Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP

Descrição e Análise do Fenômeno da Contradição para a Sumarização Automática Multidocumento

Naira Lícia da Silva
Ariani Di Felippo

NILC-TR-14-03

Setembro, 2014

Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional
NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil

Resumo

Neste trabalho, realizou-se a descrição dos casos de contradição no *corpus* multidocumento de referência do português, o CSTNews (CARDOSO et al., 2011), e a delimitação de características linguísticas desse fenômeno. Com isso, geraram-se subsídios linguísticos que podem fundamentar a identificação e o tratamento do fenômeno da contradição pelos sumarizadores automáticos multidocumento do português, posto que, no geral, busca-se produzir sumários automáticos livres de informações contraditórias. A referida descrição consistiu na caracterização ou tipificação dos casos de contradição do CSTNews. A pesquisa ora descrita foi realizada em uma iniciação científica que compreendeu o período de 01/09/2013 a 31/08/2014.

Este trabalho contou com o apoio financeiro
da FAPESP (2013/12521-5).

1. Introdução

Na subárea do Processamento Automático das Línguas Naturais (PLN) denominada Sumarização Automática Multidocumento (SAM), objetiva-se gerar automaticamente sumários (ou resumos) a partir de coleções de textos, provenientes de fontes distintas, que abordam um mesmo assunto.

A SAM tem sido motivada pela enorme quantidade de informação textual disponível na *web* e pelo pouco tempo que as pessoas têm para assimilar tanta informação (MANI, 2001).

De uma forma geral, a SAM tem focado a produção de extratos (ou seja, sumários formados pela justaposição de sentenças integralmente extraídas dos textos-fonte) informativos e genéricos, produzidos a partir de coleções de notícias jornalísticas (KUMAR, SALIM, 2012). Por serem informativos e genéricos, os extratos multidocumento contêm a informação principal dos textos-fonte de tal forma que a sua leitura por uma audiência genérica exclui a necessidade da leitura das notícias-fonte.

Para tanto, os sistemas de SAM precisam processar textos-fonte que, por tratarem do mesmo assunto, podem apresentar grande quantidade de informações repetidas, e, por serem de fontes distintas, podem apresentar várias informações contraditórias, geradas por diferentes atualizações da notícia ou pontos de vista (MANERFEEE, 2012).

A identificação de informações contraditórias ou conflitantes na SAM faz-se importante porque a ausência de contradição é um dos fatores que garantem a informatividade e a coerência dos sumários. Esse fenômeno, entretanto, é bastante complexo, pois sua identificação por depender de processos profundos de inferência.

Assim, seguindo os trabalhos de Marnefee et al. (2008) e Marnefee (2012), realizou-se a descrição/análise dos casos de contradição no único *corpus* multidocumento de referência do português, o CSTNews (CARDOSO et al., 2011). Com isso, geraram-se subsídios linguísticos que podem fundamentar a identificação e o tratamento desse fenômeno pelos sumarizadores multidocumento do português.

A descrição e análise englobaram a caracterização ou tipificação da contradição no CSTNews e a proposta de estratégias para subsidiar a detecção automática dos diferentes tipos de contradição do *corpus* em língua portuguesa. Em vista disso, este trabalho se inseriu no cenário do projeto SUSTENTO[1] (FAPESP 2012/13246-5/ CNPq 483231/2012-6), que busca gerar subsídios para a SAM em português.

Na seção 2, revisam-se os conceitos básicos de SAM e o fenômeno da contradição no PLN. Na seção 3, apresenta-se o *corpus* CSTNews. Na Seção 4, caracteriza-se o fenômeno no *corpus*. Na Seção 5, propõem-se estratégias para a detecção automática da contradição e, por fim, na Seção 6, tecem-se comentários finais sobre os resultados e apontam-se trabalhos futuros.

[1] http://www.nilc.icmc.usp.br/arianidf/sustento/

2. Revisão da literatura

Os tópicos abordados na revisão da literatura foram: (i) conceitos básicos sobre sumarização automática multidocumento e (ii) o fenômeno da contradição.

2.1. A Sumarização Automática Multidocumento

Atualmente, vive-se a era da "explosão da informação", sendo que grande parte dela é veiculada por agências de notícias *on-line* na forma de texto. Com isso, o processamento da informação textual torne-se difícil para os humanos e também para as máquinas. Enquanto humanos não têm capacidade e tempo para ler/apreender as informações de interesse, as máquinas perdem em precisão e desempenho.

Com isso, as pesquisas sobre SAM têm adquirido relevância nos últimos anos, pois a SAM consiste na produção de um único sumário (resumo) a partir de uma coleção de textos-fonte sobre determinado tópico provenientes de fontes distintas (MANI, 2001). A SAM surgiu como uma extensão natural da tradicional sumarização monodocumento, que visa à produção de um sumário a partir de um único documento.

De uma forma geral, as pesquisas sobre SAM têm focado a produção de extratos informativos e genéricos de coleções de notícias jornalísticas (KUMAR, SALIM, 2012). A natureza extrativa dos sumários é decorrente do fato de estes serem formados pela justaposição de sentenças integralmente extraídas dos textos-fonte[2]. Por serem informativos[3] e genéricos[4], os extratos multidocumento contêm a informação principal dos textos-fonte de tal forma que a sua leitura por uma audiência genérica exclui a necessidade da leitura das notícias-fonte.

Para produzir tais sumários, tem-se desenvolvido métodos/sistemas de SAM baseados nas abordagens superficial e profunda. Na superficial, faz-se uso de pouco ou nenhum conhecimento linguístico, sendo mais escaláveis e robustos. Normalmente, eles fazem uso de estatística e conhecimentos empíricos. Nos métodos/sistemas profundos, usa-se muito conhecimento linguístico, codificado em gramáticas, repositórios semânticos e modelos de discurso, e, por isso, seu desenvolvimento é caro e sua aplicação é restrita, mas seus sumários são mais informativos e gramaticais.

Os métodos/sistemas de SAM superficiais e profundos apresentam uma arquitetura composta por três etapas, como ilustrado na Figura 1 elaborada com base em Sparck Jones (1993) e Mani e Maybury (1999).

---

[2] Há também os sumários abstrativos, que contêm certa reescrita do material dos textos-fonte.

[3] Há também os indicativos e os críticos. Os indicativos apenas indicam o conteúdo dos textos-fonte (p.ex.: índices de livros). Os críticos adicionam crítica ao conteúdo textual (p.ex.: resenhas de livros)

[4] Em função da audiência, há também os sumários específicos, voltados para interesses particulares dos usuários (os quais são normalmente especificados via uma *query* ou consulta dos usuários).

Figura 1: Arquitetura genérica de um sistema de SAM.

A primeira das etapas é a "análise", que corresponde à interpretação dos textos-fonte e geração de uma representação do conteúdo linguístico expresso em termos computacionais.

A segunda é a "transformação", etapa em que o conteúdo formalizado dos textos-fonte é selecionado e condensado em uma representação computável, ou seja, não-textual. Especificamente, a transformação é comumente realizada por meio dos seguintes passos básicos (RADEV et al., 2004): (i) calcular a importância de cada sentença dos textos-fonte com base em sua redundância na coleção[5]; (ii) ranquear as sentenças em função de sua importância; (iii) selecionar a 1ª sentença do ranque para iniciar o sumário, e (iv) selecionar as demais sentenças do ranque até que o tamanho desejado do sumário seja alcançado.

A terceira etapa é a "síntese". Nela, o conteúdo condensado é expresso em língua natural na forma de um sumário. Para tanto, métodos de justaposição, ordenação, fusão e correferenciação dos segmentos textuais selecionados podem ser utilizados.

As três etapas que compõem a arquitetura dos métodos ou sistemas de SAM são guiadas pela taxa de compressão, ou seja, o tamanho desejado do sumário; um sumário com taxa de compressão de 70% apresenta tamanho equivalente a 30% do tamanho do texto original (em geral, medido em número de palavras).

Ao partir de uma coleção de textos, provenientes de fontes distintas, sobre um mesmo assunto, a SAM caracteriza-se pela ocorrência de informações redundantes, complementares e contraditórias. Daí dizer que a redundância, complementaridade e contradição são "fenômenos multidocumento".

Idealmente, na etapa de análise dos métodos/sistemas de SAM, deve-se produzir uma interpretação dos textos-fonte que revele esses fenômenos para que, na transformação, apenas as informações mais relevantes da coleção (ou seja, as mais redundantes) sejam selecionadas para compor o respectivo sumário sem que haja repetição ou contradição entre elas.

