, H.T. (2005). Overview of DUC 2005. In the Proceedings of the Document Understanding Conference (HLT/EMNLP Workshop on Text Summarization), 2005.

Di-Felippo, A. (2016). "Revisão de sumários baseada em conhecimento: transformando extratos multidocumento em *abstracts*". Relatório de Bolsa de Pesquisa no Exterior (FAPESP #2015/01450-5). http://www.nilc.icmc.usp.br/nilc/index.php/team?id=23.

Friedrich, A., Valeeva, M., Palmer, A. (2014). "LQVSumm: a corpus of linguistic quality violations in multi-document summarization". In the Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC), Reykjavik/ISL, pp. 1591-1599.

Kaspersson, T.; Smith, C.; Danielsson, H.; Jönsson, A. (2012). This also affects the context - Errors in extraction based summaries. In the Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC), Istanbul/TU, pp.173-8.

Lin, C-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In the Proceedings of the ACL Workshop on Text Summarization Branches, p. 74-81.

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

Nenkova, A.; McKeown, K. (2003a). Improving the Coherence of Multi-document Summaries: a Corpus Study for Modeling the Syntactic Realization of Entities, Columbia University, CS Department Technical Report, CUCS-001-03.

Nenkova, A.; McKeown, K. (2003b). Improving the Coherence of Multi-document Summaries: a Corpus Study for Modeling the Syntactic Realization of Entities, Columbia University, CS Department Technical Report, CUCS-001-03.

Nenkova A.; Mckeown. K. (2011). Automatic summarization. In *Foundations and Trends in Information Retrieval*, 5(2-3), pages 103–233.

Pardo, T. A. S. (2005) GistSumm - GIST SUMMarizer: Extensões e novas funcionalidades, Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.

Ribaldo, R.; Akabane, A. T.; Rino, L. H. M.; Pardo, T. A. S. (2012). Graph-based Methods for Multidocument Summarization: Exploring Relationship Maps. Complex Networks and Discourse Information. In the Proceedings of the 10th International Conference on Computational Processing of Portuguese (LNAI 7243),Coimbra/Portugal, pp. 260–271.

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

Siddharthan, A., Nenkova, A., McKeown, K. (2011). Information status distinctions and referring expressions: An empirical study of references to people in news summaries. In *Computational Linguistics* 37(4), pages 811–842.