Programa de Pós-Graduação em Lingüística

PPGL

APLICAÇÃO DE CONHECIMENTO LÉXICO-CONCEITUAL NA SUMARIZAÇÃO MULTIDOCUMENTO MULTILÍNGUE

FABRICIO ELDER DA SILVA TOSTA

Universidade Federal de São Carlos

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE EDUCAÇÃO E CIÊNCIAS HUMANAS
PROGRAMA DE PÓS-GRADUAÇÃO EM LINGUÍSTICA

APLICAÇÃO DE CONHECIMENTO LÉXICO-CONCEITUAL NA SUMARIZAÇÃO MULTIDOCUMENTO MULTILÍNGUE

FABRICIO ELDER DA SILVA TOSTA
Bolsista: CAPES

Dissertação apresentada ao Programa de Pós-Graduação em Linguística da Universidade Federal de São Carlos para o Exame de Defesa, como parte dos requisitos para a obtenção do título de Mestre em Linguística.

Orientadora: Profa. Dra. Ariani Di Felippo

Coorientador: Prof. Dr. Thiago A. S. Pardo

São Carlos - São Paulo – Brasil

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

T716ac

Tosta, Fabricio Elder da Silva.
Aplicação de conhecimento léxico-conceitual na sumarização multidocumento multilíngue / Fabricio Elder da Silva Tosta. -- São Carlos : UFSCar, 2015.
116 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2014.

1. Linguística. 2. Sumarização automática. 3. Sumarização multidocumento multilíngue. 4. Conhecimento léxico-conceitual. 5. Estratégias de seleção de conteúdo. I. Título.

CDD: 410 (20ª)

UNIVERSIDADE FEDERAL DE SÃO CARLOS
Centro de Educação e Ciências Humanas
Programa de Pós-Graduação em Linguística

BANCA EXAMINADORA DA DISSERTAÇÃO DE MESTRADO DE FABRICIO ELDER DA SILVA TOSTA

Profa. Dra. Ariani Di Felippo
Orientadora e Presidente
UFSCar/São Carlos

Profa. Dra Maria das Graças Volpe Nunes
Membro titular
USP/ São Carlos

Profa. Dra. Gladis Maria de Barcellos Almeida
Membro titular
UFSCar/São Carlos

Submetida a defesa pública em sessão realizada em: 27/fevereiro/2014.
Homologada na 64ª reunião da CPGL, realizada em 25/02/2014.

Carlos Piovezani
Coordenador
PPGL/UFSCar

À minha amada mãe, com quem aprendi que,
na vida, não devemos jamais desistir dos desafios,
por mais difíceis que pareçam,
sobretudo, quando esse desafio for o desafio pela própria vida.

Agradecimentos

À minha mãe, Célia e meu pai, Beto, pelo amor e apoio incondicionais, por cada palavra de sabedoria, incentivo e confiança.

À minha querida e inesquecível avó, Antônia, que sempre torceu por mim.

Às minhas amigas de todas as horas, Paula, Amanda e Renata, pela alegria e prazer da companhia e por toda ajuda imensurável.

Aos amigos e professores da UFSCar, pela amizade e apoio.

À minha orientadora Ariani Di Felippo e ao meu coorientador Thiago A. S. Pardo, pelos ensinamentos e paciência constantes.

Aos integrantes do NILC, pela amizade e pelos anos de companheirismo no laboratório, sempre com a usual disposição em ajudar. Em especial às queridas amigas Lucía e Lianet.

Aos colegas e amigos que colaboraram para este trabalho pela paciência e dedicação, nas consecutivas tarefas em grupo: Paula, Márcio, Verônica, Amanda, Renata, Erick, Fernando, Lucía, Jackson, Pedro, Matheus, Andressa, Rafael, Jader e Claudinha.

À minha amiga do peito, Ana Paula Romano, sempre pronta a ajudar, apesar da distância física, com suas palavras de carinho e apoio.

Às minhas queridas amigas Ana Lígia e Ana Paula Cavaguti, pela amizade e carinho.

E por fim, à CAPES, pelo apoio financeiro.

RESUMO

Tradicionalmente, a Sumarização Automática Multidocumento Multilíngue (SAMM) é uma aplicação que, a partir de uma coleção de textos sobre um mesmo assunto em ao menos duas línguas distintas, produz um sumário (extrato) informativo e genérico em uma das línguas-fonte. Os métodos mais simples realizam a tradução automática (TA) dos textos-fonte e, a partir de uma coleção monolíngue, aplicam estratégias superficiais e/ou profundas de seleção de conteúdo. Dessa forma, a SAMM precisa não só identificar a informação principal da coleção para compor o sumário, evitando-se a redundância, mas também lidar com os problemas causados pela TA integral dos textos-fonte. Buscando alternativas para esse cenário, investigaram-se dois métodos (Método 1 e 2) que, uma vez pautados em conhecimento profundo do tipo léxico-conceitual, evitam a TA integral dos textos-fonte, gerando sumários informativos e coesos/coerentes. Neles, a seleção do conteúdo tem início com a pontuação e o ranqueamento das sentenças originais em função da frequência de ocorrência na coleção dos conceitos expressos por seus nomes comuns. No Método 1, apenas as sentenças mais bem pontuadas na língua do usuário e não redundantes entre si são selecionadas para compor o sumário até que se atinja a taxa de compressão. No Método 2, as sentenças originais mais bem ranqueadas e não redundantes entre si são selecionadas para compor o sumário sem que se privilegie a língua do usuário; caso sentenças que não estejam na língua do usuário sejam selecionadas, estas são automaticamente traduzidas. Para a produção dos sumários automáticos segundo os Métodos 1 e 2 e subsequente avaliação dos mesmos, construiu-se o *corpus* CM2News, que possui 20 coleções de notícias jornalísticas, cada uma delas composta por 1 texto original em inglês e 1 texto original em português sobre um mesmo assunto. Os nomes comuns do CM2News foram identificados via anotação morfossintática e anotados com os conceitos da WordNet de Princeton de forma semiautomática, ou seja, por meio do editor gráfico MulSen desenvolvido para a tarefa. Para a produção dos sumários segundo o Método 1, somente as sentenças em português mais bem pontuadas foram selecionadas até que se atingisse determinada taxa de compressão. Para a produção dos sumários segundo o Método 2, as sentenças mais pontuadas foram selecionadas sem privilegiar a língua do usuário. Caso as sentenças selecionadas estivessem em inglês, estas foram automaticamente traduzidas para o português pelo tradutor Bing. Os Métodos 1 e 2 foram avaliados de forma intrínseca, considerando-se a qualidade linguística e a informatividade dos sumários. Para avaliar a qualidade linguística, 15 linguistas computacionais analisaram manualmente a gramaticalidade, a não-redundância, a clareza referencial, o foco e a estrutura/coerência dos sumários e, para avaliar a informatividade, os sumários foram automaticamente comparados a sumários de referência pelo pacote de medidas ROUGE. Em ambas as avaliações, os resultados evidenciam o melhor desempenho do Método 1, o que pode ser justificado pelo fato de que as sentenças selecionadas são provenientes de um mesmo texto-fonte. Além disso, ressalta-se o melhor desempenho dos dois métodos baseados em conhecimento léxico-conceitual frente aos métodos mais simples de SAMM, os quais realizam a TA integral dos textos-fonte. Por fim, salienta-se que, além dos resultados promissores sobre a aplicação de conhecimento léxico-conceitual, este trabalho gerou recursos e ferramentas importantes para a SAMM, como o *corpus* CM2News e o editor MulSen.

Palavras-chave: Sumarização Multidocumento Multilíngue. Conhecimento léxico-conceitual. Seleção de conteúdo.

ABSTRACT

Traditionally, Multilingual Multi-document Automatic Summarization (MMAS) is a computational application that, from a single collection of source-texts on the same subject/topic in at least two languages, produces an informative and generic summary (extract) in one of these languages. The simplest methods automatically translate the source-texts and, from a monolingual collection, apply content selection strategies based on shallow and/or deep linguistic knowledge. Therefore, the MMAS applications need to identify the main information of the collection, avoiding the redundancy, but also treating the problems caused by the machine translation (MT) of the full source-texts. Looking for alternatives to the traditional scenario of MMAS, we investigated two methods (Method 1 and 2) that once based on deep linguistic knowledge of lexical-conceptual level avoid the full MT of the source-texts, generating informative and cohesive/coherent summaries. In these methods, the content selection starts with the score and the ranking of the original sentences based on the frequency of occurrence of the concepts in the collection, expressed by their common names. In Method 1, only the most well-scored and non redundant sentences from the user's language are selected to compose the extract, until it reaches the compression rate. In Method 2, the original sentences which are better ranked and non redundant are selected to the summary without privileging the user's language; in cases which sentences that are not in the user's language are selected, they are automatically translated. In order to producing automatic summaries according to Methods 1 and 2 and their subsequent evaluation, the CM2News corpus was built. The corpus has 20 collections of news texts, 1 original text in English and 1 original text in Portuguese, both on the same topic. The common names of CM2News were identified through morphosyntactic annotation and then it was semiautomatically annotated with the concepts in Princeton WordNet through the Mulsen graphic editor, which was especially developed for the task. For the production of extracts according to Method 1, only the best ranked sentences in Portuguese were selected until the compression rate was reached. For the production of extracts according to Method 2, the best ranked sentences were selected, without privileging the language of the user. If English sentences were selected, they were automatically translated into Portuguese by the Bing translator. The Methods 1 and 2 were evaluated intrinsically considering the linguistic quality and informativeness of the summaries. To evaluate linguistic quality, 15 computational linguists analyzed manually the grammaticality, non-redundancy, referential clarity, focus and structure / coherence of the summaries and to evaluate the informativeness of the sumaries, they were automatically compared to reference sumaries by ROUGE measures. In both evaluations, the results have shown the better performance of Method 1, which might be explained by the fact that sentences were selected from a single source text. Furthermore, we highlight the best performance of both methods based on lexical-conceptual knowledge compared to simpler methods of MMAS, which adopted the full MT of the source-texts. Finally, it is noted that, besides the promising results on the application of lexical-conceptual knowledge, this work has generated important resources and tools for MMAS, such as the CM2News corpus and the Mulsen editor.

Keywords: Multilingual Multi-document Automatic Summarization. Lexical-conceptual knowledge. Content selection.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE SIGLAS

Adj – Adjetivo
AM – Aprendizado de Máquina
C – *Cluster* ou coleção
CM2News – *Corpus* Multidocumento Bilíngue de Textos Jornalísticos
CM3News – Corpus Multidocumento Trilíngue de Textos Jornalísticos
CST – *Cross-document Structure Theory*
D – Documento
DEMS – *Dissimilarity Engine for Multi-document Summarization*
DUC – *Document Understanding Conference*
IDC – *International Data Corporation*
IN – Inglês
L – Língua
MEAD – *Multi-document Summarization Environment*
MINDS – *Multi-lingual Interactive Document Summarization*
MulSen – *Multilingual Sense Estimator from NILC*
MUSE – *Multi-lingual Sentence Extractor*
N – Nome
NILC – Núcleo Interinstitucional de Linguística Computacional
PT – Português
PLN – Processamento de Língua Natural
ROUGE – *Recall-Oriented Understudy of Gisting Evaluation*
RST – *Rethorical Structure Theory*
SA – Sumarização Automática
SAM – Sumarização Automática Multidocumento
SAMM – Sumarização Automática Multidocumento Multilíngue
SN – Sintagma Nominal
SNs – Sintagmas Nominais
SPrep – Sintagma preposicional
SUCINTO – *Summarization for Clever Information Access*
SUMMAC – *Text Summarization Evaluation Conference*
SUSTENTO – *Generation of Linguistic Knowledge for Multi-document Summarization*
Synset – Synonym set (conjunto de formas sinônimas)
TA – Tradução automática
TAC – *Text Analysis Conference*
WN.Pr – WordNet de Princeton
Wol – Word Overlap

ÍNDICE

1 INTRODUÇÃO

1.1 Contextualização

Devido ao uso generalizado da *web*, cresce cada vez mais o volume de informação disponível aos internautas. Segundo o informe da *International Data Corporation* (IDC) (GANTZ, REINSEL, 2011), foram disponibilizados aproximadamente 1.8 zettabytes[1] de informação na *web* em 2011, quantidade nove vezes maior que a produzida em 2006.

Em resumo, trata-se da era da "explosão da informação", sendo que grande parte dessa informação circula em formato textual e em diferentes línguas, e é produzida e veiculada por agências de notícias *on-line*, *blogs*, *microblogs* e redes sociais (p.ex.: *Facebook*, Orkut, *Myspace*, *Twitter*)

Nesse cenário, reconhece-se cada dia mais, a necessidade do desenvolvimento de aplicações de SA (Sumarização Automática) capazes de lidar com o crescente volume de informação e, sobretudo, com a multiplicidade de idiomas, permitindo o acesso às informações na língua do usuário ou em uma língua na qual seja proficiente. Em, especial, a tarefa de SA permite que, a partir de uma fonte, sejam retiradas informações mais importantes de forma a apresentá-las de forma condensada e sensitiva a um usuário e/ou aplicação (MANI, MAYBURY, 1999).

O desenvolvimento da SA em contexto multilíngue já era foco dos pesquisadores do PLN no final da década de 1990, como evidencia o trabalho de Cowie et al.(1998), mas tem recebido maior atenção nos últimos anos (p.ex.: EVANS et. al., 2004; SAGGION, 2006; EVANS et. al., 2005; FUNG, NGAI, 2006; GEY et al., 2006; STEIBERGER, TURCHI, 2012; TOSTA et. al., 2013).

No contexto multilíngue, as aplicações de SA podem ser desenvolvidas com base em 3 abordagens distintas, as quais classificam os métodos/sistemas em: (i) *cross-language,* (ii) multilíngue e (iii) independentes de língua (ORĂSAN, 2009). Nelas, tem-se focado a produção de extratos (isto é, sumários compostos por sentenças extraídas integralmente dos textos-fonte) informativos (isto é, que veiculam a ideia central dos textos-fonte a ponto de substituir a leitura dos originais) e genéricos (isto é, voltados para uma audiência ampla).

[1] Unidade de medida de informação que corresponde aproximadamente a 2^{70} Bytes.

Os métodos/sistemas *cross-language* caracterizam-se pela sumarização de um texto-fonte ou de um conjunto de textos-fonte em uma língua de entrada L(x) e geração do sumário em uma língua de saída L(y), podendo ser mono e multidocumento, p. ex: (ORĂSAN; CHIOREAN, 2008), (WAN; LI; XIAO, 2010) e (BOURDIN et al., 2011).

Os métodos/sistemas multilíngue caracterizam-se por serem multidocumento; daí, tem-se a SA multidocumento multilíngue (SAMM). Em especial, eles partem obrigatoriamente de um conjunto composto pelo menos por um texto em uma Lx e outro texto em uma língua Ly que abordam o mesmo assunto e geram o sumário em uma dessas línguas-fonte (Lx ou Ly) (p.ex.: EVANS et al., 2004).

Os métodos/sistemas independentes de língua diferenciam-se dos *cross-language* e multilíngue pelo fato de não se basearem em conhecimento linguístico (LITVAK et al*.,* 2010). Geralmente, a seleção do conteúdo que irá compor o sumário é feita com base em características que, idealmente, podem ser aplicadas a qualquer língua natural. Nesses casos, tais métodos pautam-se em conhecimento empírico-estatístico (p.ex.: COWIE et al*.,* 1998, RADEV et al., 2004, LITVAK et al*.,* 2010).

Os métodos/sistemas *cross-language* englobam uma etapa central, que é a etapa de tradução automática (TA). Nesses métodos, a TA pode ser realizada de duas formas, antes ou depois do processo de extração de conteúdo.

Na primeira delas, os textos-fonte são sumarizados com base em métodos multidocumento existentes para a língua-fonte e, na sequência, o sumário gerado ainda na língua-fonte é automaticamente traduzido para a língua-alvo. Segundo Wan, Li e Xiao (2010), esse tipo de SA é denominada *late translation.*

Na outra, os textos-fonte são traduzidos automaticamente para a língua-alvo e, posteriormente, sumarizados com base em métodos exclusivamente multidocumento desenvolvidos para a língua-alvo em questão. Segundo Wan, Li e Xiao (2010), esse tipo de SA é denominada *early translation.*

Tanto em uma abordagem como na outra*,* tem-se o(s) texto(s)-fonte a serem sumarizados em uma única língua de partida. Assim, pode-se aplicar qualquer dos métodos de SA mono e multidocumento baseados em conhecimento linguístico simples (p.ex.: localização das sentenças nos textos-fonte) ou em conhecimento empírico/estatístico (cf. GUPTA, LEHAL, 2010). Os métodos com conhecimento empírico/estatístico possuem a vantagem de poder contar com um processamento mais independente de conhecimento linguístico. No caso dos métodos baseados em conhecimento linguístico mais profundo (sintático, semântico e discursivo) (cf.

KUMAR, SALIM, 2012), sua aplicação dependerá da disponibilidade de recursos linguísticos (p.ex.: léxicos e *corpora*) e ferramentas (p.ex.: *parsers* sintáticos e discursivos) capazes de viabilizar o processamento da língua de entrada.

Potencialmente, os métodos da abordagem multilíngue também podem englobar uma etapa de TA, que é realizada antes do processo de SA, pois, a partir de uma coleção de partida composta por mais de uma língua, objetiva-se uniformizar os textos-fonte em uma única língua. Por ser exclusivamente multidocumento, qualquer método de Sumarização Automática Multidocumento (SAM) poderá ser aplicado.

Quando realizam a etapa de TA, os métodos/sistemas de SA *cross-language* e SAMM herdam os da TA. Autores como Vilar et al. (2006); Popovic; Burchardt (2011) e Martins; Caseli, (2013) já categorizaram e/ou utilizaram tipologias variadas de erros de TA nos diferentes níveis linguísticos. Martins et al. (2013) descreve uma tipologia de erros, divididos em diferentes categorias, a saber:

a) **Erros morfossintáticos**: englobam apenas uma palavra, cujo lema é correto, mas a forma superficial errada. Envolvem concordância de número, gênero, flexão verbal, mudança de categoria lexical, entre outras (p. ex. "Na sexta-feira, o vídeo de uma manifestantes pacíficos [...]" (C6) e "Brian Flynn do Eurocontrol explica como a nuvem de cinzas está sendo monitorizada [...] (C9).[2]

b) **Erros lexicais**: englobam apenas uma forma lexical que não compartilha a forma base (lema) com nenhuma palavra de referência. Esses erros englobam (i) palavras "extras" (que não possuem nenhuma correspondência na sentença original ou fonte, (ii) palavras ausentes (que existem na sentença original, mas não possuem correspondentes na sentença traduzida), (iii) palavras não traduzidas, (iv) palavras incorretamente traduzidas e (v) palavras com erro de grafia (ou seja, a tradução correspondente acontece, mas há algum tipo de erro de grafia) (p. ex. "Em uma opinião concurring, justiça Antonin Scalia escreveu a porção do DNA[...]" (C11) e "O Tribunal revogou patentes detidas por uma firma baseada em Utah em dois genes ligados ao cancro da mama e ovário [...] (C11)).

[2] Os exemplos de erros de tradução automática foram retirados de sumários gerados pelo Método 2, investigado neste trabalho (c.f. 4.1.2, pág. 85), a partir de sua aplicação ao *corpus* CM2News.

c) **N-grama errado**: engloba várias palavras que formam uma expressão, seja ela semântica ou não. Esse tipo de erro é comum quando o sistema de TA não reconhece fraseologias e a tradução destas ocorrem de maneira inadequada, geralmente de forma literal (p. ex. "Uma porta-a-busca de porta da área danificada começou na manhã de segunda-feira [...]" (C9) e "Correspondente do BBC Sport, futebol sul-americano Tim Vickery disse a BBC World Service : sociedade brasileira foi explicitamente disse em 2007 que todo o dinheiro gasto com estádios seria dinheiro privado [...] (C17) e "Nos EUA, os comentadores têm também pegou em tom mais escuro do filme e a falta de humor [...]", (C 19)).

d) **Ordem errada**: o erro engloba uma ou mais palavras com erros que não se enquadram nas categorias anteriores. A ordem das palavras na sentença traduzida é incorreta em relação à sua sentença de referência (p.ex. "*Ministério das relações exteriores do Reino Unido pediu senhor Rouhani conjunto Irã em um curso diferente para o futuro* : preocupações internacionais sobre o programa nuclear iraniano[...]" (C13); "Professor Gareth Evans, do centro de mama de Manchester, na Grã-Bretanha, disse que os dois genes BRCA 1 e BRCA 2 (câncer de mama 1 e 2) foram que os dois primeiros majorly mama câncer pre-eliminação de genes que foram identificados e estão também ligados a um risco aumentado de câncer de ovário[...]" (C15)).

e) **Erros fonte-alvo**: o erro engloba uma sequência de palavras que são corretas na língua alvo, não sendo possível identificar qual palavra foi traduzida incorretamente. A saída da TA é correta (legível, gramatical e coerente), no entanto, não condiz com a sentença-fonte. (p. ex: uma tradução como: "O garoto estava jogando bola", quando na realidade, a tradução "correta" (referência) deveria ser "Os garotos estavam jogando bola".[3]

Levando em conta a existência desses e de outros erros de TA, os métodos/sistemas de SAMM podem gerar sumários com baixa legibilidade, problemas agramaticais e incoerências (EVANS et al., 2005; WAN; LI; XIAO, 2010).