Suponha-se que, dado um ranque das sentenças de certa coleção, a 1ª sentença é selecionada para o sumário sem atingir a taxa de compressão. Por conseguinte,

[5] A redundância é o critério de relevância mais empregado na SAM para selecionar o conteúdo a compor o sumário. Esse critério é linguisticamente motivado, pois pauta-se no fato comprovado em pesquisas sistemáticas de *corpus* de que o conteúdo mais recorrente nos textos-fonte é selecionado pelos humanos para a produção de um sumário multidocumento (MANI, 2001; NENKOVA, 2006; CAMARGO, 2013).

seleciona-se a 2ª do ranque, cujo tamanho, somado ao tamanho da 1ª sentença, atinge a taxa. Caso os fenômenos tenham sido identificados na análise, verifica-se há redundância ou contradição entre a 1ª e a 2ª sentença. A 2ª sentença somente comporá o sumário se não apresentar redundância ou contradição em comum com a 1ª sentença. Caso contrário, o método/sistema seleciona a 3ª sentença do ranque e realiza as mesmas verificações quanto à presença de informações redundantes ou contraditórias entre ela e a 1ª sentença até atingir a taxa de compressão desejada.

Dessa forma, vê-se a importante da identificação dos fenômenos multidocumento pelas métodos/sistemas de SAM. A seguir, apresenta-se resumidamente as principais estratégias de detecção dos fenômenos multidocumento para a SAM em português.

2.2. O fenômeno da contradição e a sua detecção automática

Para entender esse fenômeno, destaca-se Condoravdi et al. (2003), que foram os primeiros autores a reconhecer que a identificação da contradição é importante para a compreensão do significado textual no PLN e a propor regras computacionalmente tratáveis para identificá-la.

Nesse trabalho, os autores partem de uma definição lógica, de acordo com princípios da Semântica Formal. Assim, duas sentenças (S1 e S2) são contraditórias quando não há um mundo possível em que S1 e S2 sejam ambas verdadeiras[6]. Em outras palavras, S1 e S2 são contraditórias entre si quando se uma for verdadeira, a outra tiver de ser falsa; ou então quando não houver situação alguma que possa ser descrita simultaneamente por ambas as sentenças (CHIERCHIA, 2003).

Para ilustrar, considerem-se as sentenças em (3), extraídas de Condoravdi et al. (2003, p.1), as quais são contraditórias porque, se S1 for verdadeira, S2 tem de ser falsa.

(3) Nenhum civil foi morto no ataque suicida in Najaf. (S1)
(*No civilians were killed in the Najaf suicide bombing.*)
Dois civis morreram no ataque suicida em Naja. (S2).
(*Two civilians died in the Najaf suicide bombing.*)

Para identificar automaticamente a contradição assim definida, os autores propuseram uma série de regras lógicas, baseadas em condições de verdade, as quais, contudo, não foram implementadas.

Com base nessa definição de contradição, sentenças como as do exemplo (4) não são consideradas contraditórias, pois é possível que ambas sejam simultaneamente verdadeiras, apesar de que intuitivamente os humanos classifiquem a informação veiculada por ambas como contraditória (MARNEFEE, 2012, p. 65).

[6] "[...] *sentences A and B are contradictory if there is no possible world in which A and B are both true*" (CONDORAVDI et al., 2003, p. 1).

(4) John pensa que ele é incompetente. (S1)
(*John thinks that he is incompetente*.)
Seu chefe acredita que a John não está sendo dada a oportunidade. (S2)
(*His boss believes that John is not being given a chance*)

Diante disso, Marneffe et al. (2008) salientam que o conceito de contradição da lógica é restrito e não combina com a intuição humana sobre esse fenômeno. Consequentemente, os autores trabalham com uma definição mais frouxa, que busca capturar as intuições de incompatibilidade sobre um mesmo evento, a saber: "a contradição ocorre quando é pouco provável que duas sentenças, S1 e S2, sejam simultaneamente verdadeiras"[7]. Assim, pares de sentenças como (5) de Marnefee et al. (2008, p. 1040) são classificados como contraditórios mesmo que um possa ter vendido um barco ao outro.

(5) Sally vendeu um barco ao John. (S1)
(*Sally sold a boat to John.*)
John vendeu um barco a Sally. (S2)
(*John sold a boat to Sally.*)

A partir dessa definição que busca capturar a intuição humana sobre a contradição, Marneffe et al (2008) identificaram um conjunto de tipos de contradição com base na descrição e análise de *corpus* em língua inglesa. No Quadro 1, a tipologia de Marneffe et al. (2008) é apresentada e ilustrada em português.

Quadro 1: Exemplificação em português dos tipos de contradição de Marneffe (2012).

	Tipo	**Sentença 1**	**Sentença 2**
1	Antonímia	A pena de morte é um catalizador para mais crime.	A pena de morte é um impedimento para o crime.
2	Negação	A Suprema Corte disse que os júris e não os juízes devem impor uma sentença de morte.	O Supremo Tribunal Federal decidiu que os juízes podem impor a pena de morte.
3	Número, data e tempo	A tragédia da explosão em Qana, que matou mais de 50 civis, apresentou a Israel um dilema.	Uma investigação sobre o ataque em Qana identificou 28 mortos confirmados até agora.
4	Modalidade	O terrorista pode ter entrado na embaixada.	O terrorista entrou na embaixada.
5	Estrutura	Jacques Santer sucedeu Jacques Delors como presidente da Comissão Europeia, em 1995.	Delors sucedeu Santer na presidência da Comissão Europeia.

[7] "[...] *contradiction occurs when two sentences are extremely unlikely to be true simultaneously*." (MARNEFEE et al., 2008, p. 1040).

6	Léxico	Maitur Rehman, uma paquistanês de 29 anos, é referido como o atual chefe de Jundullah.	Fontes dizem que Maitur Rehman é um militante de baixo escalão que opera no Waziristão do Sul.
7	WK	Microsoft Israel, uma das primeiras filiais da Microsoft fora dos EUA, foi fundada em 1989.	A Microsoft foi fundada em 1989.

Segundo Marneffe (2012), as sentenças do exemplo (1) do Quadro 1 são contraditórias devido à presença das unidades lexicais "catalisador" ("que inicia ou acelera algo") (S1) e "impedimento" (S2), consideradas antônimas.

Em (2), a contradição caracteriza-se pela presença da palavra "não" (S1), que nega o conteúdo de S2.

Já em (3), as sentenças apresentam uma discrepância numérica sobre o mesmo fato; no caso, S1 veicula a morte de 50 civis e S2, de 28.

A contradição ilustrada em (4) emerge da diferença de modalidade (isto é, estratégia pela qual o falante expressa seu relacionamento com o conteúdo proposicional), pois S1 apresenta o verbo auxiliar modal "pode", que indica a incerteza do falante sobre o fato ("entrada do 6 terrorista na embaixada"), e S2 não.

Em (5), o sujeito de S1 é *Jacques Santer*, enquanto em S2, o sujeito é *Delors*, o que sugere que as duas são incompatíveis.

Em (6), a contradição emerge da oposição conceitual entre a unidade lexical "chefe", em S1, e a expressão "militante de baixa escalão" de S2.

Por fim, em (7), a incompatibilidade surge devido ao nosso conhecimento de mundo sobre empresas matrizes e filiais.

Segundo os autores, os tipos de contradição do Quadro 1 podem ser classificados como "simples" ou "complexos". As contradições simples caracterizam-se pela ocorrência da antonímia, negação e incompatibilidades numéricas. As complexas envolvem modalidade, conhecimento lexical e estrutural e também conhecimento de mundo (*world knowledge*, WK). Em um *corpus* composto por 131 pares de sentenças contraditórias manualmente coletadas de textos jornalísticos em inglês, os autores verificaram que a frequência dos tipos descritos no Quadro 1 é a descrita na Tabela 1.

Tabela 1: Porcentagem de distribuição dos tipos de contradição em Marnefee (2012).

Classe	Tipo	Frequência
Simples	Antonímia	9,2%
	Negação	17,2%
	Numérico	29%
Complexa	Modalidade	6,9%
	Estrutura	3,1%
	Léxico	21,4%
	WK	13%

Observa-se na Tabela 1 que as contradições simples são ligeiramente mais frequentes, totalizando aproximadamente 55,5%. Destacam-se nesse grupo as do tipo "negação" (17%) e "numérica" (29%). Quanto às complexas, observa-se que a do tipo lexical também é frequente (21,4%).

Para Marnefee (2012), essa distribuição confirma a hipótese de que em um *corpus* composto por textos jornalísticos sobre um mesmo assunto, a contradição ocorre principalmente por duas razões: as informações são atualizadas/publicadas conforme o conhecimento sobre o evento/fato é adquirido ao longo do tempo (p.ex.: número de mortos) ou as notícias são escritas sob diferentes pontos de vista.

Quanto à identificação automática contradição, Harabagiu et al. (2006) forneceram os primeiros resultados empíricos. Considerando a contradição uma relação que pode ser caracterizada por marcas semântico-discursivas específicas, os autores analisaram um *corpus* em inglês e constataram que a negação, a antonímia e informações semântico-pragmáticas, como tempo, modalidade, factividade e atos de fala, são essenciais para localizar diversas contradições.