[3] A sentença exemplificada não no item em questão ("erros fonte-alvo") tem apenas caráter ilustrativo, uma vez que não foram encontradas ocorrências reais desse erro no *corpus*.

Dessa forma, a SA em contexto multilíngue precisa encontrar maneiras de lidar com tais problemas. Na literatura, identificam-se ao menos duas estratégias que buscam driblar os problemas de legibilidade dos sumários causados por erros de TA.

Em uma delas, a seleção das sentenças dos textos-fonte é feita pela combinação de dois fatores: (i) informatividade e (ii) qualidade de tradução. (p. ex. WAN; LI; XIAO, 2010 e BOURDIN et al., 2011)

Em outra estratégia, em que se parte de uma coleção composta por textos traduzidos em uma Lx e textos originais nessa mesma língua, somente os textos traduzidos são submetidos à seleção de conteúdo. Especificamente, as sentenças são selecionadas em função de algum método de relevância e, na sequência, as sentenças traduzidas selecionadas são comparadas às dos textos originais. Caso haja sentenças originais similares às traduzidas que foram selecionadas, as originais passam a compor o sumário. Evans et al. (2005) é um exemplo de trabalho em que se utiliza a similaridade para a substituição das sentenças traduzidas por originais.

Para o português, tem-se conhecimento de uma única investigação sobre a SAMM. No caso, trata-se do trabalho de Tosta et al. (2013), em que se testaram 2 métodos (Método 1 e Método 2) *baseline* capazes de sumarizar uma coleção composta por 3 textos jornalísticos, cada um deles em uma língua distinta (inglês, espanhol e português), e gerar um sumário em português. Para tanto, construiu-se o *corpus* denominado CM3News (*Corpus Multidocumento Trilíngue de Textos Jornalísticos*), composto por 10 coleções de textos jornalísticos, advindos de fontes distintas e nas línguas mencionadas (TOSTA et al., 2012).

Os métodos investigados por Tosta et al. (2013) caracterizam-se por englobar a etapa de TA dos textos-fonte em espanhol e inglês para o português antes do processo de seleção de conteúdo. Dessa forma, tais métodos seguem a abordagem *early-translation* e se baseiam em conhecimento linguístico superficial para a seleção do conteúdo a compor o sumário multidocumento. Uma vez traduzidos para o português, os autores aplicam aos textos-fonte métodos clássicos de SA (GUPTA, LEHAL, 2010), comumente utilizados no cenário multidocumento, a saber: (i) localização e (ii) frequência.

No Método 1, a seleção de conteúdo pauta-se na localização da informação nos textos-fonte (BAXENDALE, 1958). Esse método apoia-se na hipótese de que as informações localizadas no início dos textos jornalísticos veiculam a informação mais relevante.

No Método 2, a seleção é feita com base na frequência de ocorrência das unidades lexicais na coleção (LUHN, 1958). Assim, as sentenças dos textos-fonte recebem uma pontuação resultante da soma da frequência de ocorrência na coleção de suas palavras de classe aberta, a partir da qual são ranqueadas em ordem decrescente. Assim, o topo do ranque é ocupado pelas sentenças compostas pelas palavras mais frequentes da coleção.

Tendo em vista que as coleções de Tosta et al. (2013) eram multidocumento, houve a necessidade de eliminar a (i) redundância, um dos fenômenos típicos da multiplicidade de textos-fonte que abordam o mesmo assunto, e as (ii) sentenças com problemas resultantes da TA.

Para ambos, adotou-se a medida *word-overlap (Wol)*, que calcula a redundância ou similaridade entre sentenças com base na sobreposição das palavras de classe aberta idênticas (JURAFSKY, MARTIN, 2001). Com base na similaridade calculada por essa medida, selecionou-se do ranque de cada um dos métodos investigados, idealmente, apenas sentenças não-redundantes e sem problemas de tradução. Especialmente, para a criação de cada sumário, (i) selecionou-se a primeira sentença do ranque, (ii) comparou-se a sentença em questão com a próxima do ranque e, caso não fosse redundante com a sentença previamente selecionada, era selecionada para o sumário, (iii) compararam-se, da mesma forma, cada uma das sentenças selecionadas com a próxima do ranque, até que a taxa de compressão fosse atingida. Nos casos em que sentenças traduzidas eram selecionadas, estas foram substituídas por sentenças redundantes originais em português, quando existentes.

Para verificar o desempenho dos 2 métodos, os autores submeteram os sumários gerados manualmente a uma avaliação intrínseca manual, que consistiu na análise da legibilidade (ou fluência) dos sumários de 5 das 10 coleções do CM3News.[4] Os sumários foram analisados por 1 especialista em função dos 5 parâmetros utilizados na DUC 2007 (*Document Understanding Conference*)[5]: (i) gramaticalidade, (ii) não-redundância, (iii) clareza referencial, (iv) foco (temático), e (v) estrutura/coerência.

Nessa avaliação, constatou-se que o Método 1, pautado na localização associado ao tratamento da redundância e da tradução, obteve em média as mais altas pontuações quanto aos 5 parâmetros.

---

[4] Salienta-se que o corpus CM3News não se trata de um corpus paralelo, dessa forma, os textos de uma mesma coleção não são traduções uns dos outros.

[5] Em 2008, a DUC tornou-se parte de outra conferência, denominada *Text Analysis Conference* (TAC). O site <http://duc.nist.gov/> engloba as informações referentes à DUC de 2001 a 2007.

Idealmente, o cálculo de similaridade entre as sentenças objetivou substituir sentenças traduzidas automaticamente com problemas de TA por sentenças similares segundo o cálculo de *Wol*, e que fossem provenientes dos textos originais em português. Constatou-se que, apesar da realização desse procedimento, os sumários ainda apresentavam problemas de gramaticalidade, posto que, dentre os 5 parâmetros da DUC, esse foi o que obteve as médias mais baixas. Uma possível explicação reside no fato de que alguns sumários apresentam algumas sentenças traduzidas que não eram redundantes, mas que possuíam problemas de tradução.

Tendo em vista que o único trabalho sobre SAMM que envolve o português investigou métodos tidos como *baseline,* já que englobavam a TA dos textos-fonte e critérios de seleção superficiais (localização e frequência), traçam-se os objetivos descritos na próxima seção, que visam contribuir para o avanço das pesquisas sobre SAMM.

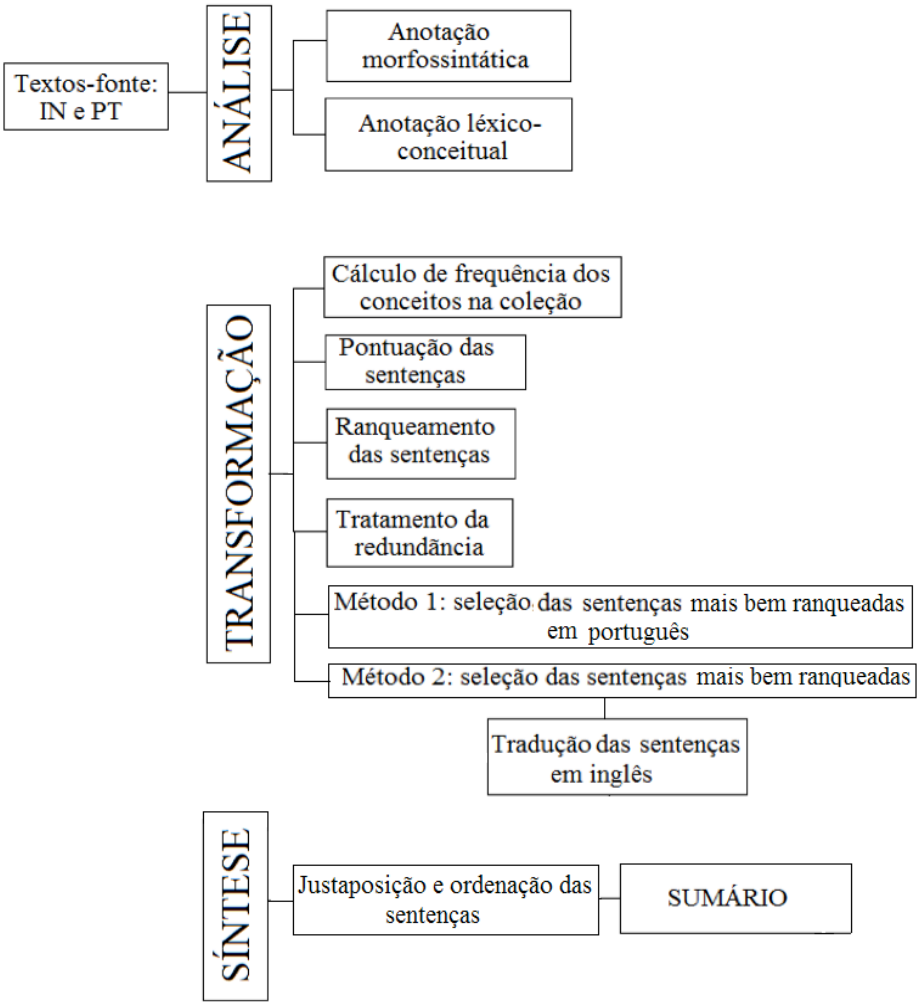
1.2 Objetivos e hipóteses

Como objetivo geral deste trabalho, investigaram-se métodos extrativos de SAMM baseados em conhecimento profundo (em especial, conhecimento léxico-conceitual) e que, por isso, não englobem a problemática fase de TA dos textos-fonte na íntegra. Como objetivo específico, propuseram-se 2 métodos extrativos de SAMM capazes de sumarizar uma coleção composta por textos em português e em inglês com base em conhecimento linguístico de nível léxico-conceitual.

Especificamente, os 2 métodos investigados, doravante Método 1 e Método 2, são ambos pautados na frequência de ocorrência dos conceitos nominais da coleção de textos-fonte. O Método 1 caracteriza-se por focar a língua do usuário, no caso, o português. Assim, as sentenças são ranqueadas em função da frequência dos conceitos (em toda a coleção bilíngue) e somente as sentenças em português mais bem pontuadas são selecionadas para o sumário até que se atinja a taxa de compressão, ou seja, até que o tamanho desejado do sumário seja atingido. O Método 2 caracteriza-se por selecionar as sentenças apenas em função do ranque (na coleção), sem levar em consideração a língua do usuário. Consequentemente, o sumário pode conter sentenças em língua estrangeira, as quais são automaticamente traduzidas para o português.

A Figura 1 ilustra o esquema geral dos métodos investigados baseado no modelo trifásico de SA (MANI, MAYBURY, 1999):

Figura 1– Esquema geral dos métodos 1 e 2.



Fonte: Elaborarada pelo autor.

Tendo em vista que, os métodos baseados no paradigma profundo podem apresentar melhores resultados, visou-se nesta investigação a aplicação de conhecimento profundo, em especial, léxico-conceitual na SAMM envolvendo o português, de tal forma que a etapa de TA dos textos-fonte não fosse realizada.

Esses objetivos (geral e específico) foram traçados com base em 7 hipóteses, sendo a hipótese 1 e 2 consideradas "fracas" e as de 3 a 6 "fortes":

1. É possível realizar a SAMM sem traduzir integralmente os textos-fonte ou os sumários, com base em conhecimento léxico-conceitual;

2. O conhecimento léxico-conceitual reflete os fenômenos multidocumento (redundância, complementariedade e contradição);
3. A ocorrência de conceitos em múltiplos textos reflete a informação principal;
4. As divergências lexicais entre as línguas não terão impacto na SAMM, devido à representação conceitual;
5. Um sumário composto exclusivamente por sentenças originais em português reflete as informações mais relevantes da coleção, posto que os conceitos que ocorrem no texto em inglês são levados em consideração no processo de ranqueamento das sentenças;
6. Um sumário composto por sentenças originais em português e por sentenças traduzidas individualmente para o português apresenta menos problemas de TA do que um sumário produzido a partir de uma coleção composta por textos traduzidos de forma integral para o português e originais nessa mesma língua [6];

Os objetivos formulados, aliás, integram os do projeto maior denominado SUSTENTO (FAPESP 2012/13246-5/ CNPq 483231/2012-6), que objetiva gerar conhecimento linguístico que possa subsidiar o enriquecimento de métodos e/ou a proposição de novos métodos, principalmente no que tange ao processamento do português[7]. As contribuições deste projeto para o projeto SUSTENTO são elencadas mais detalhadamente na seção 6.2.

Os resultados do projeto SUSTENTO, que incluem os desta pesquisa, poderão ser utilizados em outro projeto, denominado SUCINTO[8] (FAPESP 2012/03071-3), cujo objetivo é produzir recursos, ferramentas e sistemas de SA que, além da contribuição científica, possam ser disponibilizados para uso de pesquisadores e usuários finais. Ambos estão sendo desenvolvidos no Núcleo Interinstitucional de Linguística Computacional (NILC)[9].

[6] A hipótese 6 leva em consideração o fato de que, quanto maior a complexidade textual a ser traduzida de forma automática, maior a probabilidade de erros de tradução. Nesse sentido, um texto extenso traduzido integralmente possui maior probabilidade de possuir erros de tradução do que uma sentença traduzida individualmente.

[7] Disponível em: <http://www.nilc.icmc.usp.br/arianidf/sustento>

[8] A página eletrônica do projeto SUCINTO (*Summarization for Clever Information Acess*) está disponível em: <http://www.icmc.usp.br/~taspardo/sucinto/>.

[9] A página eletrônica do NILC (Núcleo Interinstitucional de Linguística Computacional) está disponível em:< http://www.nilc.icmc.usp.br/nilc/>.

1.3 Metodologia

Para alcançar os objetivos e conferir as hipóteses, o trabalho foi equacionado nas seguintes etapas metodológicas:

- Revisão da literatura: consistiu na leitura constante da bibliografia fundamental e de demais referências encontradas durante a pesquisa que fossem pertinentes. A bibliografia foi composta basicamente por trabalhos sobre SA em contexto monolíngue (mono e multidocumento) e multilíngue.

- Construção do *corpus*: essa etapa consistiu na construção de um *corpus* que satisfizesse às necessidades da investigação. No caso, construiu-se o *corpus* CM2News que possui as seguintes características: (i) multidocumento, (ii) multilíngue (português-inglês) e (iii) jornalístico. Especificamente, os textos são de domínios e assuntos variados, totalizando 24.724 palavras.

- Seleção das unidades lexicais: essa etapa consistiu na delimitação das unidades lexicais a serem anotadas em nível conceitual. Para tanto, alguns critérios foram investigados, como (i) a categoria gramatical e/ou (ii) a frequência. Ao final, optou-se pelo critério da classe de palavra. Assim, apenas os nomes comuns foram anotados em nível conceitual, posto que estes carregam boa parte da carga semântica de um texto.

- Seleção da ontologia: essa etapa consistiu no estudo e na seleção da ontologia com base na qual as unidades lexicais do *corpus* foram anotadas. Tendo em vista a não existência de uma ontologia digital de língua geral em português que satisfizesse a necessidade de anotação de conceitos de domínio variados, a escolha pela WordNet de Princeton (WN.Pr) (FELLBAUM, 1998) pautou-se em justificativas teóricas e práticas.

- Anotação semântica: essa etapa consistiu na anotação das unidades lexicais da classe dos nomes com os conceitos da WN.Pr, os quais estão codificados em *synonym sets* ou *synsets* (isto é, conjuntos de formas sinônimas). Esse processo foi

semiautomático, posto que foi feito por meio da utilização de um editor que recebeu a denominação MulSen.[10] Esse editor engloba os processos automáticos de (i) anotação morfossintática dos textos-fonte em inglês e português para a identificação dos nomes, (ii) sugestão de equivalentes de tradução em inglês para as unidades lexicais a serem anotadas provenientes dos textos em português, posto que os conceitos na WN.Pr são representados por unidades da língua inglesa, (iii) sugestão de conceitos, dentre os armazenados na WN.Pr, para a anotação final das unidades lexicais em português e inglês.

- Proposta e Aplicação de métodos de SAMM: essa etapa consistiu na proposição e aplicação dos 2 método(s) extrativo(s) de SAMM baseado(s) na anotação dos textos-fonte em nível conceitual. Para tanto, os textos-fonte foram segmentados em nível sentencial e as sentenças pontuadas e ranqueadas em função da ocorrência dos conceitos na coleção. No método 1, foram selecionadas do ranque, apenas as sentenças mais bem pontuadas que foram extraídas de textos em português. No Método 2, as sentenças foram selecionadas apenas em função da sua classificação no ranque. Especificamente, nos casos em que sentenças dos textos-fonte em inglês foram selecionadas para compor o sumário, estas foram traduzidas de forma automática.

- Avaliação: essa etapa consistiu na avaliação intrínseca dos métodos investigados, ou seja, na avaliação dos sumários por eles gerados. Os métodos foram avaliados pela qualidade linguística e informatividade dos seus sumários. A avaliação da qualidade linguística ocorreu de forma manual, a partir dos parâmetros da DUC 2007 (*Document Understanding Conference*)[11] e a avaliação da informatividade foi avaliada automaticamente pelo pacote de medidas da ROUGE (*Recall-Oriented Understudy of Gisting Evaluation*).

[10] Essa ferramenta foi desenvolvida por Fernando Antônio Asevedo Nóbrega, doutorando do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional do ICMC/USP/São Carlos e pesquisador do NILC, com a supervisão do Prof. Dr. Thiago A. S. Pardo (ICMC-USP), que coorienta este trabalho. Ressalta-se que essa ferramenta foi fundamental para o desenvolvimento desta pesquisa e, por isso, agradece-se a todos os envolvidos diretamente na sua construção.

[11] Em 2008, a DUC tornou-se parte de outra conferência, denominada *Text Analysis Conference* (TAC). O site http://duc.nist.gov/ engloba as informações referentes à DUC de 2001 a 2007.

1.4 Estrutura da Dissertação

Em termos formais, esta dissertação organiza-se em 7 Seções. Na Seção 2, apresenta-se a revisão da literatura. Na Seção 3, apresentam-se o processo de construção e anotação do *corpus* CM2News, bem como a configuração do editor de anotação MulSen e sua interface gráfica. Na Seção 4, descrevem-se os métodos de SAMM propostos neste trabalho. Na Seção 5, apresentam-se os procedimentos de avaliação dos métodos e seus resultados. Por fim, na seção 6, algumas considerações finais são feitas, apresentando as contribuições dadas por este trabalho, suas limitações e trabalhos futuros.

2 REVISÃO DA LITERATURA

Na Seção 2.1, apresentam-se os conceitos gerais sobre a SA. Especificamente, a revisão sobre os conceitos gerais da SA engloba a descrição das etapas que constitui esse processo, dos fatores que o afetam e dos diferentes tipos de avaliação. Nas Seções 2.2 e 2.3, apresenta-se, respectivamente, uma revisão sobre a SA monolíngue e a SA que envolve mais de uma língua. Ao se apresentar os trabalhos monolíngue, dar-se-á maior atenção aos multidocumento, posto que o foco deste trabalho é a SAMM.

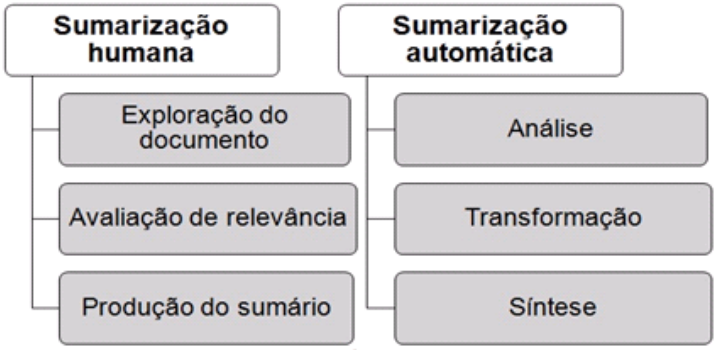
2.1 Noções básicas de Sumarização Automática

O Processamento Automático das Línguas Natural (PLN) é uma área multidisciplinar que busca desenvolver sistemas capazes de realizar tarefas linguísticas específicas, como a correção ortográfica e gramatical, a tradução, a extração de informação, entre outras (DIAS-DA-SILVA et al., 2007).

Na subárea do PLN denominada Sumarização Automática (SA), automatiza-se a produção de sumários (ou resumos) principalmente a partir de textos (MANI, 2001). A SA é motivada pela enorme quantidade de informação disponível e pelo pouco tempo que as pessoas têm para assimilar tanta informação. Os sistemas que realizam essa tarefa de PLN são denominados sumarizadores automáticos (SPARCK JONES, 2007).

Buscando emular na máquina as etapas de sumarização humana, Cremmins (1996) e Endres-Nieggemeyer (1998) sugerem que a SA envolva idealmente 3 processos: (i) análise dos textos-fonte, (ii) transformação e (iii) síntese. O paralelo entre as etapas humanas e os processos automáticos de sumarização é apresentado na Figura 2.

Figura 2 – Etapas de sumarização humana e automática.



Fonte: Sparck Jones (1993) com adaptação de Endres-Niggemeyer (1998).

A análise visa interpretar os textos-fonte e extrair uma representação formal dos mesmos.

A transformação é a etapa principal, pois, a partir da representação gerada na análise, o conteúdo dos textos-fonte é condensado em uma representação interna do sumário. Essa etapa engloba a seleção do conteúdo que irá compor o sumário e, para tanto, é necessário ranquear os segmentos dos textos-fonte em função de sua relevância e selecionar os de maior pontuação (isto é, que contêm as ideias centrais do texto) até que o tamanho desejado do sumário seja atingido. O ranqueamento pode seguir diferentes critérios.

A síntese visa à construção do sumário em língua natural a partir da representação interna gerada na transformação. Para tanto, métodos de justaposição, ordenação, fusão e correferenciação dos segmentos textuais selecionados podem ser utilizados (SPARCK JONES, 1993). No caso da produção de extratos compostos por sentenças, as mesmas são comumente justapostas na ordem em que são selecionadas do ranque para formar o sumário.

A SA, comumente realizada de acordo com os 3 processos sugeridos por Sparck Jones (1993), é influenciada por uma série de fatores. Dentre eles, estão: (i) taxa de compressão, (ii) audiência, (iii) função, (iv) forma, (v) gênero, (vi) número de textos-fonte, (vii) quantidade de línguas e outros (MCKEOWN, RADEV, 1995; MANI, 2001).

A taxa de compressão é o fator que determina o tamanho desejado dos sumários; se a taxa de compressão é estipulada em 70%, um sumário produzido a partir de um único texto deverá apresentar tamanho equivalente a 30% desse texto-fonte (em geral, medido em número de palavras). Assim, essa taxa determina o volume de informação a ser selecionado durante a transformação para a composição do sumário.

Com relação à audiência a que se destinam, os sumários podem ser genéricos ou focados nos interesses dos usuários. Os sumários genéricos veiculam as informações mais importantes dos textos-fonte sem se preocupar com um perfil específico de usuário. Os sumários focados nos interesses dos usuários, por sua vez, customizam as informações que veiculam em função do conhecimento prévio dos usuários ou de uma consulta (*query*) realizada por eles. Por exemplo, se o usuário for leigo no assunto de um texto-fonte, pode ser interessante que o sumário desse texto contenha informações contextuais; se, por outro lado, o usuário for um especialista no assunto, as informações contextuais já não são mais relevantes, sendo pertinente, por exemplo, a veiculação de informação nova.

Quanto à função, pode-se objetivar a produção automática de sumários informativos, indicativos ou críticos. Os sumários indicativos (p.ex.: índices de livros e outros) não substituem o texto-fonte, apenas dizem do que ele trata, podendo funcionar como ponto de partida para a seleção de uma leitura mais profunda sobre determinado tópico que o usuário julgar interessante. Os informativos contêm as informações principais de um texto-fonte de forma coerente e coesa a ponto de substituir a leitura do mesmo (p.ex.: *abstracts* de artigos científicos). Os sumários críticos apresentam as informações principais dos textos-fonte acrescidas de avaliações sobre elas; as resenhas de obras como livros e filmes são exemplos de sumários críticos.

Quanto à forma, a SA pode produzir extratos ou *abstracts*. Os extratos são sumários compostos por trechos (comumente, sentenças) extraídos na íntegra dos textos-fonte, não havendo, portanto, nenhum tipo de modificação ou reescrita do conteúdo do *input*. Em contraposição, os *abstracts* são sumários produzidos por meio da reescrita do conteúdo veiculado pelos textos-fonte, havendo, portanto, material linguístico que não estava presente no *input*. Para a produção de *abstracts,* é preciso aplicar processos bastante complexos de sumarização, como a fusão e generalização de informação.

O gênero dos sumários é outro fator que afeta diretamente a SA, pois, dependendo do gênero, diferentes estratégias de seleção são aplicadas na transformação.

Sobre o número de textos-fonte, a SA pode ser mono ou multidocumento. Na primeira, produz-se um sumário a partir de um único texto-fonte e, na SA multidocumento (SAM), o sumário é produzido a partir de uma coleção de textos, de fontes distintas, que abordam o mesmo assunto.

Com relação ao número de línguas, a SA pode envolver apenas uma língua ou mais de uma língua. No primeiro caso, a modalidade de SA é denominada monolígue e se caracteriza pela geração de um sumário em uma língua *x* a partir de um ou mais textos na mesma língua *x.* No segundo caso, a SA pode ser de 3 tipos: (i) *cross-language,* em que, a partir de um ou mais textos em uma língua *x,* produz-se um sumário em uma língua *y,* (ii) multilíngue, em que, a partir de uma coleção de textos em diferentes línguas, produz-se um sumário em uma das línguas dos textos de entrada, e (iii) independente de língua (ORĂSAN, 2009).

Além dos fatores mencionados, ressalta-se que a SA também é influenciada pela quantidade e pelo nível de conhecimento linguístico envolvidos no processo, os quais determinam a abordagem segundo a qual a SA é realizada.

Segundo Mani (2001), a SA pode ser realizada com base em pouco ou nenhum conhecimento linguístico. Nesse caso, tem-se a abordagem superficial, uma vez que o tratamento dos textos-fonte pauta-se comumente em dados estatísticos. Por essa razão, esses sistemas geram extratos e apresentam as seguintes características positivas: (i) baixo custo de desenvolvimento e (ii) altas robustez e escalabilidade. Por outro lado, os métodos superficiais podem produzir sumários menos coerentes, coesos e informativos[12].

Quando o SA envolve o uso massivo de conhecimento linguístico codificado em gramáticas, repositórios semânticos e modelos de discurso, diz-se que esta segue a abordagem profunda. O desenvolvimento de métodos de SA segundo a abordagem profunda é mais caro em relação aos superficiais e sua aplicação é mais restrita. O desempenho, no entanto, pode ser superior, uma vez que os sumários produzidos podem ser mais coerentes, coesos e informativos. Os sumarizadores profundos podem gerar não só extratos, mas também *abstracts* (isto é, sumários produzidos pela reescrita dos textos-fonte).

Ressalta-se que as abordagens superficiais e profundas podem ser mescladas, originando abordagens híbridas.

Para avaliar os métodos/sistemas de SA, a comunidade do PLN tem realizado conferências internacionais dedicadas, como a SUMMAC[13] (*Text Summarization Evaluation Conference*) e a TAC[14] (*Text Analysis Conference*), o que evidencia a importância e a necessidade da avaliação e das dificuldades inerentes à SA.

A avaliação de métodos/sistemas de SA pode ser intrínseca ou extrínseca. Na intrínseca, avalia-se o desempenho dos métodos/sistemas pela análise de seus resultados (sumários). Na extrínseca, avalia-se a utilidade dos sumários em tarefas específicas, por exemplo, a recuperação de informação (SPARCK JONES; GALLIERS, 1996).

A avaliação da qualidade dos sumários automáticos é tradicionalmente realizada por humanos, pois o foco reside na análise de aspectos relativos à gramaticalidade (p.ex.: ortografia e gramática) e à textualidade (p.ex.: coesão e coerência) (p.ex.: SAGGION; LAPALME, 2000; WHITE et al*.,* 2000), os quais dificilmente podem ser avaliados automaticamente.

---

[12] Entende-se por robustez a capacidade de um sistema em lidar com entradas que não respeitam as regras definidas. Assim, a robustez de um sistema garante seu funcionamento adequado mesmo na presença de entradas que se afastam do padrão esperado (MENZEL, 1995). Por escalabilidade, entende-se a habilidade de um sistema em conseguir funcionar em vários domínios sem sofrer perdas.

[13] Disponível em: < http://www-nlpir.nist.gov/related_projects/tipster_summac/>.

[14] Disponível em: < http://www.nist.gov/tac/about/index.html>.

Na SAM, a DUC (2007), por exemplo, propõe uma série de atributos linguísticos que buscam avaliar a qualidade dos sumários em função de gramaticalidade, não-redundância, clareza referencial, foco, estrutura e coerência. Sobre a SAM que envolve o português, Castro Jorge e Pardo (2011) realizaram a avaliação humana com base em critérios bastante semelhantes aos da DUC a saber: informatividade, coerência, coesão, redundância e gramaticalidade. A avaliação da informatividade consiste em identificar o quanto de informação relevante dos textos-fonte o sumário automático incorpora. Essa identificação é feita pela comparação automática entre os sumários automáticos e os sumários humanos (denominados "sumários de referência"). Para isso, utiliza-se o pacote de medidas da ROUGE, que calcula a informatividade pela coocorrência de n-gramas entre os sumários automáticos e os humanos e expressa essa informatividade pelas medidas de precisão, cobertura e medida-f (LIN; HOVY, 2003). Uma vez que não há consenso sobre a melhor forma de se avaliar um sistema dessa natureza, diversos autores investigaram outras estratégias (p.ex.: SPARCK JONES, 1999; SAGGION et al., 2002; LOUIS, NENKOVA, 2013).

Saggion et al. (2002), por exemplo, propuseram 3 métodos de avaliação baseado em conteúdo que medem a similaridade entre os sumários: (i) similaridade do cosseno,[15] (ii) sobreposição de unidades lexicais (unigrama ou bigrama) e (iii) sobreposição da maior subsequência de unidades lexicais.

Louis e Nenkova (2013) apresentam 3 métricas de avaliação para a sumarização, conhecidas como modelo da pirâmide,[16] (i) similaridade entre textos-fonte e sumários, isto é, métricas que consideram que quanto mais similar o sumário é dos seus textos-fonte, melhor o seu conteúdo, (ii) adição de pseudomodelos, ou seja, aos sumários humanos de referência, acrescentam-se sumários automáticos escolhidos por humanos, e (iii) sumários automáticos como modelo, isto é, as autoras consideram que os sumários automáticos são bons o suficiente para servirem de sumários de referência e, portanto, não utilizam nenhum esforço humano para a criação dos mesmos.

---

[15] A medida de similaridade do cosseno faz com que os resultados das comparações entre dois documentos sejam normalizados. Os pares de documentos são comparados e os documentos são considerados similares a partir de um limiar previamente estabelecido.

[16] O modelo de avaliação da "pirâmide" atribui valor ao sumário por meio da similaridade entre suas *summarization content units* (SCUs), ou seja, a SCU que aparecer em todos os sumários de referência sob avaliação recebe o maior peso, baseado na quantidade de sumários em que ocorreu, e ocupa a última camada da pirâmide. Nesse sentido, é possível prever o conteúdo ideal que deve conter em um sumário, visto que no topo se encontram as unidades mais importantes.

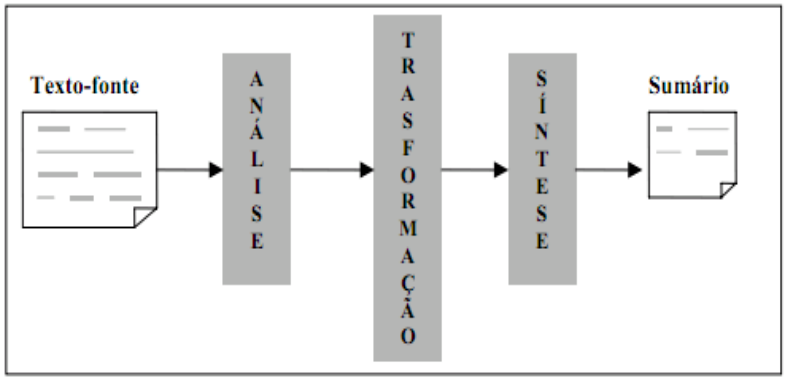
2.2 A Sumarização Automática Monolíngue

Focalizando-se apenas uma língua, a SA pode ser monodocumento ou multidocumento.

2.2.1 A Sumarização Automática Monodocumento

A SA monodocumento monolíngue é uma modalidade tradicional e, por isso, tem sido foco de pesquisa desde a década de 1950 (p.ex.: LUHN, 1958; EDMUNDSON, 1969; O'DONNELL, 1997; SALTON et al., 1997; MARCU, 2000; CONROY; O'LEARY, 2001; PARDO; RINO, 2002; PARDO et al., 2003; RINO et al., 2004; SVORE et al.., 2007; UZÊDA et al., 2010; CLARKE; LAPATA, 2010; LOUIS et al., 2010, etc.). Na Figura 3, ilustra-se, com base em Sparck Jones (1993), a arquitetura típica de um sumarizador monodocumento.

Figura 3 – Arquitetura genérica de um sumarizador monodocumento.



Fonte: Sparck Jones (1993).

Na SA monodocumento, o foco é, tradicionalmente, a produção de extratos informativos e genéricos. Para a seleção das sentenças que compõem o sumário, há várias estratégias de seleção de conteúdo (sentenças) empregadas de forma individual ou em conjunto (GUPTA; LEHAL, 2010). Tais estratégias podem e são também utilizadas em outras modalidades de SA, como a SAM.

Uma das estratégias é a seleção de informação que se relaciona com as palavras contidas no título/subtítulo dos textos-fonte. Essa estratégia pressupõe a identificação das palavras que compõem o título, subtítulo e tópicos para selecionar sentenças que

contenham as ideias principais dos textos. No caso, a existência de subtítulos e tópicos depende do tipo/gênero e do tamanho dos textos a ser sumarizado.

Outra estratégia consiste na seleção de conteúdo com base nas palavras-chave dos textos-fonte. As palavras-chave são comumente as de classe aberta mais frequentes dos textos-fonte. A utilização desse atributo pressupõe que as palavras mais frequentes expressam o conteúdo principal de um texto.

Além da seleção de conteúdo com base nas palavras do título/subtítulo e palavras-chave, destaca-se que o tamanho ou extensão (em número de palavras) das sentenças dos textos-fonte também são um critério comumente utilizado, com base no qual as sentenças de tamanho médio são selecionadas para compor um sumário.

Outra estratégia é a seleção de sentenças que contêm expressões-chave ou indicativas de conteúdos que caracterizam os componentes da estrutura discursiva dos gêneros. Um texto científico, por exemplo, apresenta uma estrutura composta pelos componentes "resumo", "introdução", "materiais/métodos", "resultados", "discussão" e "conclusão", os quais são introduzidos nos textos por certas expressões, que, para a seleção de conteúdo, funcionam como pistas; a expressão "o objetivo deste trabalho é", por exemplo, indica a expressão do conteúdo "meta/objetivo".

Além das estratégias mencionadas, a seleção do conteúdo pode ser feita com base na localização das sentenças no texto. Para a geração de sumários jornalísticos, seleciona-se a sentença localizada no início do texto-fonte, pois esta expressa o *lead,* ou seja, a informação principal veiculada em um texto jornalístico (LUHN, 1958; EDMUNDSON, 1969).

As estratégias mencionadas e outras têm subsidiado o desenvolvimento de métodos superficiais e profundos de SA monodocumento.

Nos métodos superficiais, as estratégias de seleção de conteúdo são traduzidas em atributos linguísticos simples, os quais guiam a seleção das sentenças de um texto-fonte para a geração de seu respectivo sumário extrativo (genérico e informativo). Em um trabalho clássico da SA monodocumento, Baxendale (1958) propôs um método superficial em que um sumário científico é produzido pela seleção das sentenças localizadas no início e final dos parágrafos do seu respectivo texto-fonte. Em outro trabalho clássico, Luhn (1958) propôs um método superficial em que as sentenças são pontuadas e ranqueadas com base nas palavras mais frequentes do texto.

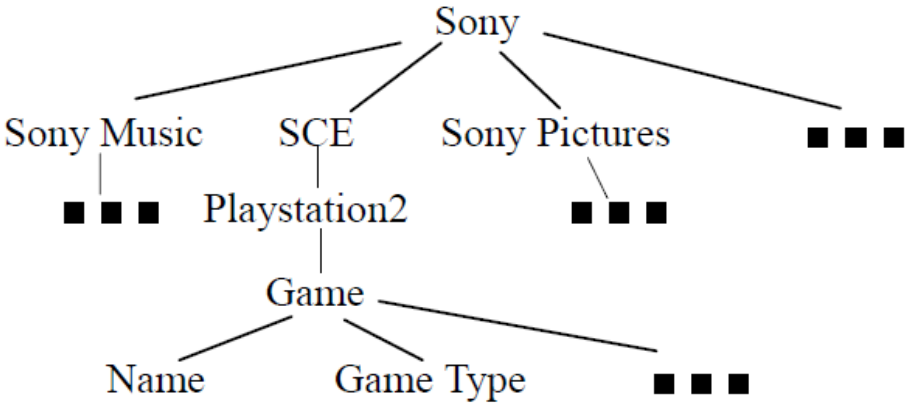
No cenário dos métodos profundos de SA monodocumento, destacam-se os métodos de Wu e Liu (2003) e Henniget al. (2008), que se baseiam em conhecimento

léxico-conceitual. Para ilustração, descreve-se com mais detalhes o método de Wu e Liu (2003).

Especificamente, o método de SA monodocumento de Wu e Liu (2003) baseia-se na identificação dos principais tópicos e subtópicos de um texto-fonte para, a partir deles, selecionar os parágrafos que contêm tais informações topicais para compor o sumário. A identificação topical é feita pela comparação dos termos que ocorrem nos parágrafos aos termos de uma ontologia[17].

Para a proposição do método, os autores construíram um *corpus* e uma ontologia de domínio. O *corpus* é composto por 51 artigos publicados no *New York Times* ou no *The Wall Street Journal,* os quais foram compilados por meio da *query* (isto é, termo de busca) SONY. A ontologia, construída de forma manual, possui 142 termos organizados hierarquicamente, na forma de uma árvore. No caso, diz-se que se trata de uma ontologia de domínio que armazena, por exemplo, (i) termos/conceitos (p.ex.: Sony, Sony Music e Sony Pictures), e (ii) relações de subsunção (p.ex.: *Sony* subsume *Sony Music* e *Sony Pictures*). Por se tratar de uma árvore conceitual, diz-se que os conceitos são os nós ou folhas e as relações são os galhos. A Figura 4 ilustra os conceitos mais genéricos[18] da referida ontologia (WU; LIU, 2003).

Figura 4 – *Top-level ontology* do domínio *Sony Corporation*.



Fonte: Wu e Liu (2003).

[17] No PLN, ontologia pode ser definida como um recurso ou base de conhecimento que fornece um inventário de conceitos, propriedades e relações entre conceitos que representam "uma interpretação da realidade" (isto é, o conhecimento de mundo compartilhado pelos membros de uma comunidade linguística) (GRUBER, 1995).

[18] Os conceitos mais genéricos, dispostos nos níveis superiores de uma ontologia, constituem uma *top-ontology*.

Para que os parágrafos sejam pontuados em função da informação topical que expressam, é preciso comparar os termos que neles ocorrem aos termos da ontologia.

No trabalho de Wu e Liu (2003), os termos de um texto-fonte que não estão armazenadas na ontologia são descartados. Caso o termo esteja presente na ontologia, é feita a indexação do mesmo à ontologia e o elemento da ontologia é pontuado.

Quando se pontua um termo/conceito na hierarquia, seus termos/conceitos superiores são automaticamente pontuados. Por exemplo, na ontologia em questão, "Spider-man" é um nó-filho do nó-pai "movie"; assim, se um parágrafo contiver o termo "Spider-man", ambos os termo/conceitos, "Spider-man" e "movie", são pontuados na ontologia.

Com base na indexação e pontuação, o conceito mais genérico que inicia a *top-ontology* (p.ex.: Sony) terá sempre a pontuação mais elevada, enquanto os conceitos do segundo nível, que representam subtópicos, terão pontuações diferentes. Com isso, apenas os conceitos mais bem pontuados do segundo nível da hierarquia são selecionados para representar os subtópicos do texto. Na sequência, os conceitos com maior pontuação são, então, selecionados como os principais tópicos do documento de origem e cada parágrafo é pontuado em função desses tópicos. Os parágrafos são selecionados até que o tamanho desejado do sumário seja alcançado.

Dessa forma, pode-se dizer que o método de Wu e Li (2003) é uma versão mais sofisticada do método da palavra-chave, pois busca identificar o conteúdo de um texto-fonte por meio da frequência dos termos/conceitos organizados em uma ontologia.

Ainda quanto aos métodos profundos de SA monodocumento, destacam-se os que se baseiam especificamente em uma modelagem discursiva do texto-fonte. Neles, busca-se refletir a estratégia de seleção nessa modelagem. Por exemplo, ao se modelar um texto-fonte de acordo com a teoria *Rhetorical Structure Theory* (RST) (MANN; THOMPSON, 1987), gera-se uma árvore retórica em que as unidades de conteúdo (p.ex.: sentenças) são representadas por nós e as relações semântico-discursivas (p.ex.: *Circumstance, Background, Concession,* etc.) entre as unidades são representadas por arestas. Quando da SA de um texto jornalístico, a primeira sentença é geralmente a mais nuclear em uma árvore RST bem construída do mesmo texto e, por isso, selecionada para compor o sumário. Nesse caso, por uma árvore RST codificar conhecimento semântico-discursivo, a localização no topo dessa árvore é tida como um atributo profundo da sentença.