Como exemplo, a contradição entre S1 *"Joachim Johansson revidou dramaticamente o ataque do atual campeão Andy Roddick"* e S2 *"O atual campeão Andy Roddick nunca enfrentou Joachim Johansson"* evidencia-se pela presença do advérbio "nunca" em S2. Apesar de S1 e S2 serem semanticamente similares, a ocorrência de "nunca" inverte o valor de verdade de S2 pela negação. Assim, os autores propõem comparar a ocorrência de expressões de negação diretas (p. ex.: "não", "nada", etc.) e indiretas (p. ex.: verbos como "negar", "falhar", preposições como "sem", "exceto", etc.) entre as sentenças de um par.

Outro exemplo é a contradição entre S1 "*Tais normas impossibilita a venda de armas para estados como Líbia*" e S2 "*Mas estados como Ruanda, antes de sua crise atual ainda seria capaz de comprar armas legalmente*". Nesse caso, ela que se manifesta pela antonímia direta entre "venda"/"vender" e "comprar" e pela antonímia indireta entre "impossibilitar" e "ser capaz". Para detectar os contrastes, os autores usam a WordNet de Princeton (FELLBUAM, 1998), base lexical que armazena oposições conceituais entre nomes, adjetivos, verbos e advérbios.

Quanto ao trabalho Harabagiu et al. (2006), ressalta-se que as estratégias de detecção da contradição pautam-se em conhecimento linguístico variado, os quais são obtidos por diversas ferramentas de PLN (p.ex.: etiquetadores morfossintáticos e analisadores sintático-semânticos e pragmáticos, etc.) e recursos linguístico-computacionais externos (p.ex.: WordNet). De um modo geral, tais estratégias obtiveram aproximadamente 62% de precisão na detecção da contradição.

Com base na tipologia do Quadro 2, Marneffe (2012) especificaram várias estratégias ou métodos para a identificação automática da contradição entre duas sentenças. Tais estratégias buscam capturar a contradição com base nos atributos linguísticos referentes a cada um dos tipos do Quadro 2. Alguns deles, aliás, assemelham-se aos de Harabagiu et al. (2006), como a presença de palavras antônimas em dado par de sentenças. De uma forma geral, para os tipos simples, a precisão obtida por Marneffe (2012) foi próxima à de Harabagiu et al. (2006),

girando em torno de 65%. Para os tipos mais complexos, o desempenho das estratégias foi mais modesto.

Em seguida, descreve-se o *corpus* em que a contradição está sendo investigado.

3. Seleção e Recorte do *corpus*

Para a realização desta pesquisa, foi necessário um *corpus*. Por definição, um *corpus* é um conjunto de dados linguísticos sistematizados de acordo com determinados critérios, de maneira que possa ser processado por computador com a finalidade de propiciar resultados vários e úteis para a descrição e análise (SINCLAIR, 2005).

Considerando os objetivos ora traçados, selecionou-se o CSTNews, um *corpus* multidocumento de referência em português (CARDOSO et al., 2011).

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

Especificamente, cada coleção do CSTNews contém: (i) 2 ou 3 textos sobre um mesmo assunto/evento compilados de diferentes fontes jornalísticas; (ii) sumários humanos mono e multidocumento; (iii) sumários automáticos multidocumento, e (iv) diferentes tipos de anotações linguísticas intra e intertextuais.

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

Ademais, as coleções estão categorizadas pelos rótulos das "seções" dos jornais dos quais os textos foram compilados. Assim, o corpus é composto por coleções das seguintes categorias: "esporte" (10 coleções), "mundo" (14 coleções), "dinheiro" (1 coleção), "política" (10 coleções), "ciência" (1 coleção) e "cotidiano" (14 coleções).

Os textos-fonte no interior de cada coleção do CSTNews foram manualmente alinhados em nível sentencial, formando-se pares, como ilustrado na Figura 2.

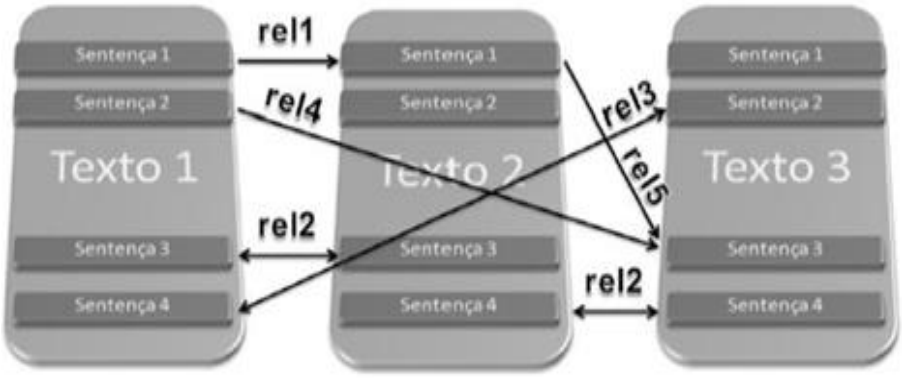

Figura 2: Esquema genérico de alinhamento multidocumento.
Fonte: Maziero (2012, p. 14)

Os alinhamentos foram rotulados por relações semântico-discursivas baseadas na *Cross-document Structure Theory* (CST) (RADEV, 2000), teoria/modelo que permite estruturar o discurso pela conexão das sentenças (ou outras unidades textuais) advindas de diferentes, capturando os diferentes fenômenos multidocumento.

Segundo Maziero et al. (2010), as relações CST podem ser classificadas em dois grupos: relações de conteúdo (isto é, que ligam o conteúdo das sentenças) e relações de forma (ou seja, relações que ligam sentenças com base na forma), como ilustrada a Figura 3. As relações de conteúdo podem indicar "redundância", "complemento" ou "contradição". As relações de forma podem ser do tipo "fonte/autoria" ou "estilo".

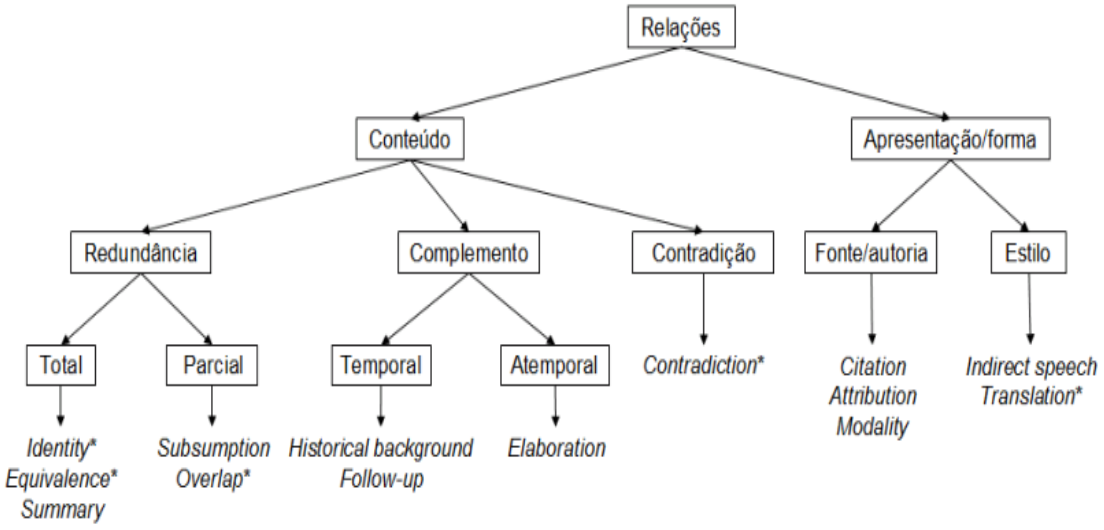

Figura 3: Tipologia das relações CST de Maziero et al. (2010).

O reconhecimento ou anotação manual da relação *Contradiction* no CSTNews foi feita com base em uma definição bastante genérica do fenômeno, segundo a qual "*Contradiction* é uma relação de conteúdo com direcionalidade nula que ocorre entre

duas sentenças, S1 e S2, quando estas divergem sobre algum de seus elementos[8] (MAZIERO et al., 2010).

Tomando-se como base a anotação CST do *corpus*, fez um recorte no CSTNews, que consistiu na seleção de todos os pares de sentenças associadas pela relação CST *Contradiction*, que resultou em um *subcorpus* composto por 46 pares de sentenças[9]. Ressalta-se que os erros de digitação, como "pro" ("por") no par 30, foram mantidos, assim como os pares repetidos, como 9 e 10. A repetição de pares decorre do fato de o *corpus* ser multidocumento, ou seja, composto por mais de uma notícia sobre o mesmo evento/acontecimento.