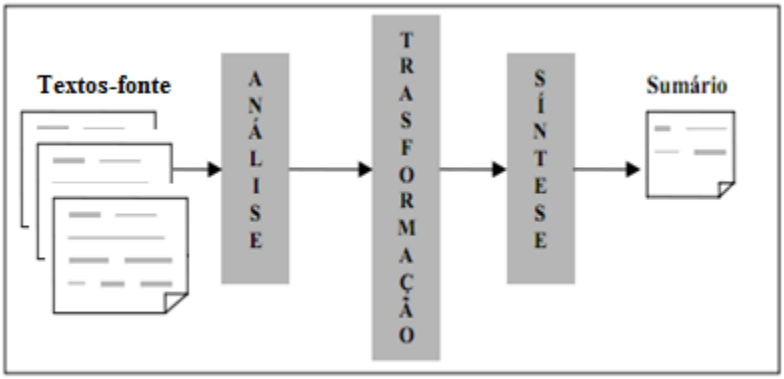
2.2.2 A Sumarização Automática Multidocumento

A SAM monolíngue estabeleceu-se como uma modalidade de SA de destaque em resposta à demanda por novas tecnologias de gerenciamento do enorme volume de informação que está em constante crescimento e mudança na *web* (MANI, 2001).

Especificamente, a SAM iniciou nos anos de 1990 (p.ex.: MCKEOWN; RADEV, 1995; RADEV; MCKEOWN, 1998; CARBONELL; GODLSTEIN, 1998) e tem adquirido relevância nos últimos anos (p.ex.: RADEV et al., 2000; ZHANG et al., 2002; OTTERBACHER et al., 2002; MCKEOWN et al., 2005; NENKOVA, 2005a, 2005b; WAN; YANG, 2006; AFANTENOS et al., 2004, 2008; WAN, 2008; HAGHIGHI; VANDERWENDE, 2009; CASTRO JORGE; PARDO, 2010, 2011; CELIKYILMAZ; HAKKANI-TUR, 2011, entre outros).

A SAM diferencia-se da SA (monodocumento) na medida que se parte não de um único texto, mas de uma coleção de textos (de fontes distintas) que abordam o mesmo assunto para a produção de um sumário. A arquitetura típica de um sumarizador multidocumento é a ilustrada na Figura 5, proposta com base em Sparck Jones (1993).

Figura 5 – Arquitetura genérica de um sumarizador multidocumento monolíngue.



Fonte: Sparck Jones (1993).

A SAM apresenta alguns fenômenos típicos que não ocorrem normalmente na SA(monodocumento) como a ocorrência de informações complementares, contraditórias e redundantes. A redundância, em especial, é o fenômeno mais comum na SAM, visto que a multiplicidade de textos-fonte, sobre um mesmo assunto, resulta em uma grande quantidade de informações redundantes ou repetidas.

Assim, os vários sistemas/métodos superficiais e profundos de SAM monolíngue, propostos principalmente para produzir extratos informativos e genéricos, pontuam e ranqueiam as sentenças dos textos-fonte em função de um critério de relevância e selecionam as mais relevantes para o sumário, desde que haja pouca similaridade ou redundância entre elas (MANI, 2001). Em outras palavras, os métodos de SAM englobam um **fator de redundância**, o qual busca garantir que as sentenças selecionadas não são redundantes entre si (JURAFSKY, MARTIN, 2001).

Como critérios de relevância para a pontuação/ranqueamento das sentenças (e, por conseguinte, para a seleção do conteúdo), utilizam-se as mesmas estratégias da SA monodocumento (p.ex.: localização) e/ou a redundância, pois a informação mais repetida entre os textos-fonte de uma coleção é tida como a mais importante para constar em um sumário multidocumento (MANI, 2001).

Os métodos superficiais podem ser organizados em 3 grupos de acordo com o tipo de conhecimento linguístico que utilizam para a seleção de conteúdo (GUPTA, LEHAL, 2010; KUMAR, SALIM, 2012).

O primeiro engloba os que se baseiam em atributos linguísticos (*feature-based methods*), que podem variar em número e combinação (p.x.: LIN, HOVY, 2002; SCHILDER, KONDADADI, 2008) e apresentar pesos diferentes em função do tipo/gênero dos textos-fonte (cf. BOSSARD, RODRIGUES, 2011; SUANMALI et al., 2011).

Um exemplo de atributo bastante aplicado é a frequência de ocorrência das palavras (de classe aberta). Em um método que se baseia nesse atributo, a análise é relativamente simples, consistindo na segmentação sentencial e no cálculo da frequência de ocorrência de cada palavra dos textos-fonte na coleção. A transformação consiste em pontuar e ranquear as sentenças em função da soma da frequência de suas palavras constitutivas e, na sequência, selecionar as mais pontuadas que não sejam redundantes[19] entre si para compor o sumário até que se atinja a taxa de compressão. O sumário então é sintetizado pela justaposição das sentenças na ordem em que aparecem nos textos-fonte.

O segundo grupo engloba os trabalhos baseados nos conceitos de *cluster* (grupo) e centroide (*cluster-based methods*) (p.ex.: RADEV et al*.,* 2004). Neles, a análise consiste em agrupar as sentenças de dada coleção em *clusters* (conjuntos) com base na

[19] Nesses casos, a redundância ou similaridade entre duas sentenças pode ser calculada por medidas como a *word ovelap*. A medida em questão é mais detalhadamente explicitada na página 51.

similaridade lexical. Assim, os *clusters* são formados por sentenças semelhantes entre si, que veiculam os "tópicos" da coleção. Cada *cluster* é representado por um centroide, ou seja, um conjunto de palavras estatisticamente importantes. Assim, seleciona-se, em cada *cluster,* a sentença que contém o maior número de palavras do centroide.

O terceiro grupo de métodos superficiais engloba aqueles cuja análise consiste em modelar os textos-fonte como grafos (*graph-based methods*) (p.ex.: SALTON et al., 1997; MIHALCEA, TARAU, 2005; WAN, 2008). Especificamente, as sentenças são modeladas como nós e a similaridade entre elas é modelada como arestas que conectam os nós. Assim, as sentenças mais fortemente conectadas a outras são extraídas para a construção do sumário.

Quanto à abordagem profunda, os métodos também podem ser organizados em 3 grupos de acordo com o tipo de conhecimento linguístico predominante (MANI, 2001).

No primeiro grupo, reúnem-se os métodos baseados em conhecimento sintático. Em Barzilay et al. (1999), por exemplo, a análise consiste em segmentar as sentenças e analisar sua estrutura sintática por meio de um *parser* (analisador sintático). A partir da análise sintática, as estruturas predicativas (predicado-argumento) similares são agrupadas, as quais teoricamente expressam tópicos, e as mais recorrentes são selecionadas. As estruturas predicativas são reordenadas na síntese, gerando as sentenças dos sumários abstrativos.

O segundo grupo inclui os métodos baseados em conhecimento semântico-conceitual, como os de Mani e Bloedorn (1997) e Li et al. (2010).

No trabalho de Mani e Bloedorn (1997), por exemplo, cada texto-fonte é modelado em um grafo na fase de análise, no qual as palavras são representadas por nós e a similaridade distribucional entre elas por arestas. No caso, dois nós com arestas similares representam palavras sinônimas e, portanto, expressam um conceito. Diante dos conceitos mais importantes da coleção, selecionam-se as sentenças dos textos-fonte que contêm as palavras que os expressam para compor o sumário.

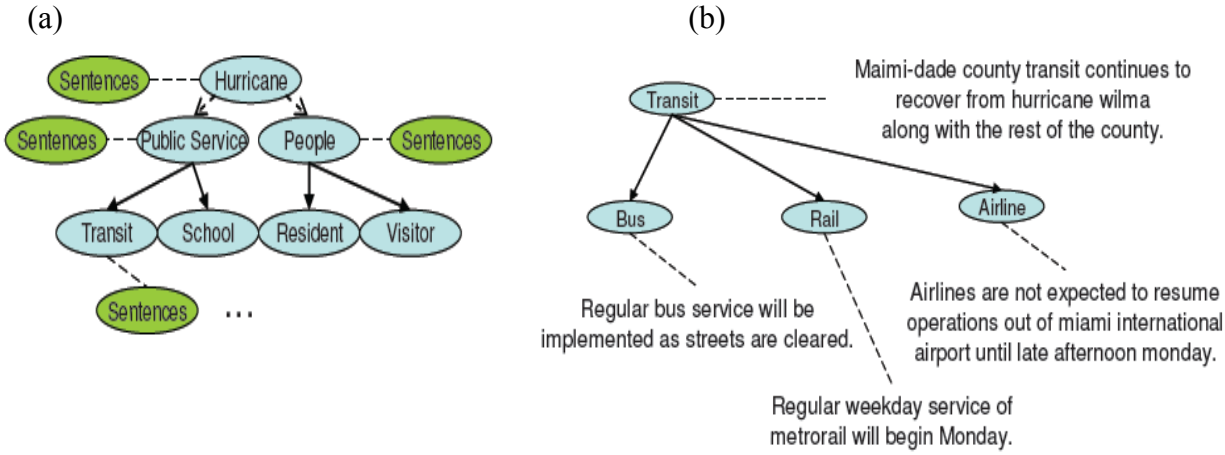
No trabalho de Li et al. (2010), especificamente, os autores mapeiam ou indexam as sentenças de uma coleção aos conceitos de uma ontologia. Dada a *query* de um usuário, a qual também é mapeada na ontologia, o sistema seleciona para compor o sumário apenas as sentenças dos textos-fontes indexadas aos mesmos conceitos a que as unidades lexicais da *query* foram mapeadas e/ou aos conceitos mais específicos. Para tanto, os autores utilizaram uma ontologia do domínio *desastre* construída manualmente e cujos conceitos são expressos por rótulos únicos.

Na Figura 6, vê-se, em (a), que as sentenças dos textos de uma coleção (elipses verdes) que versam sobre a passagem do furacão *Wilma* por Atlanta em 2005 foram indexadas a vários conceitos da ontologia (elipses azuis) (*Hurricane, Public Service, People* e *Transit*).

Na expansão do conceito *Transit* em (b), observa-se que há sentenças também indexadas a seus subordinados (*Bus, Rail* e *Airline*).

Assim, diante de uma *query* como "*get all the information related to transit in Miami-Dade County after Hurricane Wilma passed"* ("obter todas as informações relacionadas ao trânsito em *Miami-Dade County* após a passagem do furacão *Wilma*"), que também seria indexada ao conceito "Transit", apenas as sentenças indexadas a esse conceito e a seus subordinados seriam selecionadas para compor o sumário.

Figura 6 – Exemplo da indexação sentencial a uma ontologia em Li et al. (2010).



Fonte: Li et al. (2010).

No terceiro grupo, reúnem-se os métodos que utilizam conhecimento semântico-discursivo, que comumente modelam os textos-fonte na fase de análise segundo a teoria/modelo linguístico-computacional CST (do inglês, *Cross-document Structure Theory*) (RADEV, 2000).

Para tanto, utiliza-se um analisador discursivo, como o CSTParser (MAZIERO, 2012). Inspirada principalmente na RST (do inglês, *Rhetorical Structure Theory*) (MANN, THOMPSON, 1987), a CST permite estruturar o discurso por meio da conexão das sentenças (ou outras unidade textual, como palavras, sintagmas, etc.) provenientes de diferentes documentos (RADEV, 2000). Tal conexão é rotulada por

várias relações. Originalmente, propôs-se um conjunto de 24 relações discursivas para o relacionamento intertextual, as quais estão listadas no Quadro 1.

Quadro 1 – Conjunto original de relações da CST.

Identity	*Modality*	*Judgment*
Equivalence	*Attribution*	*Fulfillment*
Translation	*Summary*	*Description*
Subsumption	*Follow-up*	*Reader profile*
Contradiction	*Elaboration*	*Contrast*
Historical background	*Indirect speech*	*Parallel*
Cross-reference	*Refinement*	*Generalization*
Citation	*Agreement*	*Change of perspective*

Fonte: Adaptado de Radev (2000).

Trabalhos posteriores refinaram as relações, produzindo conjuntos mais compactos (p.ex.: ZHANG et al., 2003; MAZIERO et al., 2010).

A seguir, em (1), ilustra-se a relação *equivalence* (correspondente à paráfrase) entre os segmentos (a) e (b), provenientes de textos diferentes (RADEV, 2000, p. 6):

(1) a. *Ford's program will be launched in the United States in April and globally within 12 months.*

b. *Ford plans to introduce the program first for its employees in the United States, then expand it for workers abroad.*

A relação como a ilustrada em (1) permite indicar os fenômenos multidocumento, desde similaridades e diferenças de conteúdo, até questões de estilo de escrita. Assim, a seleção das sentenças nos métodos baseados na CST pode ser guiada pelo tipo das relações estabelecidas por elas. Como exemplo, a relação *Equivalence* entre duas sentenças de textos distintos indica que há informação redundante e, por isso, as sentenças são relevantes para compor o sumário. O número de relações CST das sentenças também pode ser utilizado como critério de seleção, ou seja, sentenças com mais relações CST são consideradas mais relevantes (MANI, 2001).

O primeiro trabalho a utilizar relações discursivas na SAM foi o de Radev e McKeown (1998), no qual as bases da CST foram formuladas.

Zhang et al. (2002) propuseram a troca de sentenças pouco relacionadas por sentenças que apresentam maior número de relações CST, o que resultou em melhora significativa na qualidade dos sumários.

Afantenos et al. (2004, 2008) propuseram teoricamente um método em que a seleção de conteúdo baseia-se na identificação de relações CST entre *templates*. Dada a consulta *"What was the performance of Georgeas during the first three rounds*?(Qual foi a performance de *Georgeas* durante as primeiras três rodadas?)*,* por exemplo, o sistema identifica o *template "performance*", composto pelos argumentos *entity*, *in_what*, *time_span* e *value*, o qual deve ser preenchido por informações extraídas dos textos-fonte para que a pergunta seja respondida por meio do sumário. Para ilustração, considera-se que a sentença 1 do texto 1 e a sentença 1 do texto 2 de dada coleção preenchem os argumentos do *template* com elementos idênticos, p.ex.: *performance* (*georgeas*, *general*, *round_1*, *excellent*). Segundo o método, a relação CST *Identity* conecta os *templates* preenchidos por tais sentenças e, com base no conteúdo redundante codificado no *template*, gera-se uma sentença para compor o sumário.

Quanto a métodos de SAM híbridos, destaca-se o de Schiffman et al. (2002), que se caracteriza por unificar informações superficiais (localização da sentença nos textos-fonte e tamanho da sentenças) e conhecimento léxico-conceitual. Além dos atributos superficiais, os autores relacionam as palavras dos textos-fonte pela sinonímia e hiponímia para delimitar os conceitos mais representativos da coleção. As relações são identificadas pela indexação das palavras à WN.Pr, base léxico-conceitual do inglês norte-americano (FELLBAUM, 1998).

Para o português, há métodos/sistemas de SAM desenvolvidos segundo as abordagens superficial, profunda e híbrida.

Segundo a abordagem superficial, destacam-se os trabalhos de Pardo (2005) e Akabane (2012).

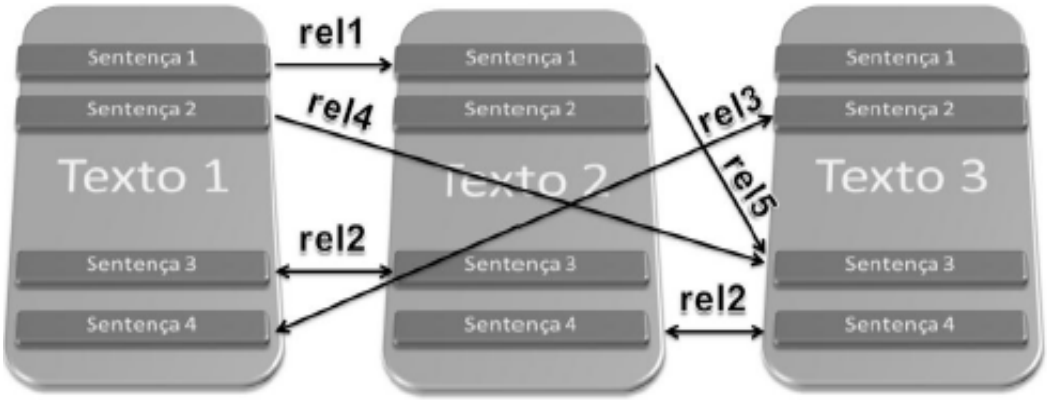
Pardo (2005) desenvolveu o sumarizador superficial GistSumm (PARDO, 2005). O GistSumm pontua e ranqueia as sentenças dos textos de uma coleção com base na frequência de ocorrência de suas palavras na coleção. A sentença de maior pontuação é considerada a *gist sentence* (isto é, sentença que expressa o conteúdo principal da coleção) e selecionada para iniciar o sumário. As demais sentenças que compõem o sumário satisfazem a dois critérios: (i) conter pelo menos um radical em comum com a *gist sentence* e (ii) ter pontuação maior que a média das pontuações de todas as sentenças.

O método desenvolvido por Akabane (2012) e implementado no sistema CNSumm enquadra-se no grupo dos “baseados em grafos”, posto que se pauta na modelagem dos textos-fonte em grafos e redes complexas.

Quanto aos métodos da abordagem profunda para o português, desconhecem-se trabalhos que utilizam informações sintáticas e semântico-conceituais. Os métodos/sistemas de SAM desenvolvidos para o português com base em conhecimento linguístico utilizam majoritariamente conhecimento semântico-discursivo codificado nas relações da CST, como o CSTSumm (CASTRO JORGE; PARDO, 2010).

Nesses trabalhos, a análise consiste na modelagem dos textos-fonte em um grafo, em que as sentenças são representadas por nós e as relações CST por arestas (Figura 7):

Figura 7 – Esquema genérico de análise multidocumento.



Fonte: Maziero (2012).

Na transformação, as sentenças são pontuadas e ranqueadas com base no número de conexões no grafo, sendo que as sentenças com mais conexões ocupam o topo do ranque. Sobre esse ranque, aplicam-se operadores de seleção de conteúdo que codificam preferências do usuário, como “apresentação de informação contextual”. Uma vez ativado, um operador reordena as sentenças no ranque, privilegiando a informação relevante para o usuário. As sentenças são selecionadas a partir do novo ranque até que a taxa de compressão seja atingida.

Segundo a abordagem híbrida, destaca-se o método de Ribaldo et al. (2012),que implementado no sistema RSumm, explora medidas estatísticas aplicadas a grafos e redes em combinação com as relações CST.

Para as pesquisas sobre SAM a partir de textos em português, aliás, conta-se com um *corpus* de referência denominado CSTNews (CARDOSO et al*.,* 2011).

Especificamente, esse *corpus* é composto por 50 coleções ou grupos de textos, sendo que cada coleção versa sobre um mesmo tópico. Os textos são do gênero discursivo "notícias jornalísticas", pertencentes à ordem do relatar (DOLZ; SCHNEWLY, 2004).[20] 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 ou tema compilados de diferentes fontes jornalísticas; (ii) sumários humanos (*abstracts*) mono e multidocumento; (iii) sumários automáticos multidocumento; (iv) extratos humanos multidocumento; (v) anotações CST e RST dos textos, dentre outros dados.

As fontes jornalísticas das quais os textos foram compilados correspondem aos principais jornais *online* do Brasil, a saber: *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).

Quanto aos sumários humanos multidocumento, especificamente, ressalta-se que estes foram construídos manualmente de forma abstrativa, ou seja, com reescrita do conteúdo dos textos-fonte. Além disso, a produção dos mesmos foi guiada por uma taxa de compressão de 70%. Consequentemente, os sumários contêm, no máximo, 30% do

[20] Dolz; Schneuwly (1996) propuseram cinco agrupamentos de gêneros com base em três critérios: domínio social da comunicação a que pertencem; capacidades de linguagem envolvidas na produção e compreensão desses gêneros e sua tipologia geral. Tais critérios são agrupados segundo: a o ordem do narrar, ordem do argumentar, ordem do expor, ordem do descrever ações e ordem do relatar. Em especial, a ordem do relatar comporta os gêneros pertencentes ao domínio social da memorização e documentação das experiências humanas, situando-as no tempo. Exemplos: relato de experiências vividas, diários íntimos, diários de viagem, notícias, reportagens, crônicas jornalísticas, relatos históricos, biografias, autobiografias, testemunhos etc.

número de palavras do maior texto-fonte da coleção. Do ponto de vista da audiência, os sumários do CSTNews são genéricos, pois não foram construídos com o foco em leitores específicos e, do ponto de vista funcional, são informativos, pois contemplam as informações principais de seus textos-fonte, substituindo a leitura dos mesmos.