4. Descrição/análise da contradição

A partir do recorte do CSTNews, procedeu-se, segundo a metodologia originalmente proposta, à descrição/análise dos casos de contradição, que englobou a (i) tipificação e (ii) a tabulação da frequência dos casos. Com base nessas duas tarefas, realizou-se, na sequência, a identificação de atributos ou características sentenciais que podem subsidiar a detecção automática da contradição.

4.1. Tipificação dos casos de contradição do CSTNews

A tipificação dos 46 pares do *corpus* de contradição foi feita com base na definição mais ampla de Marnefee et al. (2008), que busca capturar a intuição humana sobre o fenômeno. Ademais, os casos foram tipificados em função da tipologia de Marnefee (2012) (cf. Quadro 2).

Para a tipificação, fez-se uma análise manual prévia dos 46 casos do CSTNews, cujas sentenças foram manualmente conectadas pela relação CSTNews *Contradiction.* Essa análise visava verificar se todos os casos extraídos do CSTNews encaixavam-se na definição de contradição adotada neste trabalho.

Após a análise, o conjunto inicial de 46 casos foi reduzido para 40, pois 6 casos foram considerados não-contraditórios, os quais são descritos no Quadro 2.

Quadro 2: Pares do CSTNews não-contraditórios anotados com *Contradiction*

No.		**Pares de sentenças**
8	a	Ao menos 300 policiais de Amapá, Distrito Federal, Mato Grosso, Acre e Rondônia trabalharam na Operação Dominó.
	b	Mais de 300 policiais federais do Amazonas, Distrito Federal, Mato Grosso, Acre e Rondônia participaram das buscas e prisões.
19	a	Os feridos, vários deles em estado grave, foram levados para seis hospitais da região onde aconteceu o acidente, um dos piores ocorridos no Egito desde 2002, disse El-Gabaly, citado pela agência egípcia de notícias Mena.

[8] http://www.icmc.usp.br/pessoas/taspardo/sucinto/supplementary_material.html.
[9] Os 46 pares cujas sentenças foram anotadas com a relação *Contradiction* compõem o Anexo 1.

	b	O acidente ocorreu no delta do Nilo, ao norte de Cairo, no Egito.
39	a	Entre outras conquistas, a seleção nacional soma dois títulos do Campeonato Mundial, uma medalha de ouro nos Jogos Olímpicos de Atenas-04 e a Copa do Mundo do Japão, em 2003.
	b	A conquista reforça o status de melhor time de vôlei do mundo, uma vez que o Brasil soma cinco títulos consecutivos da Liga Mundial, além do título da Copa do Mundo e das Olimpíadas de Atenas, em 2004.
40	a	O terremoto deixou até o momento nove mortos, todos idosos, mais de mil feridos e cerca de 13 mil desabrigados, assim como grandes danos materiais.
	b	Ao menos 9 pessoas morreram e cerca de 700 se feriram.
42	a	O grande destaque da prova foi Nicolas Oliveira, quarto nadador brasileiro a cair na água.
	b	Thiago, que abriu o revezamento, foi o grande destaque do quarteto nas piscinas do Parque Aquático Maria Lenk.
44	a	O tremor atingiu a região às 10h13 (horário local, 22h13 de domingo, em Brasília) e seu epicentro foi localizado a 260 km da costa de Niigata, ao nordeste da capital, Tóquio, onde também foi sentido.
	b	Duas mulheres na faixa dos 80 anos morreram quando suas casas ruíram durante o tremor de magnitude 6,8, na área de Niigata, cerca de 250 km noroeste de Tóquio, informou a imprensa japonesa.

A suposta contradição no par 8 se refere às expressões “ao menos 300 policiais” e “mais de 300 policiais”, nas sentenças (a) e (b), respectivamente. Esse par foi consideração não-contraditório porque a expressão quantitativa “ao menos 300”, da sentença (a), não contradiz a expressão “mais de 300” da sentença (b) e vice-versa. Em outras palavras, “mais de 300” pressupõe que “ao menos 300” policiais trabalharam na operação e, por outro lado, “ao menos 300” pode indicar que o contingente de polícias englobava “mais de 300”. Assim, ambas as sentenças podem ser verdadeiras.

O par 19 foi considerado não-contraditório porque envolve uma relação de especificação/generalização. Quanto ao local, a sentença (a) é mais genérica que a sentença (b), pois (b) expressa a região do Egito em que o acidente em questão ocorreu. Essa diferença de granularidade não configura contradição de acordo com a definição em que este trabalho se baseia, já ambas as sentenças de 19 são verdadeiras.

As sentenças do exemplo 39 também não apresentam informações conflitantes, já que todos os títulos que a seleção masculina de vôlei ganhou nos referidos campeonatos estão corretos. Esse equívoco na anotação das sentenças com a relação *Contradiction* talvez se justifique pelo fato de que sua correta interpretação depende de conhecimento de mundo/domínio dos esportes. A maioria das pessoas

não sabe, por exemplo, que Campeonato Mundial (39a) e Liga Mundial (39b) são eventos distintos, nos quais o Brasil possuía, na época, 2 e 5 títulos, respectivamente.

Quanto ao par 40, ressalta-se que os trechos "mais de mil feridos", em (a), e "cerca de 700" não são contraditórios.

Sobre o par 42, ressalta-se que a contradição inicialmente anotada não se sustenta porque o nadador Nicholas Santos foi o destaque de uma prova específica, enquanto Thiago Pereira foi o destaque de toda a competição (que envolve várias provas). Por conseguinte, ambas veiculam informação verdadeira.

Por fim, a não-contradição das sentenças do par 44 justifica-se pelo fato de que "260 km" (a) não contradiz a informação veiculada pela sentença (b), já que "260 km" pode ser interpretado como "cerca (aproximadamente) a 250 km".

Assim, dos 46 originais, 40 permaneceram para a descrição/análise. Para a tipificação dos 40, criou-se uma tabela no formato Excel, em que os pares foram alocados na primeira coluna, um em cada linha, e os tipos de contradição nas demais colunas. Para cada par, assinalou-se com "1" o respectivo tipo. Caso mais de um tipo caracteriza-se a contradição, cada um deles recebeu o valor "1".

No Quadro 3, apresenta-se a tipificação dos 40 pares. As linhas cinzas indicam os casos excluídos. Quanto à tipificação, seguem algumas observações:

- Dos 9 pares de contradição do tipo "negação" no CSTNews, 4 deles (ou seja, 11, 12, 13, 14 e 15) são repetidos e 1 é parcialmente repetido (par 11); nos casos repetidos, as sentenças compartilham o mesmo verbo principal "confirma" (e sujeito), sendo que, em uma delas, ele é negado pela ocorrência do advérbio "não". No caso parcialmente repetido, ocorre o advérbio de negação "não" em uma das sentenças.
- A contradição do par 5 caracteriza-se por ser de dois tipos: negação e modalidade. A negação se caracteriza pela ocorrência do advérbio "não" em uma das sentenças e a modalidade pela ocorrência do verbo auxiliar "teria (apresentado)", no futuro do pretérito/condicional, na outra sentença.
- A negação dos pares 24, 32 e 43 caracteriza-se pela ocorrência, respectivamente, dos advérbios "ainda", "nenhuma" e "não", em apenas uma das sentenças de cada par.
- Dos 26 casos de contradição do tipo "número, data ou tempo", 2 envolvem a expressão de porcentagem (3 e 4), 2 envolvem velocidade (km/hora) (6 e 7), 2 casos envolvem a expressão de distância (33 e 34) e 4 envolvem informação de tempo (25, 26, 37 e 38). Nos demais 12 casos (9, 10, 16, 17, 18, 20, 21, 23, 27, 28, 29, 31, 32, 41, 45 e 46), a informação numérica que mais se destaca na geração da contradição é a discrepância quanto ao número de pessoas feridas/mortas em acidentes/atentados.
- O caso de contradição do par 2 foi tipificado como "léxico" porque há nomes distintos para uma mesma entidade ("*Air Traset*" / "*Trapset Congo*").
- A contradição do par 22 é do "léxico" por causa da possível divergência entre os conceitos "ferido" e "morto". No caso, interpretou-se que o conceito "ferido" não

pressupõe o conceito “morto”, gerando, assim, a contradição entre as sentenças do par, pois uma delas veicula apenas a informação sobre os “feridos” do referido acidente, enquanto a outra veicula informação sobre “feridos” e “mortos.

- Os casos de contradição do tipo WK surgem por conflitos relativos a conhecimento de mundo variado. Em outras palavras, a contradição ocorre pelo conflito de informação sobre: (i) a causa da queda de um avião (1); (ii) o índice de intenção de voto em um processo eleitoral (3); (iii) (7); (iv) o andar de um prédio em que ocorreu uma explosão (29); (v) período de tempo (parte/todo) (30); (vi) posse de bola em um jogo de futebol (35); (vii) comportamento de um time de futebol (36); (viii) direitos de uma comissão de investigação (43), e (ix) (46) placar de um jogo de vôlei.