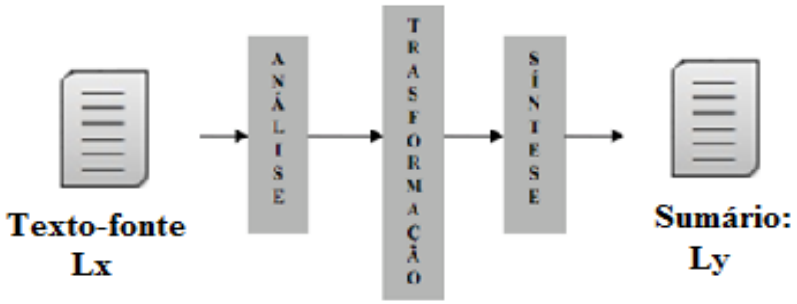
2.3 A Sumarização Automática e a Multiplicidade de línguas

Dada a grande quantidade de informação disponível em várias línguas na *web*, o multilinguismo também tem sido focado no âmbito da SA com o objetivo de permitir o acesso à informação que inicialmente tenha sido veiculada em uma língua distinta da do usuário. Os métodos/sistemas de SA que envolvem mais de uma língua podem ser organizados em 3 grupos: (i) *cross-language*, (ii) multilíngue e (iii) independentes de língua (ORĂSAN, 2009).

2.3.1 Os métodos/sistemas cross-language

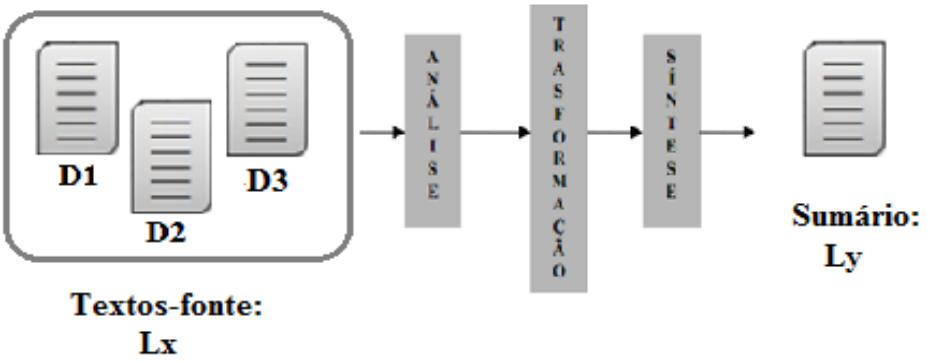
Os métodos/sistemas *cross-language* podem ser monodocumento (Figura 8) ou multidocumento (Figura 9). Em ambos os casos, a língua do sumário é diferente da língua do(s) texto(s)-fonte. Em outras palavras, o(s) texto(s)-fonte estão em uma língua (Lx) e o sumário em outra língua (Ly):

Figura 8 – Esquema da sumarização *cross-language* monodocumento.



Fonte: Figura elaborada pelo autor.

Figura 9 – Esquema da sumarização *cross-languae* multidocumento.



Fonte: Elaborado pelo autor.

De um modo geral, a SA *cross-language* pode seguir duas abordagens, a abordagem *early translation* ou *late translation,* assim, o processo de sumarização envolve necessariamente a etapa de tradução automática (TA) dos textos-fonte ou dos sumários.

Ressalva-se, no entanto, que a qualidade dos textos traduzidos automaticamente ainda está longe de ser satisfatória, apesar do avanço dos últimos anos. Ainda hoje, as traduções apresentam, no geral, inadequações linguísticas, agramaticalidade e baixa legibilidade, causadas, na maioria das vezes, por (i) a falta de conhecimento linguístico, (ii) a formulação de hipóteses erradas dos sistemas sobre o funcionamento das línguas e (iii) a complexidade inerente à tarefa de tradução (WAN; LI; XIAO, 2010).

Assim sendo, como os tradutores automáticos não são aplicações de PLN totalmente precisas, produzindo, por conseguinte, textos com os problemas mencionados, a abordagem *late translation* apresenta certa vantagem frente à *early translation*, pois os textos-fonte não são traduzidos na íntegra, mas sim apenas as sentenças selecionadas para compor o sumário (WAN; LI; XIAO, 2010). Ademais, ressalta-se que os problemas gerados pela TA dos textos-fonte na abordagem *early translation* podem influenciar na aplicação dos métodos de SA mono e multidocumento.

Um exemplo de método/sistema *cross-language* monodocumento é o de Wan et al. (2010), o qual gera um sumário em chinês a partir de um texto-fonte jornalístico em inglês. Para tanto, utiliza-se a abordagem *late translation* e, por isso, o texto-fonte em

inglês é sumarizado e, em seguida, o sumário é traduzido para o chinês por meio do serviço *on-line Google Translate*[21].

As etapas realizadas pelo sistema de SA em questão subdividem-se em:

(i) determinar a predição da qualidade de tradução de cada sentença do texto-fonte em inglês, a qual se baseia em um conjunto de atributos linguísticos superficiais (p.ex.: tamanho da sentença, porcentagem de substantivos e adjetivos e ocorrência de *wh-words*, como *who, what, where*, etc.) e profundos (p.ex.: quantidade de sintagmas nominais e verbais);
(ii) calcular a informatividade de cada sentença dos textos-fonte por meio do sistema MEAD (RADEV et al., 2004);
(iii) selecionar as sentenças do texto-fonte em inglês com base na combinação da predição da qualidade de tradução e o cálculo da informatividade, e
(iv) traduzir automaticamente as sentenças selecionadas para o chinês.

Com base nas etapas de SA do método de Wan et al. (2010), observa-se que a predição da qualidade de tradução das sentenças é um critério de seleção previsto para minimizar os problemas gerados pela TA.

Quanto aos sistemas de SA *cross-language* multidocumento, em que se parte de uma coleção de textos em uma língua Lx que abordam um mesmo assunto para produzir um sumário em uma língua Ly, destacam-se os trabalhos de Orăsan e Chiorean (2008) e Bourdin et al. (2011).

Orăsan e Chiorean (2008) apresentam um sistema em que uma coleção de textos-fonte em romeno é sumarizada por um método superficial de SAM e o sumário é traduzido para o inglês por um tradutor de livre acesso romeno-inglês[22]. Dessa forma, a abordagem característica desse método é a *late translation*.

O método de Orăsan e Chiorean (2008) é *query-based* e, por isso, o sumário deve apresentar a informação da coleção que satisfaz a uma consulta do usuário. Assim, as sentenças dos textos-fonte são pontuadas e ranqueadas por meio da medida estatística do cosseno, que captura a similaridade lexical entre elas e a consulta do usuário. No caso, as sentenças mais similares à consulta ocupam o topo do ranque. Para compor o sumário, somente as mais bem pontuadas que apresentam pouca similaridade ou

[21]Disponível em: <http://translate.google.com/>.
[22]Disponível em: <http://www.etranslator.ro>.

redundância entre si são selecionadas. Dessa forma, pode-se dizer que as sentenças são ranqueadas com base em um critério combinado de relevância e novidade de informação.

Outro método *cross-language* multidocumento é o de Bourdin et al (2011), em que um sumário em francês é gerado a partir de uma coleção de textos-fonte jornalísticos em inglês. No caso, ao contrário de Orăsan e Chiorean (2008), os textos-fonte em inglês são primeiramente traduzidos para o francês pelo *Google Translate* e, posteriormente, sumarizados. Assim, a abordagem *cross-language* é a *early translation.*

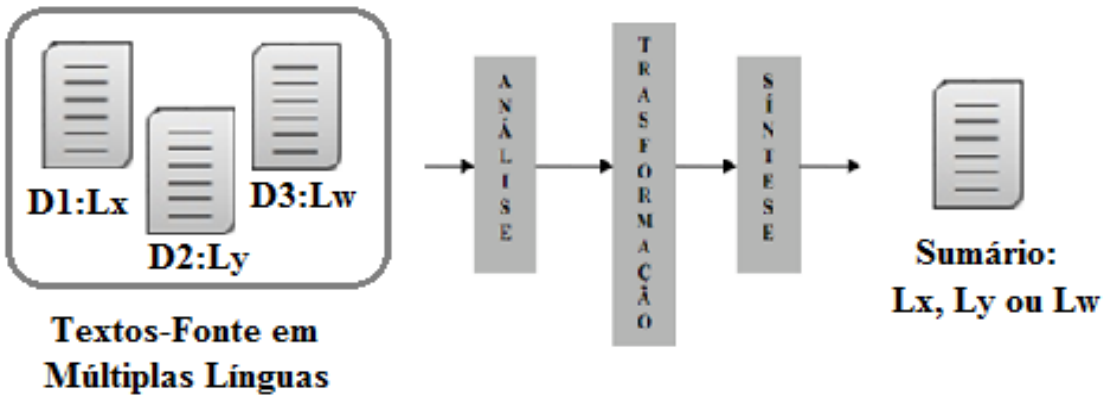
Assim como em Wan et al. (2010), a seleção de conteúdo em Bourdin et al (2011) é baseada na determinação da qualidade de tradução e informatividade das sentenças. Como o método *cross-language* multidocumento de Bourdin et al (2011) é superficial baseado em grafo (*graph-based methods*), a qualidade da tradução e a informatividade são critérios inseridos no grafo.

Especificamente, nesses grafos, as sentenças são codificadas em nós e a similaridade entre elas é modelada como arestas que conectam os nós. Assim, as sentenças são pontuadas pela sua conectividade, que indica informatividade, e pela qualidade de tradução.

2.3.2 Os métodos/sistemas multilíngue

Os métodos/sistemas de SA classificados como multilíngue partem necessariamente de uma coleção de textos, em diferentes idiomas, que abordam o mesmo assunto. Assim, tem-se a chamada SA multidocumento multilíngue (SAMM), como na Figura 10:

Figura 10 – Esquema da Sumarização Multidocumento Multilíngue.



Fonte: Elaborado pelo autor.

Nesse grupo, destacam-se os trabalhos de Evans et al. (2004, 2005).

Evans et al. (2004) desenvolveram uma versão multilíngue para o sistema de sumarização *online* denominado *Columbia Newsblaster*. Nela, a SA tem início com uma coleção composta por textos jornalísticos em inglês e, por exemplo, em russo, sobre um mesmo evento.

Para uma mesma coleção, realizam-se dois processos de SA. Em um deles, os textos em russo são traduzidos automaticamente para o inglês e, juntamente com os textos originais em inglês, são submetidos ao processo de SAM. No outro, os textos traduzidos para o inglês e os originais em inglês são submetidos separadamente ao processo de SAM. Nesse último caso, os autores disponibilizam a visualização dos 2 sumários para o usuário familiarizado com a língua inglesa, de tal forma que este pode comparar as diferenças e semelhanças de conteúdo entre os sumários provenientes de textos em inglês e de textos em russo sobre o mesmo assunto.

Para a sumarização dos textos-fonte traduzidos (para o inglês) e dos originais em inglês, o *Columbia Newsblaster* utiliza um de dois sistemas de SAM em função da similaridade dos textos.

Quando os textos-fonte são muito similares, o sistema utilizado é o MultiGen (McKeown et al., 1999) que, a partir do agrupamento das sentenças similares em *clusters,* realiza a análise sintática das mesmas e a fusão de informação para a produção do sumário.

Quando os textos são muito distintos, o sistema utilizado no ambiente *Columbia Newsblaster* é o DEMS (*Dissimilarity Engine for Multi-document Summarization*) (SCHIFFMAN et al.,2002), que se baseia na utilização de várias estratégias superficiais e profundas para pontuar e ranquear as sentenças de um *cluster*.

As estratégias superficiais do DEMS são baseadas em:

(i) localização: segundo a qual sentenças que ocorrem no início dos textos são privilegiadas em detrimento de sentenças que ocorrem no meio e fim dos documentos;

(ii) data de publicação: segundo a qual sentenças provenientes de textos mais recentes são privilegiadas em detrimento de sentenças provenientes de textos mais antigos;

(iii) tamanho: critério segundo o qual sentenças de tamanho médio são privilegiadas em detrimento de sentenças abaixo de um limiar inferior (p. ex.: 15 palavras) e acima de um limiar superior (p.ex.: 30 palavras);

(iv) *lead words*: segundo a qual as sentenças que possuem as palavras que ocorrem nas primeiras sentenças do texto são privilegiadas porque veiculam os tópicos principais do mesmo;

As estratégias mais profundas, por sua vez, são baseadas em:

(ii) verbos semanticamente relevantes: segundo a qual sentenças que apresentam verbos plenos que ocorrem associados a sujeitos específicos tendem a transmitir uma informação completa e, por isso, são privilegiadas em detrimento de sentenças que apresentam verbos semanticamente mais vazios. Por exemplo, um verbo como "prisão" sugere uma atividade policial; mas os verbos menos específicos (como "ser" ou "fazer") ocorrem com uma grande variedade de temas e objetos.

(iii) conceitos: segundo a qual as sentenças constituídas pelos conceitos mais frequentes da coleção são privilegiadas em detrimento das demais; para identificar os conceitos subjacentes às palavras, os autores utilizam os *synsets* e a relação de hiponímia/hiperonímia da WN.Pr (FELLBAUM, 1998).

Caso sentenças traduzidas sejam selecionadas para compor o sumário, o sistema DEMS foi modificado por Evans et al. (2004) para identificar sentenças originais similares e substituí-las no sumário. Especificamente, o DEMS passou a englobar o sistema de detecção de similaridade denominado SimFinder (HATZIVASSILOGLOU et al., 2001), que se baseia no compartilhamento de: (i) nomes próprios, (ii) itens lexicais morfologicamente relacionados, (iii) itens lexicais sinônimos, (v) itens lexicais com mesmo hiperônimo e (iv) núcleos sintagmáticos.

Em Evans et al. (2005), por sua vez, o método de SAMM parte de um *corpus* bilíngue formado unicamente por textos em inglês e em árabe com o objetivo de produzir um sumário em inglês. Os textos em árabe são traduzidos para o inglês e somente as traduções são submetidas ao processo de seleção de conteúdo.

Uma vez que a coleção passa a ser monolíngue multidocumento, os autores aplicam o sistema DEMS (SCHIFFMAN et al., 2002) para selecionar a sentenças que comporão os sumários. Em Evans et al. (2005), o sistema DEMS simplifica as sentenças (ou seja, quebra as sentenças complexas em menores) antes de pontuar e ranqueá-las.

Tendo em vista que as sentenças (simplificadas) selecionadas para compor o sumário podiam apresentar problemas de gramaticalidade e/ou inteligibilidade gerados pela TA do árabe para o inglês, os autores utilizaram o sistema SimFinder (HATZIVASSILOGLOU et al., 2001) para identificar sentenças originais em inglês que são similares às traduzidas, as quais são efetivamente levadas ao sumário. Ao final, o

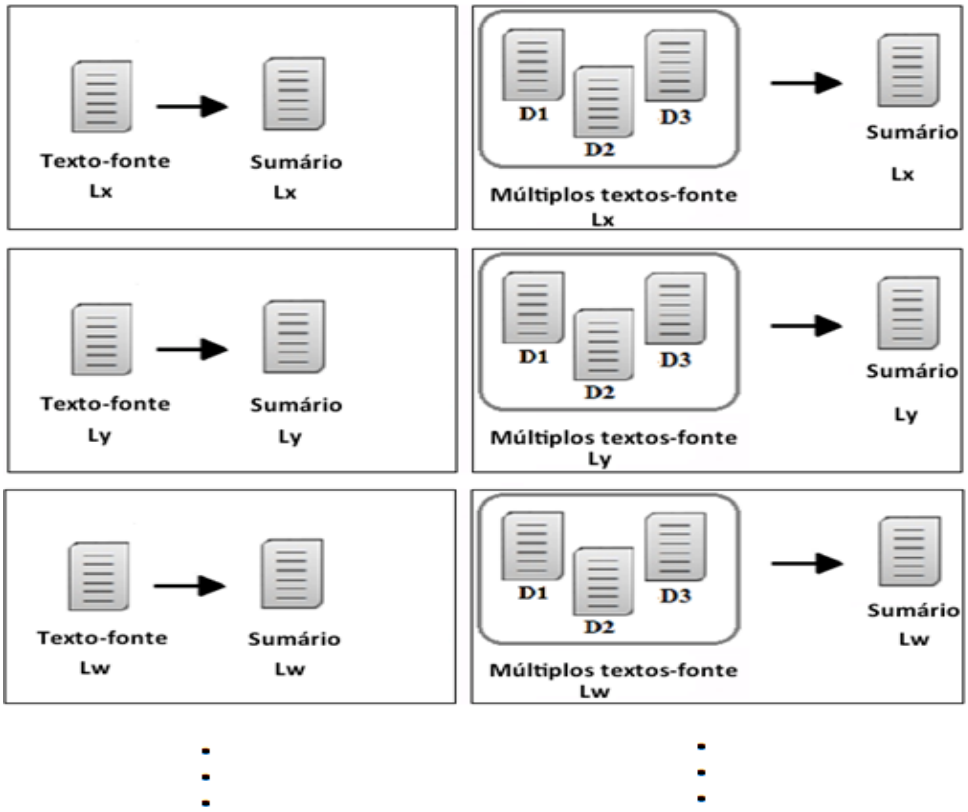
sumário é composto apenas por essas sentenças originais. Segundo Evans et al. (2005), a substituição por sentenças similares originais melhorou em 68% a inteligibilidade do sumário, sem prejudicar a informatividade.

2.3.3 Os métodos/sistemas independentes de língua

Os métodos/sistemas independentes de língua caracterizam-se por não utilizarem conhecimento linguístico profundo para a seleção de conteúdo. De maneira geral, tais métodos/sistemas utilizam apenas conhecimento linguístico superficial e/ou conhecimento empírico estatístico e, por isso, processam diferentes línguas, seguindo o pressuposto de que o conhecimento superficial/estatístico é capaz de generalizar fenômenos que são recorrentes na maioria das línguas naturais.

Assim sendo, a SA independente de língua pode ser mono ou multidocumento (Figura 11):

Figura 11– Esquema de SA independente de língua mono e multidocumento



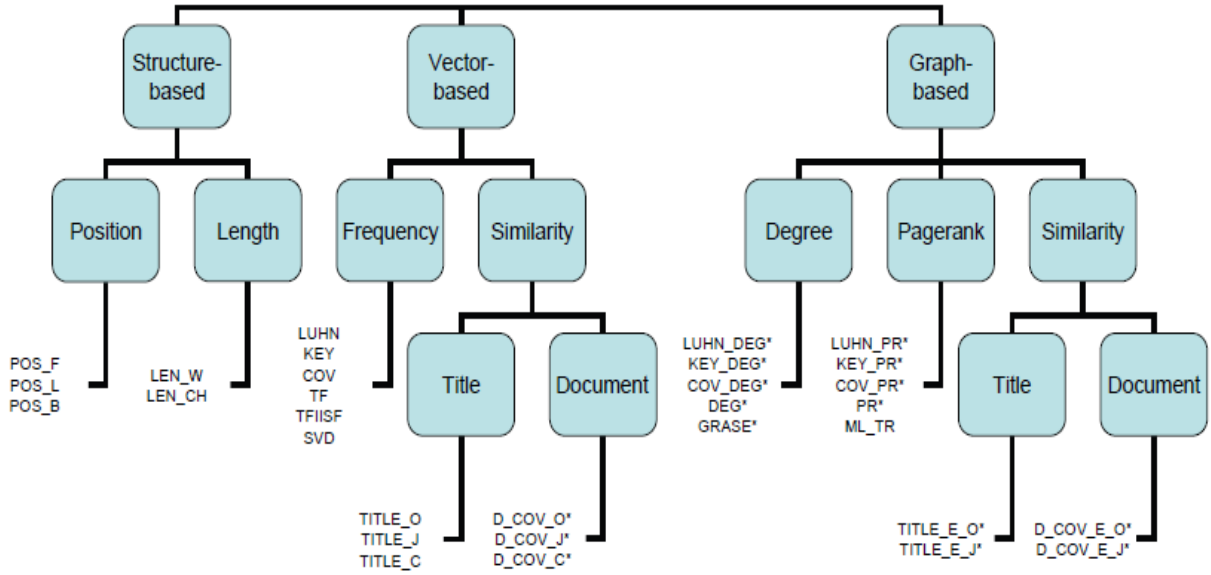
Fonte: Elaborado pelo autor.

Nesse cenário, destaca-se a plataforma *on-line* denominada MEAD[23], que disponibiliza: (i) métodos de SA independentes de língua, que podem ser aplicados em combinação ou de forma isolada, e (ii) métodos intrínsecos e extrínsecos de avaliação automática de sumários (RADEV et al., 2004).

Especificamente, os métodos de SA independentes de língua disponíveis no MEAD são baseados em: (i) centroide; (ii) localização (da sentença no texto-fonte); (iii) similaridade lexical com a primeira sentença do texto-fonte (ou com o título); (iv) tamanho da sentença, (v) palavras-chave, etc (RADEV et al., 2004).

Outros métodos independentes de língua podem ser encontrados em Litvak et al. (2010). Especificamente, os autores investigaram 31 métodos, os quais podem ser agrupados em 3 grandes classes (Figura 12): (i) baseados na estrutura (textual) (do inglês, *structure-based methods*); (ii) baseados na representação dos textos em vetores (do inglês, *vector-based methods*), e (iii) baseados na representação dos textos-fonte em grafos (do inglês, *graph-based methods*).

Figura 12 – Métodos de SA independentes de língua de Litvak et al. (2010).



Fonte: Litvak et al. (2010).

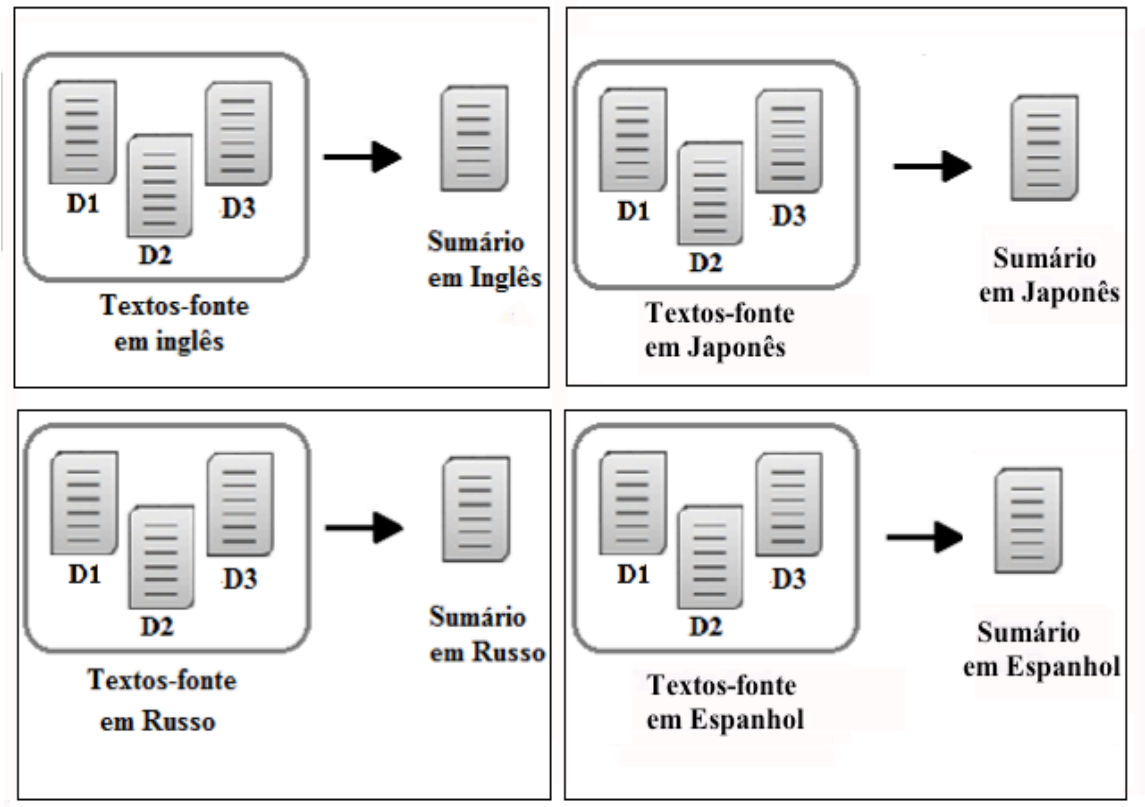
Além dos trabalhos citados, ressalta-se também o de Cowie et al. (1998), em que se desenvolveu um sistema de SA independente de língua denominado MINDS, capaz de lidar com os cenários ilustrados na Figura 13.

[23]Disponível em: <http://www.summarization.com/mead/>.

Com base na Figura 13, observa-se que o foco dos autores era efetivamente o desenvolvimento de um sistema capaz de gerar um sumário multidocumento na língua de origem dos textos-fonte, independentemente de qual fosse essa língua (inglês, russo, japonês ou japonês).

Para tanto, a seleção das sentenças no MINDS é feita com base em conhecimento linguístico superficial, como (i) localização e (ii) frequência das palavras.

Figura 13 – Sumarização independente de língua em Cowie et al. (1998)



Fonte: Elaborado pelo autor.

2.3.4 A Sumarização Automática, a Multiplicidade de línguas e o Português

Para o português, desconhecem-se trabalhos que focam o desenvolvimento de métodos/sistemas *cross-language* e independentes de língua que partem de coleções com textos em mais de uma língua.

O único trabalho conhecido é o de Tosta et al. (2013), em que foi investigada a aplicação manual de 2 métodos *baseline* de SAMM capazes de sumarizar uma coleção composta por 3 textos jornalísticos, cada um deles em uma língua distinta (inglês, espanhol e português), e gerar um sumário em português.

Para a investigação dos métodos *baseline*, os autores construíram o primeiro *corpus* multilíngue que engloba o português, o qual recebeu a denominação CM3News (*Corpus Multidocumento Trilíngue de Textos Jornalísticos*)[24]. O CM3News é composto por 10 coleções de textos jornalísticos, advindos de fontes distintas, e nas línguas mencionadas (inglês, espanhol e português) (TOSTA et al., 2012).

No Quadro 2, apresenta-se o CM3News, que totaliza 16.139 palavras.

Quadro 2 – Coleções do CM3News.

Coleção	Domínio	Assunto/Tema	Documento	Língua	Publicação (data/ hora)	No. de palavras
C1	Mundo	Ataques a Londres	D1_C1_folha	PT	11/08/2011 - 09:11	1.721
			D2_C1_bbc	IN	11/08/2011-11:10 (GMT)	
C2	Poder	Kit gay	D1_C2_folha	PT	25/05/2011-13:12	895
			D2_C2_bbc	IN	25/05/2011- 21:07 (GMT)	
C3	Saúde	Intoxicação alimentar	D1_C3_folha	PT	30/05/2011 - 18:47	1.915
			D2_C3_bbc	IN	30/05/2011 - 5:43 (GMT)	
C4	Mundo	Massacre na Noruega	D1_C4_folha	PT	08/08/2011 - 14h20	1.613
			D2_C4_bbc	IN	02/08/2011-14:52 (GMT)	
C5	Ambiente	Novo código florestal	D1_C5_folha	PT	25/05/2011– 00:43	2.093
			D2_C5_bbc	IN	25/05/2011– 09:50 (GMT)	
C6	Mundo	Conflito na universidade da CA	D1_C6_folha	PT	20/11/2011– 00:15	1.193
			D2_C6_bbc	IN	21/11/2011– 23:26 (GMT)	
C7	Saúde	Proibição do fumo em NY	D1_C7_folha	PT	24/05/2011– 13:38	1.224
			D2_C7_bbc	IN	24/05/2011– 18:36 (HKT)	
C8	Mundo	Terremoto na Nova Zelândia	D1_C8_folha	PT	05/03/2011– 05:01	1.329
			D2_C8_bbc	IN	03/03/2011– 04:45 (GMT)	
C9	Mundo	Terremoto em Missouri	D1_C9_folha	PT	23/05/2011– 08:04	1.880
			D2_C9_bbc	IN	23/05/2011– 20:21 (GMT)	
C10	Mundo	Erupção vulcânica na Islândia	D1_C10_folha	PT	24/05/2011– 12:13	2.276
			D2_C10_bbc	IN	24/05/2011– 15:51 (GMT)	
Total						**16.139**

Fonte: Tosta et al. (2012)

Os textos-fonte do CM3News foram coletados manualmente da *web*, ou seja, as páginas desejadas foram acessadas uma a uma e o textos de interesse foram salvos no computador um a um. Especificamente, Tosta et al. (2012) selecionaram as versões eletrônicas dos seguintes jornais: (i) *A Folha de São Paulo*[25], para a coleta dos textos em português; (ii) *BBC News*[26], para a coleta dos textos em inglês, e (iii*) El pais*[27], para os textos em espanhol. Tais fontes foram selecionadas por serem versões eletrônicas de

[24] Disponível em: <http://www.nilc.icmc.usp.br/arianidf/sustento/resources.html>.
[25] Disponível em: <http://www.folha.uol.com.br/>.
[26] Disponível em: <http://www.bbc.co.uk/news/>.
[27] Disponível em: <http://elpais.com/>.

jornais importantes quanto à circulação e qualidade das notícias.

Para a seleção dos domínios, os autores optaram pelos critérios "atualidade" e "variedade", buscando selecionar diferentes domínios e assuntos em circulação na mídia jornalística.

No Quadro 2, observa-se também que os textos foram nomeados segundo o padrão *documento_coleção_fonte_formato* (p.ex.: `D1_C1_folha.txt`). A numeração dos documentos na coleção indica a língua de origem do mesmo, sendo 1 para os textos em português e 2 para os textos em inglês. No exemplo em questão, a nomeação `D1_C1_folha.txt` indica que se trata do texto em português da coleção 1. Ressalta-se que os textos foram armazenados sem os seus respectivos títulos, posto que estes não são processados na SA. Os títulos dos textos foram armazenados, também em formato txt, em uma pasta específica, denomina "Títulos".

Quanto à seleção de conteúdo, Tosta et al. (2013) aplicam dois métodos que são tidos como *baseline* porque (i) partem de uma coleção composta por textos traduzidos para o português e originais e (ii) utilizam conhecimento superficial para a seleção das sentenças.

Os métodos investigados por Tosta et al. (2013) caracterizam-se por englobar a etapa de TA integral dos textos-fonte para o português antes do processo de seleção de conteúdo. Dessa forma, tais métodos seguem a abordagem *early-translation* e se baseiam em conhecimento linguístico superficial para a seleção do conteúdo a compor o sumário multidocumento.

Uma vez traduzidos para o português, os autores aplicaram aos textos-fonte métodos clássicos de SA comumente utilizados no cenário multidocumento, a saber: (i) localização (Método 1) e (ii) frequência das palavras de classe aberta (Método 2) (GUPTA, LEHAL, 2010). A seguir, descrevem-se detalhadamente os métodos.

- **Método 1:** com base no critério da localização, as sentenças são caracterizadas em função da sua posição no texto-fonte da coleção. As sentenças contidas no primeiro parágrafo de cada um dos 3 textos são especificadas com o atributo localização="inicio", as sentenças localizadas no último parágrafo com o atributo localização="fim", e as demais, com o atributo localização="meio. Assim, o topo do ranque é ocupado pelas sentenças "início", seguidas pelas sentenças "meio" e, por fim, pelas sentenças "fim". A partir do ranque, a seleção manual de conteúdo no Método 1 consiste em: (i) selecionar a sentença de maior pontuação do ranque para

iniciar o sumário; (iii) selecionar a próxima sentença do ranque; (iv) calcular a redundância entre a nova sentença candidata e a sentença já selecionada para o sumário; (v) selecionar a sentença candidata para compor o sumário se esta contiver pouca similaridade com a sentença inicialmente selecionada e não apresentar problemas de TA, (vi) substituir a sentença selecionada não-redundante com problemas de tradução por uma similar proveniente do texto-fonte original em português, (vii) repetir os passos para as demais sentenças do ranque até que a taxa de compressão de 70% fosse atingida. A similaridade, tanto para eliminar a redundância como para substituir sentenças traduzidas agramaticais por originais em português, é calculada de forma automática com base na medida estatística *wordoverlap* que se baseia na sobreposição das palavras de classe aberta idênticas (JURAFSKY, MARTIN, 2001). O cálculo *word overlap* entre sentenças é feito por meio da aplicação da fórmula, descrita em (2).

(2)

$$Wol\ (S1, S2) = \frac{\#CommonWords}{\#Words(S1)\ + \#Words\ (S2)}$$

Para calcular a *word overlap* entre um par de sentenças (S1 e S2), divide-se o número total de palavras idênticas entre as sentenças (*Common Words*) pela soma do número total de palavras de cada sentença (*Words(S1) + Words(S2)*), excluindo-se as *stopwords,* números e símbolos). O resultado obtido será entre 0 e 1, sendo que, quanto mais próximo de 1 for a *Wol,* mais redundante será o par entre si, e, quanto mais próximo de 0, menos redundante. A produção dos sumários foi manual pela justaposição das sentenças na ordem em que foram selecionadas.

- **Método 2:** dada uma coleção, as sentenças dos textos-fonte recebem uma pontuação resultante da soma da frequência de ocorrência na coleção de suas palavras de classe aberta, a partir da qual são ranqueadas em ordem decrescente. Assim, o topo do ranque é ocupado pelas sentenças compostas pelas palavras mais frequentes. A pontuação e o ranqueamento são feitos por uma funcionalidade do sumarizador *GistSumm* (Pardo, 2005). Com base no ranque, a seleção manual de conteúdo no Método 2 segue os mesmos passos do Método 1, já que engloba o tratamento da

redundância e dos problemas gerados pela TA. A produção dos sumários também é manual pela justaposição das sentenças na ordem em que foram selecionadas.

Para verificar o desempenho dos 2 métodos, os autores submeteram os sumários gerados manualmente a uma avaliação intrínseca de qualidade, que consistiu na análise da legibilidade (ou fluência) dos sumários gerados para 5 das 10 coleções do *corpus* CM3News. Os sumários foram analisados por 1 especialista em função dos 5 parâmetros utilizados na DUC: (i) gramaticalidade, (ii) não-redundância, (iii) clareza referencial, (iv) foco temático, e (v) estrutura/coerência. Na Tabela 1, esquematiza-se a média obtida por cada método em uma escala de 1 a 5.

Tabela 1 – Média das pontuações dos métodos *baseline* de Tosta et al. (2013).

Critérios	**Método 1**	**Método 2**
Gramaticalidade	**3**	2,8
Não-redundância	**3**	3
Clareza referencial	**3,2**	3
Foco temático	**4**	3,8
Estrutura e Coerência	**2,8**	2,4

Nessa avaliação, constatou-se que o Método 1, pautado na localização com tratamento da redundância e da tradução, obteve em média as mais altas pontuações quanto aos 5 parâmetros. Além disso, constatou-se que, apesar da aplicação da similaridade para a substituição das sentenças traduzidas por originais em português, os sumários ainda apresentam problemas de gramaticalidade, posto que, dentre os 5 parâmetros da DUC, esse foi o que obteve as médias mais baixas. Uma possível explicação reside no fato de que alguns sumários apresentam algumas sentenças traduzidas que não eram redundantes, mas que possuíam problemas de tradução.

Na próxima Seção, apresentam-se detalhadamente as formas de avaliação na SA.

2.4 A avaliação na Sumarização Automática

2.4.1 Avaliação intrínseca da qualidade linguística

A avaliação intrínseca na SA objetiva avaliar os métodos e/ou sistemas quanto à qualidade linguística e informatividade. Na medida em que a tarefa de sumarização torna-se mais complexa, como por exemplo, pelo aumento do número de textos-fonte e /ou aumento do número de línguas, os problemas linguísticos tendem a aumentar na mesma proporção.

Dessa forma, alguns problemas típicos da SA (monodocumento), como por exemplo, problemas relativos à coesão e coerência, somam-se aos problemas típicos da SAM (p. ex. contradição e redundância). Ao se acrescentar a multiplicidade de línguas ao processo, os problemas linguísticos mencionados recebem ainda outro agravante, típico do contexto de SAMM (quando envolvendo a TA), o problema da agramaticalidade.

A agramaticalidade, além de piorar a qualidade linguística do sumário, pode comprometer a aplicação de métodos de SA ou SAM, sobretudo na abordagem *early translation* (na qual versões integralmente traduzidas automaticamente dos textos-fonte são submetidos ao processo de sumarização). Tal abordagem, pode utilizar como entrada textos-fonte com conteúdo agramatical, o que pode comprometer a eficiência de aplicação dos diversos métodos de sumarização. Tal comprometimento ocorre pois, uma vez traduzidos automaticamente, os textos-fonte que servirão de entrada para a SA passam a possuir agramaticalidades que podem comprometer o funcionamento dos sistemas, que são elaborados comumente, para processarem textos gramaticais.

Os critérios propostos pela DUC (2007) para avaliar a qualidade dos sumários são: (i) gramaticalidade; (ii) não-redundância; (iii) clareza referencial; (iv) foco e (v) estrutura e coerência:

(i) Gramaticalidade: para uma boa avaliação neste critério, o sumário não deve ter erros referentes ao mal posicionamento de letras maiúsculas ou minúsculas, problemas de formatação, fragmentos ou pedaços de cabeçalho (p.ex.: datas e/ou subtítulos, componentes textuais perdidos, etc.), além, obviamente, da presença de sentenças agramaticais, que tornem o texto difícil de ler.

(ii) Não-redundância: no sumário não deve haver repetição desnecessária. Tal repetição pode assumir a forma de sentenças inteiras que se repetem no sumário, fatos

repetidos ou o uso repetitivo de um substantivo ou de um sintagma nominal (p. Ex. “Dilma Rousseff”) quando um pronome como “ela” pudesse ser empregado.

(iii) Clareza referencial: o sumário deve tornar claro a identificação de “quem” ou “o que” os pronomes e sintagmas nominais referem-se. Se uma pessoa ou outra entidade são mencionadas, seus papéis devem ser claros na história. Dessa forma, uma referência não é clara se ela for referida, mas sua identidade ou relação contextual permanecer obscura.

(iv) Foco: o sumário deve ter um foco (temático); as sentenças devem conter apenas informações que sejam relacionadas com o resto do sumário.

(v) Estrutura e Coerência: o sumário deve ser bem estruturado e bem organizado, e não um amontoado de informações relacionadas. É preciso que as sentenças sejam arquitetadas de forma a construir uma coerente estrutura informativa sobre um determinado tópico.

2.4.2 Avaliação intrínseca de informatividade

Especificamente, as medidas utilizadas pela ROUGE medem a coocorrência de n-gramas. O n-grama na medida ROUGE é considerado como uma palavra ou um conjunto de palavras que ocorrem em sequência, podendo variar de 1(unigrama) até 4 palavras. De acordo com o número de n-gramas, a ROUGE pode subdividida em medidas diferentes. A ROUGE-1, por exemplo, mede a coocorrência de unigramas, enquanto que a ROUGE-2, de bigramas.

Os resultados da ROUGE são dados em termos de precisão do inglês *precision* (P), cobertura (C) (*recall*) e medida-f (*f-measure*).

A precisão e a cobertura expressas pela ROUGE são representadas pela fórmula:

$$\mathbf{P=}\ \frac{\textbf{Número de n-gramas em comum com o sumário de referência}}{\textbf{Número de n-gramas do sumário automático}}$$

$$\mathbf{C=}\ \frac{\textbf{Número de n-gramas em comum com o sumário de referência}}{\textbf{Número de n-gramas do sumário de referência}}$$

A medida-f (*F-Measure*), representada pela fórmula ((PxC)/(P+C)) x 2, combina as métricas de cobertura e de precisão. O resultado da medida-f é um indicativo de que,

quanto mais próximo de 1, mais informativo é o sumário, e mais próximos de 0 demonstra que o sumário é ruim, quanto à informatividade.

Em resumo, os resultados da utilização de métricas automáticas, como as da ROUGE, viabilizam a avaliação intrínseca da informatividade de maneira automática, calculando a quantidade de informação que um sumário genérico preserva da fonte, em diferentes graus de compressão. No caso, compara-se sumários automáticos com sumários de referência, calculando quanto da informação presente no sumário de referência é preservada no sumário automático

3 CONSTRUÇÃO E ANOTAÇÃO DE CORPUS

Tendo em vista os objetivos deste trabalho, construiu-se um *corpus*. Por *corpus*, entende-se um conjunto de dados linguísticos, sistematizados de acordo com critérios determinados, de forma que possa ser processado por um computador, com a finalidade de proporcionar resultados diversos e úteis para a descrição e análise linguística (SINCLAIR, 2005). Dessa forma, vê-se que os *corpora* são recursos que devem apresentar características que satisfaçam as necessidades da pesquisa para a qual é construído.

3.1 A construção do corpus

Para este trabalho, o *corpus* devia apresentar as seguintes características: (i) multidocumento, (ii) multilíngue e (iii) jornalístico. As características (i) e (ii) resultaram do interesse em lidar com as informações quando estas são publicadas por diferentes fontes e em diferentes línguas. Sobre as línguas-fonte, ressalta-se que a escolha do português e do inglês foi feita com o objetivo de produzir sumários em português a partir de textos nessa mesma língua e em inglês, língua em que a maior parte da informação está disponível na *web*[28]. A escolha pelo gênero jornalístico foi feita em função da tradição dos trabalhos em SAM, que comumente focam esse gênero, e devido à facilidade de obtenção de textos desse gênero que versam sobre um mesmo assunto a partir de fontes distintas e em diferentes línguas (no caso, português e inglês).

Dos recursos disponíveis em português, destaca-se o CM3News (TOSTA et al., 2012; 2013) que aliás, foi o primeiro *corpus* com coleções de textos que versam sobre o mesmo assunto construído para a SAMM envolvendo o português.

Tendo em vista o objetivo de se trabalhar com coleções bilíngues (inglês e português), optou-se por construir um *corpus* a partir de um recorte e extensão do CM3News. Diz-se recorte porque, das 10 coleções trilíngues do CM3News, apenas os textos em inglês e em português foram considerados, excluindo-se, assim, os textos em espanhol, o que resultou em um conjunto inicial de 10 coleções bilíngues. Diz-se extensão porque, partindo-se das coleções bilíngues recortadas do CM3News, foram acrescidas outras 10.

[28]Disponível em: <http://pt.wikipedia.org/wiki/Uso_da_Internet_no_mundo>.