Quadro 3: Tipificação dos casos de contradição do CSTNews.

Par	Antonímia	Negação	Número/ data/ tempo	Modalidade	Estrutura	Léxico	WK
1							1
2						1	
3			1				1
4			1				
5		1		1			
6			1				
7			1				1
8							
9			1				
10			1				
11		1					
12		1					
13		1					
14		1					
15		1					
16			1				
17			1				
18			1				
19							
20			1				
21			1				
22						1	
23			1				
24		1					
25			1				
26			1				
27			1				
28			1				
29			1				1
30							1
31			1				

32		1	1				
33			1				
34			1				
35							1
36							1
37			1				
38			1				
39							
40							
41			1				
42							
43		1					1
44							
45			1				
46			1				1
Total	0	9	26	1	0	2	9

A tipificação descrita por ser vista como um tipo de anotação semântica de *corpus*. Por conseguinte, a tipificação organizada inicialmente em uma tabela Excel poderá ser codificada no formato XML (do inglês, *Extensible Markup Language*), visto que as demais anotações do CSTNews estão codificadas nessa linguagem[10].

Todos os 40 pares serão codificados sequencialmente entre as etiquetas ou *tags* `<contradiction-corpus>` e `</contradiction-corpus>`. No Quadro 4, tem-se apenas a tipificação do par (6) para ilustração. As informações referentes ao par estão codificadas entre as *tags* `<pair>` e `</pair>`. O par é especificado por duas informações: (i) identificador, codificado pelo atributo `ID` (no caso, `"6"`), e (ii) tipo de contradição, codificado pelo atributo `CONTRADICTION_TYPE` (no caso, `"numero"`).

A anotação em XML englobará ainda as sentenças que compõem cada um dos pares e algumas informações sobre elas. As informações sobre as sentenças estarão codificadas entre as *tags* <s> e </s>, sendo que essas informações englobarão os atributos `SENT` e `DOC`, além da própria sentença.

O atributo `SENT` especificará o número das sentenças no texto-fonte (no caso, `"3"`) e `DOC` especificará o texto-fonte (p.ex.: `"D1_C4_Folha.txt.seg"`). As sentenças do par serão codificadas após os atributos `SENT` e `DOC`, entre os sinais "< >".

No exemplo, tem-se a s1 `<O congestionamento esteve ainda maior às 9h, quando chegou a 113 km de extensão para uma média de 32 km>` e a s2 `<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)>`.

[10] O XML tem sido utilizado no PLN para armazenar de forma padronizada os diferentes tipos de anotação de *corpus*. Em breve, a tipificação da contradição no CSTNews estará disponível no site do SUSTENTO (http://www.nilc.icmc.usp.br/arianidf/sustento/) para futuras pesquisas linguístico-computacionais.

Quadro 4: Exemplo da representação em XML da tipificação.

```
<contradiction-corpus>
    <pair ID="6" CONTRADICTION_TYPE="numero">
          <s1 SENT="3" DOC="D1_C4_Folha.txt.seg"><O congestionamento
esteve ainda maior às 9h, quando chegou a 113 km de extensão para uma
média de 32 km.></s1>
          <s2 SENT="2" DOC="D2_C4_Estadao.txt.seg"><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>
    </pair>
<contradiction-corpus>
```

4.2. Cálculo da frequência dos tipos de contradição no CSTNews

Após a tipificação, procedeu-se ao cálculo ou tabulação da frequência de cada um dos tipos, a qual é apresentada na Tabela 2 em número absolutos e porcentagem.

Dos dados da Tabela 2, tecem-se as primeiras observações ou análises a respeito do fenômeno da contradição no *corpus* CSTNews: (i) as contradições "simples", de uma forma geral, ocorrem com mais frequência que as da classe "complexa"; (ii) os tipos "antonímia" e "estrutura" não ocorreram no *corpus*, tendo frequência 0; (iii) dentre os que ocorreram, o tipo "modalidade" foi o menos frequente (2,12%), seguido do tipo "léxico" (4,26%); (iv) os tipos "negação" e "WK" empataram em segundo lugar, com 19,15%, e (v) o tipo "número/ data/ tempo" foi o mais frequente, com mais da metade das ocorrências de contradição no CSTNews, no caso, 55,32%.

Tabela 2: Frequência dos tipos de contradição no CSTNews.

Classe	Tipo	Quantidade	Porcentagem
Simples	Antonímia	0	0%
	Negação	9	19,15%
	Número/ data/ tempo	26	55,32%
Complexa	Modalidade	1	2,12%
	Estrutura	0	0%
	Léxico	2	4,26%
	WK	9	19,15%
Total		**47**	**100%**

Comparando a frequência dos tipos de contradição no CSTNews e no *corpus* de Marnefee (2012), composto por 131 pares de sentenças de textos jornalísticos em inglês (cf. Tabela 1), vê-se que a distribuição é bastante similar. A única diferença diz respeito à frequência da contradição do tipo "léxico". Essa contradição é uma das menos frequentes no CSTNews (4,26%), ao passo que, no *corpus* de Marnefee (2012), ela é a segunda mais frequente (21,4%).

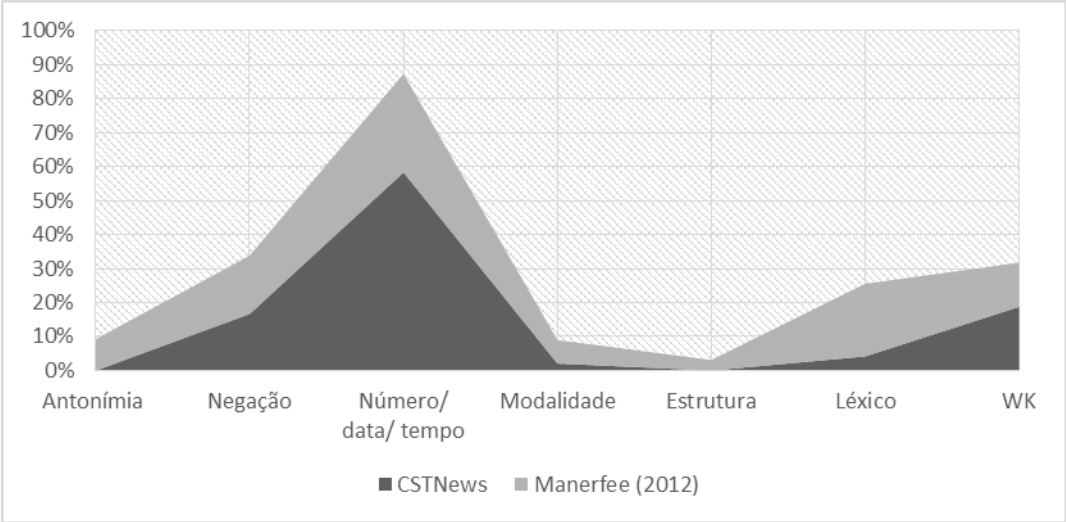

Figura 4: Comparação da ocorrência dos tipos de contradição: CSTNews *vs.* Marnefee (2012).

Apesar de o tipo "WK" ser, juntamente com o tipo "léxico", o segundo mais frequente (19,15%) dentre os casos do CSTNews, ver-se-á, na próxima subseção, que não foram propostos atributos linguísticos para a identificação os casos assim classificados, posto a subjetividade do tipo de contradição e a consequente limitação dos recursos e ferramentas de PLN. Ademais, dada a não ocorrência dos tipos "antonímia" e "estrutura" no *corpus*, também não foram identificados atributos relevantes para a identificação automática dos mesmos.

Assim, a seguir, apresentam-se os atributos linguísticos identificados como relevantes para subsidiar a identificação automática dos seguintes tipos de contradição: "negação", "número/ data/ tempo", "modalidade" e "léxico".

5. Estratégias para a detecção automática da contradição

Com base nos casos tipificados do *corpus*, especificaram-se 8 atributos para a detecção contradição no CSTNews.

5.1. Delimitação de atributos

a) Similaridade nominal

Com base na tipologia das relações CST proposta por Maziero et al (2010), ilustrada na Figura 3, a contradição (codificada pela relação CST *Contradiction*) é um fenômeno multidocumento que ocorre entre o conteúdo veiculado por 2 sentenças. Sabendo-se que as relações CST de conteúdo, como *Contradiction*, ocorrem entre sentenças que possuem certa similaridade ou redundância de conteúdo, previu-se o atributo "similaridade nominal", posto a relevância dos nomes na veiculação do conteúdo sentencial. Tal similaridade pode ser identificada entre 2 sentenças pela aplicação de uma medida estatística como a descrita em (1). No caso, trata-se de uma adaptação da medida clássica *word overlap* em função da classe nominal. Em (1),

tem-se a fórmula para o cálculo da *noun overlap* (Nol). Em (1), vê-se que, para calcular a Nol entre um par de sentenças (S1 e S2) (provenientes de textos distintos), deve-se dividir o número total de nomes idênticos entre as sentenças (*CommonNouns*) pela soma do número total de nomes de cada sentença (*Nouns(S1) + Nouns(S2)*), excluindo-se as *stopwords*[11], números e símbolos). O resultado obtido é entre 0 e 1, sendo que, quanto mais próximo de 1 for a *Nol*, mais similar será o par entre si, e, quanto mais próximo de 0, menos similar.