O produto desse processo de "recorte e extensão" gerou o *corpus* utilizado nesta investigação, o CM2News (*Corpus Multidocumento Bilíngue de Textos Jornalísticos*)[29]. O CM2News é um *corpus* com 20 coleções com 19.984 palavras. No Quadro 3, as coleções finais do CM2News são descritas:

Quadro 3 – Coleções do CM2News.

Coleção	Domínio	Assunto/ Tema	Documento	Língua	Publicação (data/hora)	No. palavras
C1	Mundo	Ataques em Londres	D1_C1_folha	PT	11/08/2011 – 09:11	1.311
			D2_C1_bbc	IN	11/08/2011 – 11:10 (GMT)	
C2	Poder	Kit gay	D1_C2_folha	PT	25/05/2011 – 13:12	516
			D2_C2_bbc	IN	25/05/2011 – 21:07 (GMT)	
C3	Saúde	Intoxicação alimentar	D1_C3_folha	PT	30/05/2011 – 18:47	1.419
			D2_C3_bbc	IN	30/05/2011 – 5:43 (GMT)	
C4	Mundo	Massacre na Noruega	D1_C4_folha	PT	08/08/2011 – 14h20	911
			D2_C4_bbc	IN	02/08/2011 – 14:52 (GMT)	
C5	Ambiente	Novo código florestal	D1_C5_folha	PT	25/05/2011– 00:43	1.217
			D2_C5_bbc	IN	25/05/2011– 09:50 (GMT)	
C6	Mundo	Conflito na universidade da CA	D1_C6_folha	PT	20/11/2011– 00:15	645
			D2_C6_bbc	IN	21/11/2011– 23:26 (GMT)	
C7	Saúde	Proibição do fumo em NY	D1_C7_folha	PT	24/05/2011– 13:38	887
			D2_C7_bbc	IN	24/05/2011– 18:36 (HKT)	
C8	Mundo	Terremoto na Nova Zelândia	D1_C8_folha	PT	05/03/2011– 05:01	948
			D2_C8_bbc	IN	03/03/2011– 04:45 (GMT)	
C9	Mundo	Terremoto em Missouri	D1_C9_folha	PT	23/05/2011– 08:04	1.169
			D2_C9_bbc	IN	23/05/2011– 20:21 (GMT)	
C10	Mundo	Erupção vulcânica na Islândia	D1_C10_folha	PT	24/05/2011– 12:13	1.476
			D2_C10_bbc	IN	24/05/2011– 15:51 (GMT)	
C11	Ciência	Patentes genes humanos	D1-C11_bbc	PT	13/07/2013- 16:34 (GMT)	963
			D2_C11_folha	IN	13/06/2013-23:50	
C12	Poder	Protestos: transporte	D1_C12_folha	PT	14/06/2013-07:25	808
			D2_C12_bbc	IN	14/06/2013-12:43 (GMT)	
C13	Mundo	Eleições do Irã	D1_C13_folha	PT	15/06/2013 – 17:57	1.266
			D2_C13_bbc	IN	16/06/2013 - 08:38 (GMT)	
C14	Saúde	Epidemia de dengue no MS	D1_C14_folha	PT	11/01/2013 1-9:03	534
			D2_C14_bbc	IN	21/01/2013- 00:21 (GMT)	
C15	Saúde	Mastectomia preventiva	D1_C15_folha	PT	15/05/2013 – 03:01	1.367
			D1_C15_bbc	IN	14/05/2013 -17:02 (GMT)	
C16	Ciência	Missão espacial chinesa	D1_C16_folha	PT	11/06/2013 – 21:06	793
			D2_C16_bbc	IN	11/06/2013-9:38 (GMT)	
C17	Poder	Protesto: copa das confederações	D1_C17_folha	PT	15/06/2013 – 14:53	918
			D2_C17_bbc	IN	16/06/2013 -13:19 (GMT)	
C18	Ciência	Viagra feminino	D1_C18_folha	PT	16/06/2013 – 03:30	975
			D2_C18_bbc	IN	17/11/2009- 9:35 (GMT)	
C19	Entretenimento	Lançamento: homem de aço	D1_C19_folha	PT	16/06/2013-13:24	898
			D2_C19_bbc	IN	11/06/2013-10:17(GMT)	
C20	Mundo	Conflito na Turquia	D1_C20_folha	PT	17/06/2013 - 09h44	963
			D2_C20_bbc	IN	17/06/2013-13:00(GMT)	
Total de palavras						**19.984**

Fonte: Elaborada pelo autor

[29] Disponível em: <http://www.nilc.icmc.usp.br/arianidf/sustento/>.

Salienta-se que os critérios de compilação do CM2News, como: “forma de coleta”, “fonte”, “atualidade dos textos” e “variedade” foram os mesmos considerados para o CM3News.

Com base no Quadro 3, observa-se que a composição final do *corpus* CM2News engloba 6 diferentes domínios de assuntos ou temas variados (mundo, poder, saúde, ambiente, ciência e entretenimento), sendo que as notícias foram publicadas no período de março de 2011 a julho de 2013.

Por fim, salienta-se que para a seleção dos textos, 2 critérios foram aplicados, em especial: (i) tamanho e (ii) originalidade. Quanto ao tamanho dos textos, buscou-se compilar textos que tivessem tamanho (medido em número de palavras) parecido. Na coleção19 (C19), por exemplo, o texto em português possui 452 palavras e o texto em inglês, 446, totalizando 898. No geral, o tamanho médio das coleções do CM2News é de 999,2 palavras. Quanto à originalidade dos textos, preocupou-se em selecionar textos que versassem sobre um mesmo assunto ou tema, mas que não fossem traduções um do outro.

3.2 A seleção das unidades lexicais

Para a anotação dos conceitos subjacentes às unidades lexicais dos textos-fonte do CM2News, necessitou-se delimitar quais unidades seriam anotadas.

Para tanto, investigaram-se dois critérios de seleção: (i) frequência das palavras de categoria aberta e (ii) categoria gramatical. Ao adotar-se o critério da frequência das palavras de categoria aberta, apenas as unidades mais frequentes na coleção seriam indexadas. Por consequência, unidades que expressam diferentes tipos de conceitos estariam envolvidas no processo, posto que, entre as mais frequentes, estão os nomes e os verbos, os quais lexicalizam, na maioria das vezes, entidades e ações, respectivamente. Essa variedade de tipos conceituais tornaria as tarefas de identificação e anotação dos conceitos mais complexas.

Para que o processo de anotação semântica fosse uma tarefa mais delimitada e controlada, optou-se pelo critério da categoria gramatical, por meio do qual apenas as unidades da categoria dos nomes foram selecionadas na coleção e efetivamente anotadas. Essa restrição se deve ao fato de que os nomes são, entre as unidades de classe

aberta, as mais frequentes, expressando, juntamente com os verbos, o conteúdo semântico principal dos textos.

3.3 A seleção da ontologia

Após a delimitação das unidades lexicais da categoria dos nomes comuns, os conceitos a elas subjacentes precisavam ser explicitados por meio da tarefa de anotação do *corpus* CM2News em nível léxico-conceitual. Para tanto, selecionou-se um conjunto de conceitos (ou ontologia[30]) que foram utilizados como rótulos ou etiquetas para anotar os nomes do CM2News.

Para o português, reconhece-se a existência de uma ontologia robusta no Dicionário Analógico da Língua Portuguesa (DOS-SANTOS-ZEVEDO, 1974), no entanto, tal ontologia não se encontra disponível em formato digital.

Assim, tendo em vista a não existência de uma ontologia suficientemente robusta de língua geral em português (isto é, que engloba conceitos de domínios variados), que seja computacionalmente tratável, optou-se por utilizar a WN.Pr (FELLBAUM, 1998), construída para o inglês americano.

Do ponto de vista teórico, a WN.Pr foi escolhida por sua adequação linguística, uma vez que busca simular o léxico mental, e abrangência, uma vez que é uma das mais extensas do inglês. Do ponto de vista prático, essa ontologia foi selecionada por ser uma das mais utilizadas no PLN.

A WN.Pr é uma rede em que as palavras e expressões, pertencentes às categorias dos nomes, verbos, adjetivos e advérbios, organizam-se sob a forma de *synsets* (do inglês, *synonym sets*). Em outras palavras, pode-se dizer que o *synset* é um conjunto de formas (do inglês, *word forms*) de uma mesma categoria gramatical que podem ser intercambiáveis em determinado contexto, p.ex.: {bicycle, bike, wheel, cycle}.

O *synset*, por definição, é construído de modo a codificar um único conceito lexicalizado por suas formas constituintes. Vale ressaltar que os *synsets* da WN.Pr também podem armazenar conceitos não-lexicalizados no inglês, ou seja, conceitos para os quais não há uma expressão lexical (isto é, expressão que se espera encontrar como entrada em um dicionário monolíngue). Incluem-se nesse grupo, por exemplo, os

[30] Por "ontologia", entende-se um inventário de conceitos, propriedades e relações entre conceitos que representam "uma interpretação da realidade", ou seja, o conhecimento de mundo compartilhado pelos membros de uma comunidade linguística (GRUBER, 1995).

conceitos codificados pelos *synsets* {natural object} e {external body parts}. A principal razão da inclusão desses conceitos é auxiliar a organização da hierarquia conceitual (VOSSEN, 1998).

Assim, se o falante não conhece o significado de uma determinada forma lexical, uma forma sinônima é suficiente para que ele identifique o conceito apropriado. Por exemplo, se o falante desconhece a forma *x* e essa forma é parte do *synset s* e o falante conhece as formas *y* e *z* desse *synset*, então, porque a forma desconhecida *x* é parte de *s*, o falante passa a ter acesso ao significado da forma *x*.

O emprego do *synset* como construto representacional pressupõe que os conceitos são ativados na mente por meio de formas lexicais sinônimas, eliminando-se a necessidade de determinar o valor semântico das unidades. A WN.Pr adotou a noção de **sinonímia contextual** para a montagem de *synsets*. De acordo com essa noção de sinonímia, "duas unidades lexicais são sinônimas em um contexto C, se a substituição de uma pela outra em C não altera o valor de verdade denotado por C" (MILLER; FELLBAUM, 1991). A sinonímia contextual contrapõe-se à **sinonímia absoluta**, segundo a qual "duas unidades lexicais são totalmente sinônimas quando são substituíveis, uma pela outra, em todos os contextos, sem que haja mudança do valor de verdade da proposição expressa pelas sentenças em que as substituições são feitas". A sinonímia absoluta é raramente encontrada na língua geral.

Entre os *synsets*, codificam-se 5 principais relações lógico-conceituais: antonímia, hiponímia, meronímia, acarretamento e causa (LYONS, 1979; CRUSE, 1986; FELLBAUM, 1998):

a) Hiperonímia/ Hiponímia: relação entre um conceito mais genérico (o hiperônimo) e um conceito mais específico (o hipônimo). Um item lexical é hipônimo de outro item lexical se o falante aceita sentenças construídas a partir da seguinte fórmula: um x é um (tipo de) y. Por exemplo, a aceitação das sentenças *a limusine é um tipo de carro* e *um carro é um tipo de veículo* identifica o possível *synset* {limusine} como hipônimo do *synset* {carro} e {carro} como hipônimo de {veículo}.[31]
b) Antonímia: relação que engloba diferentes tipos de oposição semântica. São elas: *antonímia complementar*: relaciona pares de itens lexicais contraditórios em que a afirmação do primeiro acarreta a negação do segundo e vice-versa, por exemplo: {vivo} e {morto}; *antonímia gradual*, que relaciona itens lexicais que denotam

[31] Os exemplos de *synsets* elucidados tratam-se de ilustrações do autor, não sendo exemplos fiéis retirados da WN.Pr.

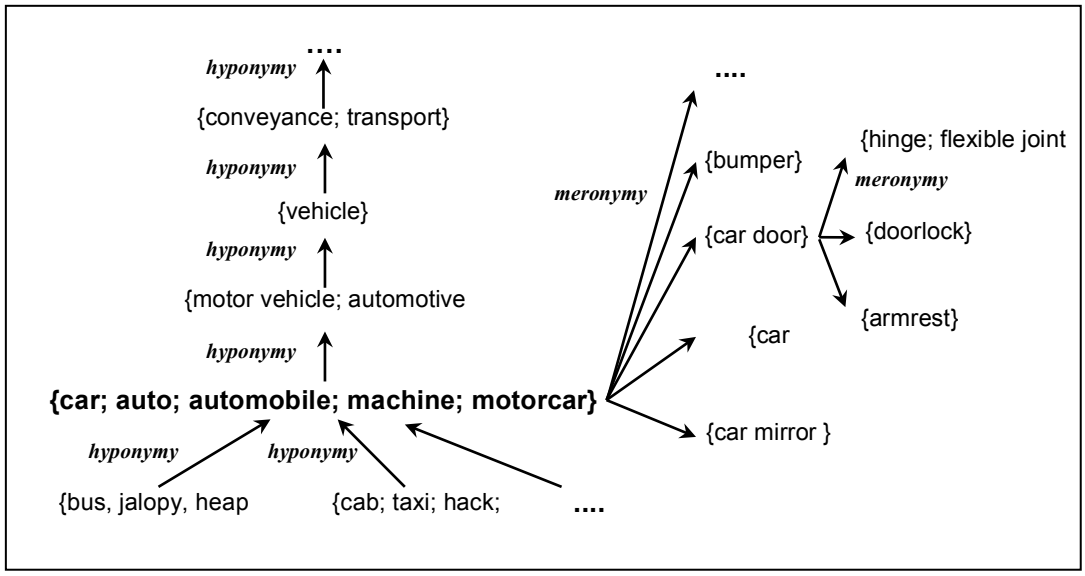
valores opostos em uma escala como, por exemplo, {pequeno} e {grande}; e "antonímia recíproca", que relaciona pares de itens lexicais que se pressupõem mutuamente, sendo que a ocorrência do primeiro pressupõe a ocorrência do segundo como, por exemplo, {comprar} e {vender}.

c) Meronímia/ Holonímia: relação entre um *synset* que expressa um "todo", o holônimo, por exemplo, o *synset* hipotético {carro}, e outros *synsets* que expressam partes do todo, os merônimos, por exemplo: {pára-choque}, {pneu}, {direção}, {câmbio}, etc.

d) Acarretamento: relação que se estabelece entre uma ação A1 e uma ação A2; a ação A1 denotada pelo verbo x acarreta a ação A2 denotada pelo verbo y se A1 não puder ser feita sem que A2 também o seja. Esse é o caso, por exemplo, da relação entre os verbos *correr* e *deslocar-se*, já que a ação de correr (A1) acarreta a ação de deslocar-se (A2); assim, estabalece-se a relação de acarretamento entre os possíveis *synsets* {correr} e {deslocar-se}. Vale salientar que o acarretamento é uma relação unilateral, isto é, por um lado *correr* acarreta *deslocar-se,* mas, por outro, o inverso não ocorre, *deslocar-se* não necessariamente acarreta *correr*.

e) Causa: relação que se estabelece entre uma ação A1 e uma ação A2 quando a ação A1 denotada pelo verbo x causa a ação A2 denotada pelo verbo y. Esse é o caso, por exemplo, da relação que se estabelece entre a ação denotada por *matar* e a ação denotada pelo verbo *morrer*.

Figura 14 – Organização dos *synsets* constituídos por nomes.



Fonte: Di-Felippo (2008).

Na Figura 14, cujo exemplo foi extraído da WN.Pr (version 2.1), exemplificam-se dois tipos de relação: a hiperonímia e a meronímia. Vê-se nessa Figura 14 que o *synset* {car; auto; automobile; machine; motorcar} está relacionado, por exemplo, a:

a) o conceito mais geral ou *synset* hiperônimo: {motor vehicle; automotive vehicle};
b) os conceitos mais específicos ou *synsets* hipônimos, p.ex.: {bus; jalopy; heap } e {cab; taxi; hack; taxicab};
c) os conceitos que indicam partes ou *synsets* merônimos, p.ex.: {bumper}, {car door}, {car mirror} e {car window}.

Observa-se ainda que cada *synset* relaciona-se novamente a outros *synsets*, por exemplo, o *synset* {motor vehicle; automotive vehicle} está relacionado à {vehicle} e {conveyance; transport}.

A WN.Pr armazena ainda uma série de informações associadas a cada *synset*:

a) um número que identifica o *synset*; por exemplo, para {bicycle; bike; wheel; cycle}, tem-se o número 02834778;
b) o tipo semântico do conceito representado no *synset*; p.ex.: o *synset* {bicycle; bike; wheel; cycle} é do tipo semântico <noun.artifact>;
c) uma glosa, isto é, uma definição informal do conceito representado no *synset*; p.ex.: "*a wheeled vehicle that has two wheels and is moved by foot pedals*" ("um veiculo rodado que tem duas rodas e é movido por pedais");
d) frases-exemplo extraídos de *corpora;*
e) um conjunto de indexadores (do inglês, *pointers*), que estabelecem as relações semântico-conceituais entre os *synsets*.

Diante da escolha da WN.Pr, os *synsets* foram então utilizados como rótulos para explicitar os conceitos subjacentes aos nomes comuns do *corpus* CM2News.

3.4 A anotação semântica

O processo de anotação dos nomes comuns com os *synsets* da WR.Pr foi realizado de forma semiautomática durante o período de 10 dias consecutivos, com encontros diários de 90 a 120 minutos. O primeiro dia de anotação foi reservado para a apresentação da tarefa, da ferramenta e do manual de anotação (itens 3.5 e 3.6). Nesse período de treino, os anotadores realizaram uma "anotação-teste", na qual duplas de

anotadores anotaram um texto-fonte escolhido aleatoriamente do *corpus*. As dúvidas geradas foram sanadas e, no segundo dia, partiu-se para a anotação definitiva dos textos.

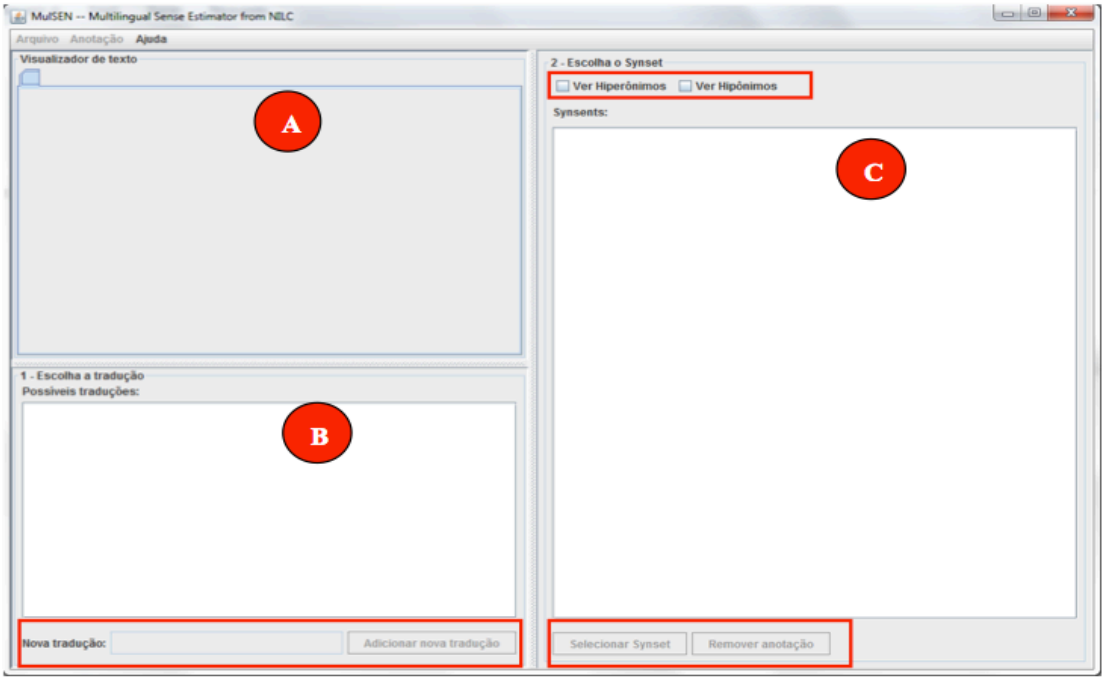
A anotação semântica do *corpus* CM2News foi feita por uma equipe de 12 linguistas computacionais. Em cada um dos 9 dias de efetiva anotação, cada dupla de anotadores ficou responsável por uma coleção distinta. A distribuição das coleções entre os anotadores ocorreu de forma aleatória, bem como a formação das duplas. A cada dia, as duplas eram modificadas.

Na sequência, apresentam-se a ferramenta de anotação semiautomática e o manual de anotação, composto por regras gerais e específicas.

3.5 O editor MulSen e o processo de anotação

Para a anotação do *corpus* CM2News em nível léxico-conceitual, desenvolveu-se um editor específico, isto é, uma ferramenta com interface gráfica de auxílio à tarefa de explicitação dos conceitos expressos pelos nomes por meio dos *synsets* da WN.Pr. Esse editor foi denominado MulSen (do inglês, *Multilingual Sense Estimator*). Na Figura 15, apresenta-se a interface do editor MulSen, composta por 3 janelas:

Figura 15 – Interface do editor MulSen.