(1) $$Nol\ (S1,S2) = \frac{\#CommonNouns}{\#Nouns\,(S1) + \#Nouns\,(S2)}$$

b) Similaridade verbal

Seguindo a justificativa do atributo de "similaridade nominal", especificou-se outro atributo lexical, a saber: "similaridade verbal". Essa similaridade também pode ser calculada entre 2 sentenças pela medida *word overlap* adaptada à classe dos verbos, a saber: *verb overlap* (Vol). Em (2), vê-se que, para calcular a Vol entre um par de sentenças (S1 e S2), divide-se o número total de verbos idênticos entre as sentenças (*CommonVerbs*) pela soma do número total de verbos de cada sentença (*Verbs(S1) + Verbs(S2)*). O resultado obtido é entre 0 e 1, sendo que, quanto mais próximo de 1 for a *Vol*, mais similar será o par entre si, e, quanto mais próximo de 0, menos similar.

(2) $$Vol\ (S1,S2) = \frac{\#CommonVerbs}{\#Verbs(S1) + \#Verbs(S2)}$$

c) Sobreposição de números

Tendo em vista a frequência dos casos de contradição que envolvem discrepâncias numéricas, especificou-se o atributo "sobreposição de números". Essa similaridade também pode ser calculada entre 2 sentenças pela medida *word overlap* adaptada aos dados numéricos, a saber: *num overlap* (NUMol). Em (3), vê-se que, para calcular a NUMol entre um par de sentenças (S1 e S2), divide-se o número total de números idênticos entre as sentenças (*CommonNumerals*) pela soma do número total de números de cada sentença (*Numerals(S1) + Numerals(S2)*). O resultado obtido é entre 0 e 1, sendo que, quanto mais próximo de 1 for a *NUMol*, mais similares serão os números do par entre si, e, quanto mais próximo de 0, menos similares.

(3) $$NUMol\ (S1,S2) = \frac{\#CommonNumerals}{\#Numerals\,(S1) + \#Numerals\,(S2)}$$

d) Discrepância de advérbios de negação

Como visto, muitos dos casos de contradição do *corpus* caracterizam-se pela presença de um advérbio de negação em apenas uma das sentenças de um par. Assim, esse atributo se mostra relevante para a identificação da contradição do tipo negação.

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

Os advérbios de negação são palavras que pertencem a uma subclasse dos advérbios e que podem ser modificadores do grupo verbal ou de constituintes do grupo verbal. Tradicionalmente, considerava-se "não" o único advérbio de negação, mas as gramáticas mais atuais já admitem outros: tampouco, nem, nunca, jamais, etc. Há também locuções que funcionam como um advérbio de negação, as quais são denominadas "locuções adverbiais de negação", como: "de modo algum", "de jeito nenhum", "de forma nenhuma".

e) Sobreposição do verbo principal

Como se observou no *corpus*, os casos com contradição do tipo "negação" são comumente formados por sentenças que compartilham o mesmo verbo principal (ainda que em formas flexionadas distintas). Em uma das sentenças do par, o verbo principal é modificado pelo advérbio de negação. Assim, a sobreposição do verbo principal parece um atributo linguístico relevante.

f) Discrepância de modo verbal

Apesar de a contradição do tipo "modalidade" ocorrer apenas 1 vez no *corpus*, foi possível observar que ela se fundamenta na ocorrência de um verbo auxiliar condicional (no caso, "ter") em apenas uma das sentenças de um par, o qual índica uma incerteza frente à afirmação da outra sentença. Assim, delimitou-se o atributo "discrepância de modo verbal". Assim, por meio desse atributo, busca-se ver se as sentenças possuam verbos auxiliares em modos distintos.

g) Discrepância de entidades nomeadas

Alguns dos casos de contradição tipificados como "léxico" ocorrem pela divergência quanto à expressão de entidades nomeadas ("*Air Traset*" vs "*Trapset Congo*"). Assim, propôs-se o atributo "discrepância de entidades nomeadas".

h) Ocorrência de opostos conceituais

Alguns dos casos de contradição tipificados como "léxico" ocorrem pela divergência quanto à ocorrência de opostos conceituais, como "feridos" vs "mortos". Assim, propôs-se o atributo "sobreposição de opostos conceituais".

2.2.5. Caracterização manual do corpus

Essa tarefa consistiu na caracterização das sentenças do *corpus*, ou seja, na descrição dos atributos linguísticos especificados na tarefa anterior. Além dos 40 pares de sentenças contraditórias extraídas no CSTNews, mais 10 pares cujas sentenças não são contraditórias foram caracterizados para fins de comparação. Pretende-se, com isso, verificar no futuro se os atributos detectam pares com contradição e sem contradição.

Assim, discriminou-se os atributos: (i) palavras pertencente à classe dos nomes (N), (ii) palavras pertencentes à classe dos verbos (V); (iii) palavras pertencentes à subclasse dos advérbios de negação (AdvN), (iv) verbo principal (VP), (v) modo (M), (vi) números (Num), (vii) entidades nomeadas (EN) e (vi)

opostos conceituais (OC). No Quadro 5, ilustra-se a caracterização de 3 pares de sentenças do *subcorpus* de contradição e 1 par sem contradição (anotado com a relação CST *Identity*).

Quadro 5: Caracterização das sentenças em função dos atributos da contradição.

Par	CST	N	V	Num.	AdvN	VP	Modo	EN	CO
1	Contradiction (Subtipo WK)	aeronoave, montanha, chama, floresta, quilômetro, distância, pista, aeroporto	chocar, cair	15	X	chocar, cair	X	X	X
		avião, tempo, pista, aterrissagem, floresta, aeroporto	prejudicar, conseguir, chegar, cair	15	não	morrer	X	Bukavu	X
2	Contradiction (Subtipo Léxico)	avião, Air Traset, passageiro, tripulante	acidentar, operar, levar	14, 3	X	levar	X	Air Traset	X
		porta-voz, avião, Soviet Antonov-28, fabricação, propriedade, companhia, Trasept Congo, carga, mineral	informar, levar	X	X	informar	X	Soviet Antonov-28, Trasept Congo	X
3	Contradiction (Subtipo WK)	José, Maria, Eymael, PSDC, Rui, Pimenta, PCO, intenção, voto	chegar, obter	1	não	obter	X	José, Maria, Eymael, PSDC, Rui, Pimenta, PCO	X
		candidato, José, Maria, Eymael, PSDC, Rui, Pimenta, PCO	pontuar	X	não	pontuar	X	José, Maria, Eymael, PSDC, Rui, Pimenta, PCO	X
41	Identity	vítima, acidente, ir, passageiro, membro, tripulação	ir	14, três	X	ser	X	X	X
		vítima, acidente, ir, passageiro, membro, tripulação	ir	14, três	X	ser	X	X	X

5.2. Verificação dos atributos entre as sentenças dos pares

Uma vez que os atributos das sentenças dos pares (50, no total) tenham sido explicitados, poder-se-á calcular a discrepância ou sobreposição dos mesmos entre as sentenças, obtendo uma tabela similar à Tabela 3.

Tabela 3: Cálculo da discrepância ou sobreposição dos atributos entre as sentenças.

Par	CST	N	V	Num	AdvN	VP	Modo	EN	CO
1	Contradiction (Subtipo WK)	0.42	0.33	1	Sim	Não	Não	Sim	Não
2	Contradiction (Subtipo Léxico)	0.15	0.4	0	Não	Não	Não	Sim	Não
3	Contradiction (Subtipo WK)	0.88	0	0	Não	Não	Não	Não	Não
41	Identity	1	1	1	Não	Sim	Não	Não	Não

Na Tabela 3, observa-se que há um grupo de atributos (N, V e Num), cujos valores obtidos no cálculo da sobreposição são numéricos e há outro grupo de atributos (AdvN, VP, Modo, EN e CO) cujos valores obtidos são categóricos, expressos por meio de "sim" ou "não".

A sobreposição dos atributos N, V e Num é expressa por meio de valores numéricos porque resultam da aplicação das variações da medida *word overlap*, como explicado anteriormente. Os valores mais próximos de 1 indicam que as sentenças do par compartilham maior número de nomes, verbos e números e os valores mais próximos de 0 indicam o contrário. Dentre os pares com contradição, as sentenças do par 3 compartilham, por exemplo, o maior número de nomes (0.88) e as do par 2, o menor (0.15). Quanto aos demais atributos AdvN, VP, Modo, EM e CO, os valores categóricos indicam se há (ou não) discrepância ou sobreposição entre as sentenças do par. Por exemplo, entre as sentenças do par 1, há discrepância quanto à "ocorrência de advérbio de negação", já que uma das sentenças possui o advérbio "não" e a outra não possui (cf. Quadro 7).

6. Considerações Finais

Sobre a investigação da contradição, ressalta-se que se sabe mais sobre esse fenômeno no *corpus* CSTNews do que antes da realização deste projeto. Especificamente, tem-se hoje a tipificação dos casos de contradição e, sobretudo, a especificação de um conjunto de atributos potencialmente relevantes para subsidiar a detecção automática da contradição em *corpus* multidocumento, ao menos naqueles com as mesmas características do CSTNews.

Pretende-se realizar, como trabalho futuro, o aprendizado e a avaliação das regras de detecção da contradição. Essas tarefas darão continuidade ao trabalho realizado neste projeto.

Agradecimentos

Os autores agradecem à FAPESP pelo apoio financeiro.

Referências bibliográficas

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

CAMARGO, R. T. **Investigação de estratégias de sumarização humana multidocumento**. São Carlos, 2013. 117p. Dissertação (Mestrado em Linguística) – Departamento de Letras, Universidade Federal de São Carlos, São Carlos, 2013.

CARDOSO, P.C.F., et al. CSTNews - A discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese. In: RST Brazilian Meeting, 3, 2011, Cuiabá, Brasil. **Proceedings…** Cuiabá, 2011, p. 88-105.

CONDORAVDI, C.; CROUCH, D.; PAVIA, V. DE; STOLLE, R.; BOBROW, D. G. **Entailment, intensionality and text understanding**. In: HLT-NAACL- WORKSHOP ON TEXT MEANING, 2003, Edmonton, Canada. **Proceedings...** 2003, p. 38-45.

CHIERCHIA, G. **Semântica.** Campinas-SP & Londrina-PR: EdUnicamp & EdUEL, 2003.

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

FELLBAUM, C. (Ed.). **WordNet: an electronic lexical database**. Ca, MA: MIT Press, 1998.

HARABAGIU, S.; HICKL, A.; LACATUSU, F. Negation, contrast, and contradiction in text processing. In: NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE, 21, 2006, Boston, Massachusetts. **Proceedings...** 2006.

KUMAR, Y.J.; SALIM, N. Automatic multi-document summarization approaches. **Journal of Computer. Science**, 8, 2012, p. 133-140.

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

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

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

MARNEFFE, M-C DE.; RAFFERTY, A. N; MANNING, C. D. Finding contradictions in text. In: ANNUAL MEETING OF THE ACL, 46, 2008. Columbus, Ohio, USA. **Proceedings...** Ohio, 2008, p. 1039-1047.

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

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

___; JORGE, M. L. C.; PARDO, T. A. S. Identifying Multidocument Relations. In: NLPCS, 7, 2010, Funchal, PT. **Proceedings...** Funchal, 2010, p. 60-69.

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

RADEV, D. A common theory of information fusion from multiple text sources, step one: cross-document structure". In: ACL Signal Workshop on Discourse and Dialogue, 1, 2000, Hong Kong, **Proceedings…** Hong Kong, 2000, p. 74-83.

RADEV, D. R. *et al.* MEAD-a platform for multidocument multilingual text summarization. In: International Conference on Language Resources and Evaluation (LREC), 4, 2004, Lisbon. **Proceedings…** Lisbon, 2004, **Proceeings…** p. 1-4.

SINCLAIR, J. Corpus and text: basic principles. In: Wynne, M. (Ed.). **Developing linguistic corpora: a guide to good practice**. Oxford: Oxbow Books, 2005. P.1-16. Disponível em: <www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter1.htm>. Acesso em: 02 ago. 2010.

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

Anexo 1 – Pares com *contradição* do CSTNews

No.	Pares de sentenças
1	A aeronave se chocou com uma montanha e caiu, em chamas, sobre uma floresta a 15 quilômetros de distância da pista do aeroporto.
	Todos morreram quando o avião, prejudicado pelo mau tempo, não conseguiu chegar à pista de aterrissagem e caiu numa floresta a 15 Km do aeroporto de Bukavu.
2	O avião acidentado, operado pela Air Traset, levava 14 passageiros e três tripulantes.
	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.
3	José Maria Eymael, do PSDC, e Rui Pimenta, do PCO, não chegaram a obter 1% das intenções de voto.
	Os candidatos José Maria Eymael (PSDC) e Ruy Pimenta (PCO) não pontuaram.
4	Num eventual segundo turno e na comparação com a pesquisa Ibope do dia 25, Lula oscila de 48% para 50%, enquanto Alckmin cai de 39% para 36%.
	No segundo turno, as intenções de voto do presidente Lula caíram de 53% em junho para 50% em julho, enquanto o candidato Alckmin subiu de 29% para 36%.
5	Em nota enviada após a exibição da reportagem, a TAM afirma "que não teve registro de qualquer problema mecânico neste avião no dia 16 de julho".
	Um dia antes do acidente, na segunda-feira, 16, o avião também teria apresentado problemas ao aterrissar em Congonhas, durante o vôo 3215, procedente de Belo Horizonte (Confins), só conseguindo parar muito próximo do final da pista.
6	O congestionamento esteve ainda maior às 9h, quando chegou a 113 km de extensão para uma média de 32 km.
	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).
7	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).
	O pico de lentidão foi registrado às 9h, com 113 km de lentidão, o dobro do registrado neste horário.
8	Ao menos 300 policiais de Amapá, Distrito Federal, Mato Grosso, Acre e Rondônia trabalharam na Operação Dominó.
	Mais de 300 policiais federais do Amazonas, Distrito Federal, Mato Grosso, Acre e Rondônia participaram das buscas e prisões.
9	O prédio da secretaria da Fazenda, no centro, foi atingido por três bombas caseiras.
	A Secretaria da Fazenda também foi atingida por uma bomba.
10	O prédio da secretaria da Fazenda, no centro, foi atingido por três bombas caseiras.
	A Secretaria da Fazenda também foi atingida por uma bomba.
11	Até o momento, as autoridades do Sri Lanka não confirmaram as mortes ou esclareceram o que acontece na cidade de Muttur.
	Quinze funcionários locais de uma organização de caridade francesa no Sri Lanka foram encontrados mortos na cidade de Muttur, no norte do país.
12	O diretor da ACF no Sri Lanka, Benoit Miribel, confirmou a morte de seus funcionários e afirmou, comovido, que a ONG "não sofreu uma perda similar em seus mais de 25 anos de existência".
	Até o momento, as autoridades do Sri Lanka não confirmaram as mortes ou esclareceram o que acontece na cidade de Muttur.
13	Até o momento, as autoridades do Sri Lanka não confirmaram as mortes ou esclareceram o que acontece na cidade de Muttur.
	O diretor da ACF no Sri Lanka, Benoit Miribel, confirmou a morte de seus funcionários e afirmou, comovido, que a ONG "não sofreu uma perda similar em seus mais de 25 anos de existência".
14	O diretor da ACF no Sri Lanka, Benoit Miribel, confirmou a morte de seus funcionários e afirmou, comovido, que a ONG "não sofreu uma perda similar em seus mais de 25 anos de existência".
	Até o momento, as autoridades do Sri Lanka não confirmaram as mortes ou esclareceram o que acontece na cidade de Muttur.
15	Até o momento, as autoridades do Sri Lanka não confirmaram as mortes ou esclareceram o que acontece na cidade de Muttur.
	O diretor da ACF no Sri Lanka, Benoit Miribel, confirmou a morte de seus funcionários e afirmou, comovido, que a ONG "não sofreu uma perda similar em seus mais de 25 anos de existência".
16	Cairo - O ministro da Saúde egípcio, Hatem El-Gabaly, informou nesta segunda-feira que 57 pessoas morreram e 128 ficaram feridas no choque entre dois trens de passageiros no delta do Nilo, ao norte do Cairo.

	CAIRO - Pelo menos 80 pessoas morreram e mais de 165 ficaram feridas nesta segunda-feira após a colisão de dois trens de passageiros no delta do Nilo, ao norte do Cairo, informaram fontes policiais e médicas.
17	Cairo - O ministro da Saúde egípcio, Hatem El-Gabaly, informou nesta segunda-feira que 57 pessoas morreram e 128 ficaram feridas no choque entre dois trens de passageiros no delta do Nilo, ao norte do Cairo.
	No entanto, o ministro da Saúde, Hatem El-Gabaly, insistiu que até o momento foram recuperados apenas 36 cadáveres e que 133 feridos foram encaminhados a hospitais da região.
18	Cairo - O ministro da Saúde egípcio, Hatem El-Gabaly, informou nesta segunda-feira que 57 pessoas morreram e 128 ficaram feridas no choque entre dois trens de passageiros no delta do Nilo, ao norte do Cairo.
	Uma colisão entre dois trens de passageiros provocou a morte de pelo menos 80 pessoas e deixou 165 feridas.
19	Os feridos, vários deles em estado grave, foram levados para seis hospitais da região onde aconteceu o acidente, um dos piores ocorridos no Egito desde 2002, disse El-Gabaly, citado pela agência egípcia de notícias Mena.
	O acidente ocorreu no delta do Nilo, ao norte de Cairo, no Egito.
20	Nove pessoas morreram, três delas crianças, e outras 25 ficaram feridas nesta segunda-feira em uma explosão ocorrida em um mercado de Moscou, informou a polícia.
	MOSCOU (Rússia) - Uma explosão causou nesta segunda-feira a morte de dez pessoas e deixou cerca de 30 feridas no mercado Cherkizov de Moscou.
21	Nove pessoas morreram, três delas crianças, e outras 25 ficaram feridas nesta segunda-feira em uma explosão ocorrida em um mercado de Moscou, informou a polícia.
	- Por causa da explosão, morreram dez pessoas e 31 foram hospitalizadas - declarou à agência "Interfax" o alto funcionário de Moscou, Vladimir Resin.
22	A maioria dos feridos, entre os quais há quatro com menos de 18 anos, foi hospitalizada.
	Anteriormente, a Polícia havia informado sobre nove mortos, sendo três deles crianças, e 25 feridos.
23	- Por causa da explosão, morreram dez pessoas e 31 foram hospitalizadas - declarou à agência "Interfax" o alto funcionário de Moscou, Vladimir Resin.
	MOSCOU (Rússia) - Nove pessoas morreram, sendo três crianças, e outras 25 ficaram feridas nesta segunda-feira em uma explosão registrada em um mercado moscovita, informou a Polícia de Moscou.
24	A bomba detonou no interior de uma cafeteria localizada no setor denominado "Evrazia" do mercado Cherkizov.
	A explosão, cujas causas ainda são desconhecidas, aconteceu às 10h40 (3h40 em Brasília) no mercado Cherkizov, localizado no nordeste da capital russa.
25	Os deputados acusados de envolvimento na máfia dos sanguessugas têm até a meia-noite desta segunda-feira para renunciar aos mandatos se quiserem escapar da cassação.
	Termina hoje, às 20 horas, o prazo para que os deputados acusados de participar do esquema dos sanguessugas renunciem para escapar da abertura de processo por quebra de decoro parlamentar.
26	Termina hoje, às 20 horas, o prazo para que os deputados acusados de participar do esquema dos sanguessugas renunciem para escapar da abertura de processo por quebra de decoro parlamentar.
	Os parlamentares acusados de envolvimento com o esquema de superfaturamento das ambulâncias têm até a meia-noite desta segunda-feira, dia 21, para renunciar ao mandato, caso queiram escapar do julgamento por quebra de decoro.
27	WASHINGTON - Ao menos 33 pessoas morreram nesta segunda-feira, 16, segundo o presidente da instituição, Charles Steger.
	Um ataque em dois lugares da Universidade Técnica da Virgínia, em Blacksburg, Estados Unidos, resultou na morte de 30 pessoas.
28	Um atirador matou ao menos 30 pessoas em dois diferentes locais da Universidade Técnica da Virgínia, em Blacksburg (Virgínia), nesta segunda-feira, no pior ataque a tiros contra um campus universitário da história dos Estados Unidos.
	WASHINGTON - Ao menos 33 pessoas morreram nesta segunda-feira, 16, segundo o presidente da instituição, Charles Steger.
29	Segundo Aimee Kanode, aluna do primeiro ano, o primeiro ataque ocorreu no primeiro andar do West Ambler Johnston, um andar acima do quarto em que ela dorme.
	Aimee Kanode, uma estudante do primeiro ano, afirmou que ouviu tiros vindos do 4º andar do prédio de dormitórios, um andar acima do seu.
30	Na terceira parte da reforma, parte dos vôos de Cumbica serão transferidos para o Aeroporto de Viracopos, em Campinas.
	Enquanto durarem as obras, os vôos serão transferidos para o Aeroporto de Viracopos, em Campinas, a 95 km da Capital, segundo informa o ministério de Nelson Jobim pro meio de nota.
31	Por conta do fechamento, das 39 partidas previstas, 19 foram canceladas, mas a Empresa Brasileira de Infra-

	Estrutura Aeroportuária (Infraero) não contabilizava atrasos superiores a uma hora, mas quase nenhum vôo saia no horário programado do aeroporto.
	De acordo com a Empresa Brasileira de Infra-Estrutura Aeroportuária (Infraero), 30 vôos foram cancelados das 57 partidas programadas, até 09h, e dois com atraso superior a 1h.
32	Segundo a Infraero, até às 8h havia registros de 21 vôos atrasados e de 16 cancelamentos.
	Segundo a Infraero, nenhuma foi cancelada.
33	RIO - Depois da queda de April Steiner, a brasileira Fabiana Murer leva a medalha de ouro no salto com vara, com 4m50 - novo recorde pan-americano.
	Com a marca de 4m60, Fabiana não só venceu a prova, como também estabeleceu o novo recorde pan-americano, 20cm mais alto do que a antiga marca de 4m40.
34	Murer - vice-campeã da Copa do Mundo de 2006 - conquistou o lugar mais alto do pódio com a marca de 4m60, contra 4m40 da norte-americana April Steiner, que ficou com a prata.
	RIO - Depois da queda de April Steiner, a brasileira Fabiana Murer leva a medalha de ouro no salto com vara, com 4m50 - novo recorde pan-americano.
35	Em desvantagem no placar, os argentinos ficaram com a bola e pressionaram.
	Com mais posse de bola, o Brasil seguiu pressionando.
36	Ao Brasil restavam os contragolpes.
	Com mais posse de bola, o Brasil seguiu pressionando.
37	Aos 26 minutos, a torcida xingava e pedia Obina na seleção, quando Kaká chutou forte de longe e Ronaldinho Gaúcho deu uma leve desviada na bola, enganando o goleiro equatoriano.
	Aos 27, Kaká arriscou de muito longe e Ronaldinho colocou o desviou o chute.
38	No rebote, Elano fez o quarto gol, aos 38min.
	A 20cm da linha de fundo ele deu dois dribles humilhantes no zagueiro equatoriano e cruzou para Elano, que fez o quarto, aos 37.
39	Entre outras conquistas, a seleção nacional soma dois títulos do Campeonato Mundial, uma medalha de ouro nos Jogos Olímpicos de Atenas-04 e a Copa do Mundo do Japão, em 2003.
	A conquista reforça o status de melhor time de vôlei do mundo, uma vez que o Brasil soma cinco títulos consecutivos da Liga Mundial, além do título da Copa do Mundo e das Olimpíadas de Atenas, em 2004.
40	O terremoto deixou até o momento nove mortos, todos idosos, mais de mil feridos e cerca de 13 mil desabrigados, assim como grandes danos materiais.
	Ao menos 9 pessoas morreram e cerca de 700 se feriram.
41	ACM já tinha sofrido infarto em 1989 e já tinha recebido três pontes de safena.
	Em 1989, ACM sofreu um infarto e, operado pelo cardiologista Adib Jatene, recebeu o implante de duas pontes de safena e duas mamárias.
42	O grande destaque da prova foi Nicolas Oliveira, quarto nadador brasileiro a cair na água.
	Thiago, que abriu o revezamento, foi o grande destaque do quarteto nas piscinas do Parque Aquático Maria Lenk.
43	O presidente do Senado argumenta que o Conselho de Ética da Casa não tem poderes para investigar seus documentos.
	Com base na resolução que disciplina a atuação do Conselho de Ética, podem assumir as investigações sem precisarem recorrer à Polícia Federal ou esperar uma definição do Supremo Tribunal Federal sobre um eventual recurso do presidente do Congresso.
44	O tremor atingiu a região às 10h13 (horário local, 22h13 de domingo, em Brasília) e seu epicentro foi localizado a 260 km da costa de Niigata, ao nordeste da capital, Tóquio, onde também foi sentido.
	Duas mulheres na faixa dos 80 anos morreram quando suas casas ruíram durante o tremor de magnitude 6,8, na área de Niigata, cerca de 250 km noroeste de Tóquio, informou a imprensa japonesa.
45	Quatro pessoas morreram e cerca de 400 ficaram feridas.
	Mais de 600 pessoas ficaram feridas, casas foram derrubadas e houve um pequeno incêndio na maior usina nuclear do mundo.
46	Mesmo com altos e baixos, a equipe brasileira não teve muito trabalho para superar o fraco time do Canadá por 3 sets a 0 (25/18, 25/19 e 25/17), em uma hora e dez minutos.
	No resultado do jogo, deu a lógica: o Brasil venceu por 3 sets a 0, com parciais de 25/19, 25/18 e 25/17.