A janela (a), denominada “Visualizador do texto”, é responsável por exibir os textos-fonte da coleção ao anotador. No caso da abertura de 2 textos-fonte, essa janela compõe-se por duas abas de exibição, uma para cada texto. Especificamente, as abas, que são intituladas “Texto 1” e “Texto 2” (cf. Figura 16, pág. 67), exibem os textos-fonte em inglês e português, respectivamente, de uma mesma coleção. A exibição dos textos não é simultânea. No entanto, o acesso a ambos pode ser feito durante todo o processo de anotação, podendo o anotador passar de um texto para o outro sempre que desejar.

Na janela (b), apresentam-se as possíveis traduções da palavra em português que se quer anotar. Para a anotação de uma palavra *w* de um texto em inglês, o MulSen exibe, na janela (b), a própria palavra *w*.

Caso o MulSen não sugira traduções adequadas para as palavras em português, o editor fornece a possibilidade de inserção pelos anotadores de um equivalente em inglês por meio do campo “Nova tradução”, localizado na parte inferior da janela (b). Somente por meio da escolha de um equivalente em inglês, o MulSen é capaz de exibir, para os anotadores, os *synsets* da WN.Pr que possuem tal equivalente e que podem representar o conceito da palavra a ser anotada.

As possíveis traduções são provenientes do acesso ao dicionário *online* WordReference[32].

A lista de *synsets* que possuem a unidade equivalente em inglês é exibida na janela (c). Na parte superior dessa janela, o editor fornece a opção de exibição dos hiperônimos (“Ver Hiperônimos”) e hipônimos (“Ver Hipônimos”) dos *synsets* listados.

Na parte inferior, estão os botões que permitem a seleção do *synset* adequado pelo anotador e a efetiva anotação, ou seja, associação do *synset* escolhido ao nome do texto-fonte.

Ressalta-se que, ao abrir os textos de uma coleção, o editor MulSen realiza 2 tarefas automáticas de pré-processamento: (i) etiquetação morfossintática (do inglês, *part-of-speech tagging*), para a identificação dos nomes, e (ii) desambiguação lexical de sentido (dos nomes) (do inglês, *word sense disambiguation*), para auxiliar na identificação do conceito.

A etiquetação morfossintática consiste na atribuição da categoria gramatical a cada uma das palavras de um texto. A categoria é explicitada por meio de rótulos ou

[32]Disponível em: <http://www.wordreference.com/>.

etiquetas.

Para os textos em português, o editor MulSen aplica o etiquetador ou *tagger* MXPOST (RATNAPARKHI, 1986) e, para os textos-fonte em inglês, o editor utiliza outra ferramenta, o *TreeTagger* (SHIMID,1994).

Para o próximo processo, que é a desambiguação lexical de sentido, apenas as palavras rotuladas com a etiqueta "nome" (comum) são consideradas.

A desambiguação lexical de sentidos (DLS) corresponde à tarefa de determinar o sentido/conceito mais adequado de uma palavra, dados o contexto em que esta foi empregada (sentença, texto, documento, etc.) e um conjunto finito de possíveis sentidos/conceitos (em dicionários, ontologias, etc.) (AGIRRE, EDMONDS, 2006).

No caso, esse processo consiste em identificar, para cada um dos nomes dos textos em português e em inglês, o *synset* da WN.Pr que mais adequadamente codifica o conceito expresso por ele nos textos-fonte. Para a identificação do conceito/*synset* mais adequado, o MulSen utiliza um dos métodos de DLS adaptados por Nóbrega (2013) para o português.

O DLS empregado por Nóbrega e Pardo (2013) atribui somente um sentido para cada palavra em uma coleção de documentos. Assim, todas as ocorrências de uma mesma palavra recebem o mesmo sentido. Essa abordagem resulta da verificação em *corpus* de que as palavras tendem a ocorrer com o mesmo sentido em textos que abordam um mesmo assunto.

Baseado no algoritmo de Lesk (1986), o algoritmo empregado por Nóbrega e Pardo (2013) no editor Mulsen, é dividido em duas partes principais.

Na primeira parte, identificam-se, para cada palavra (unidade lexical) a ser desambiguada, as duas palavras mais relacionadas a ela. As palavras "mais relacionadas" são aquelas que possuem arestas de maior peso vinculadas à palavra (unidade lexical) a ser desambiguada. As duas palavras mais fortemente relacionadas à que está em processo de desambiguação são escolhidas para representar o seu contexto.

A partir daí, na segunda parte, quando o processo de desambiguação efetivamente acontece, compara-se essas duas palavras mais relacionadas, com todas as palavras presentes nas glosas de todos os *synsets* referentes à unidade lexical a ser anotada.

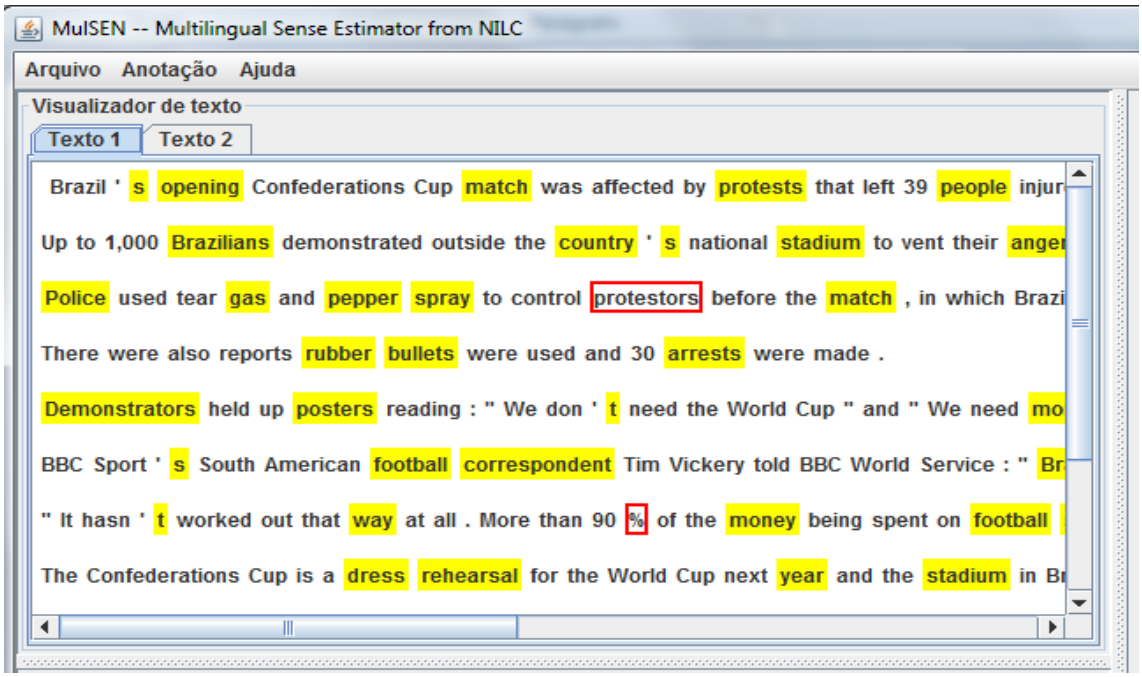
A comparação é feita a partir da sobreposição das palavras presentes nas glosas dos *synsets* com as palavras mais relacionadas com a unidade lexical (as quais foram

identicadas por suas relações de vizinhança, na primeira fase do processo).[33] O *synset* cuja glosa estabelecer maior numero de relações com as palavras em questão é selecionado, concluindo assim, o processo de desambiguação. Caso o processo não dê certo, o *synset* mais frequente é utilizado/eleito.

Somente após as tarefas de etiquetação e DLS, os textos-fonte são exibidos aos anotadores.

Na Figura 16, a aba "Texto 1" exibe, por exemplo, o texto-fonte em inglês da C17 do CM2News após os processos de etiquetação e DLS.

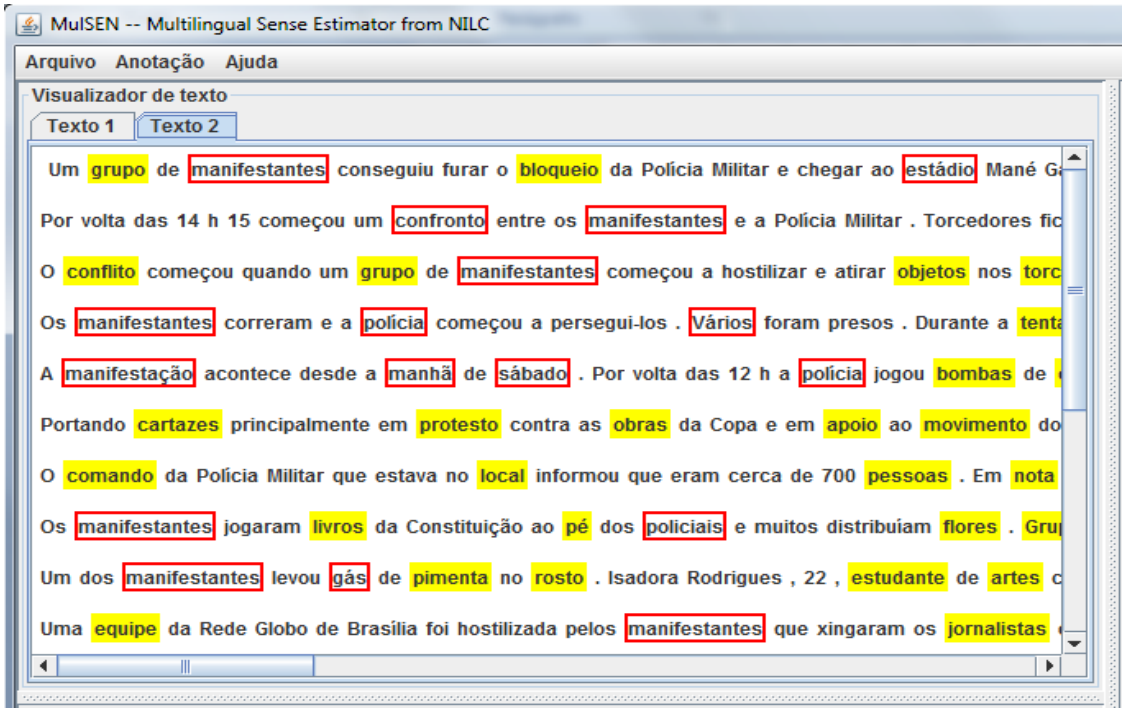
Figura 16 – Exibição do texto-fonte em inglês após a etiquetação e DLS.



[33] Para os textos em português, é necessário realizar a TA das palavras relacionadas na primeira fase do processo de desambiguação. Isso é necessário, pois na a segunda fase, (na qual ocorre a desambiguação) estabelece-se a comparação das unidades lexicais selecionadas com às unidades lexicais das glosas dos *synsets* da WN.Pr, que são em inglês.

Na Figura 17, a aba “Texto 2” exibe o texto-fonte em português da C17 do CM2News após os processos de etiquetação e DLS.

Figura 17– Exibição do texto-fonte em português após a etiquetação e DLS.



Nas abas “Texto 1” e “Texto 2”, observa-se que os nomes que ocorrem nos textos estão em destaque nas cores amarela ou vermelha. As cores indicam diferentes estatutos de pré-anotação automática das palavras.

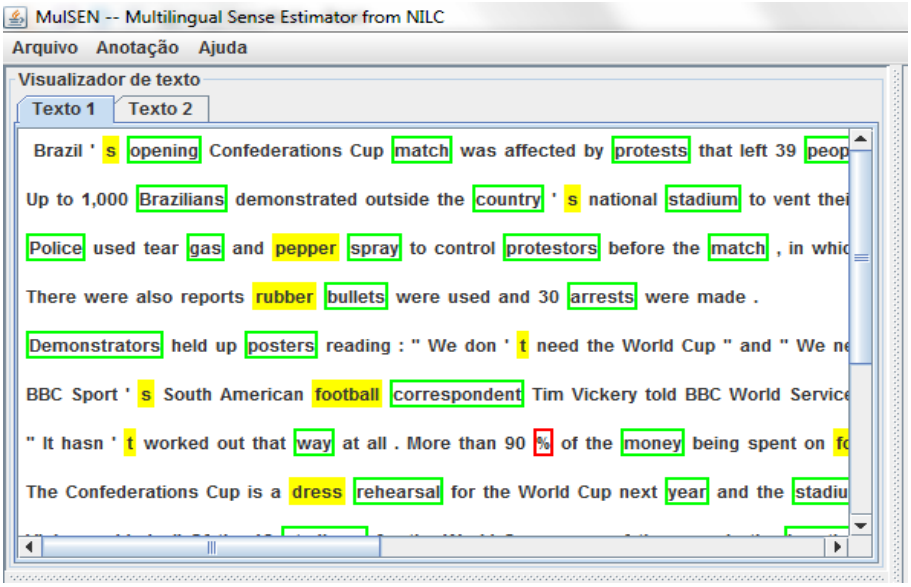
Especificamente, as unidades lexicais em amarelo foram etiquetadas como nome e desambiguadas, ou seja, foram reconhecidas como nome pelo etiquetador morfossintático e posteriormente desambiguadas (anotadas) pelo método de DLS, recebendo, assim, um *synset* a ser confirmado (ou não) pelo anotador humano.

As palavras em vermelho, por sua vez, foram apenas etiquetadas em nível morfossintático. Isso significa que, para essas palavras, o editor MulSen não conseguiu, por meio de seu método de DLS, sugerir *synsets* que potencialmente representam os seus conceitos ou não conseguiu encontrar de forma automática um equivalente de tradução para o inglês da palavra em questão. Há casos, ainda, em que o dicionário encontra um equivalente para o inglês, mas esse equivalente não está representado na WN.Pr. Nesses casos, o anotador tem a possibilidade de sugerir um novo equivalente de tradução (sinônimo ou hiperônimo) da unidade lexical em questão.

Após a seleção do *synset* adequado, seja o indicado pela ferramenta seja o escolhido pelos anotadores, e a efetiva confirmação da anotação, as palavras do texto passam a figurar em cor verde, indicando, assim, que o processo de anotação semântica da palavra foi finalizado.

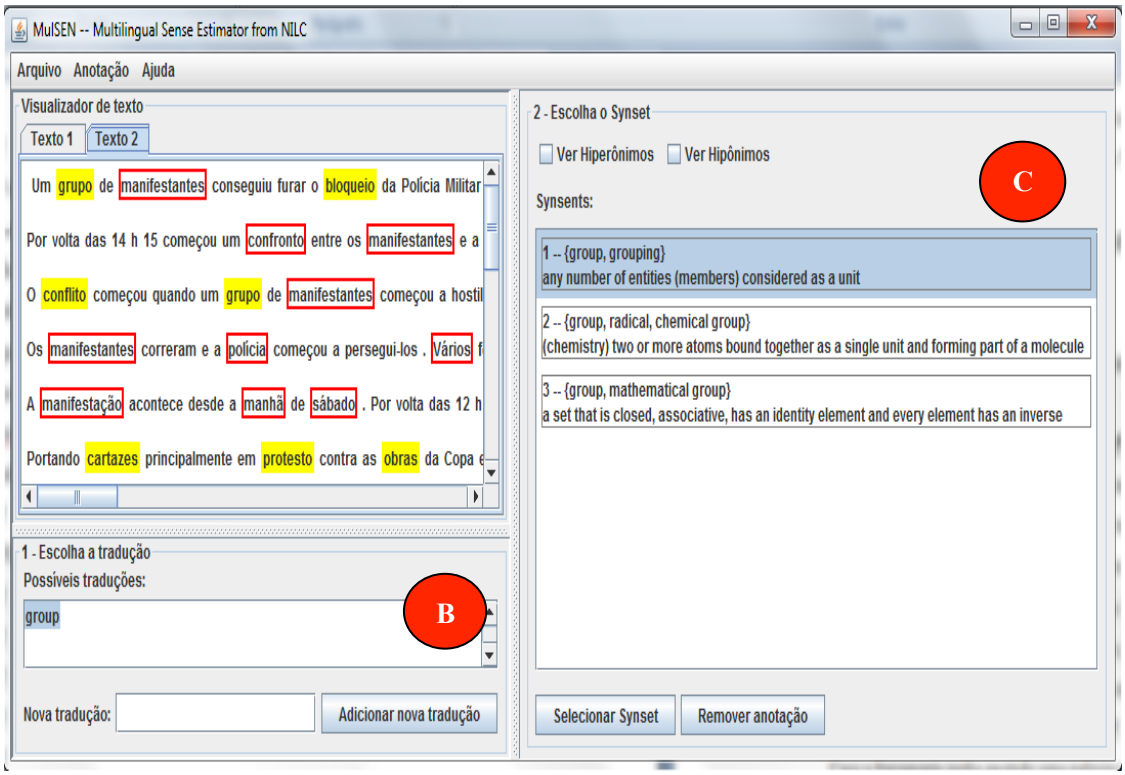
Na Figura 18, ilustra-se o texto em inglês da C17 após a finalização do processo de anotação; as palavras que ainda permanecem em amarelo foram desconsideradas no processo de anotação semântica.

Figura 18 – Exibição do texto-fonte em inglês após anotação semântica.



Para ilustrar a funcionalidade das janelas (b) e (c) da interface do editor MulSen, considera-se o primeiro nome do texto em português (Figura 19): “grupo”.

Figura 19 – Ilustração das janelas (b) e (c) da interface do MulSen.

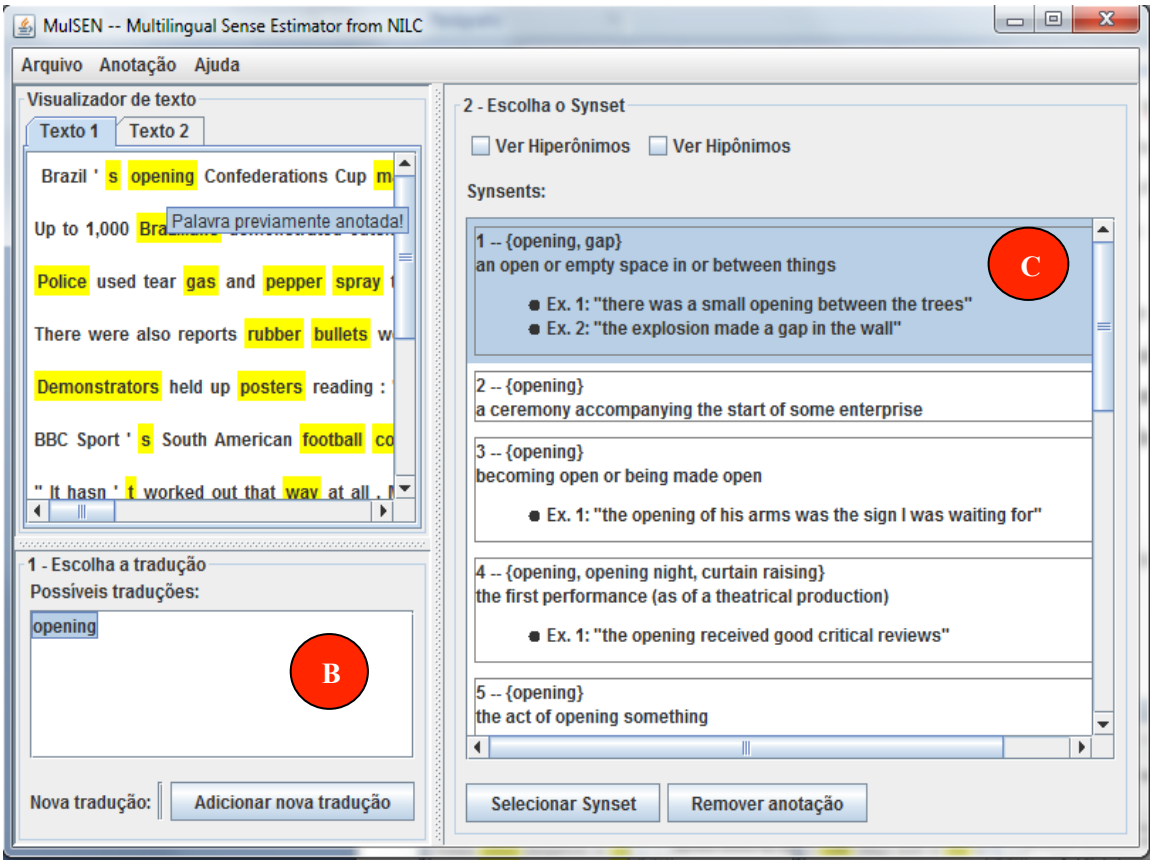


A partir da consulta ao tradutor *online* WordReference, o editor: (i) sugere, na janela (b), a palavra *group* como equivalente de tradução; (ii) exibe, na janela (c), a lista de *synsets* que possuem *group* como um de seus elementos constitutivos, e (iii) sugere, com base no método de DLS, o primeiro *synset* da lista, {group, grouping} (“*any number of entities* (*members*) *considered as a unit”*) (“qualquer número de membros, considerados como uma unidade”), como possível codificação do conceito subjacente a “grupo”, exibindo-o em cor azul (parte destacada na janela (c)).

Para a anotação semântica de uma palavra *y* de um texto-fonte em inglês, o MulSen exibe a própria palavra *y* na janela (b) de “possíveis traduções”, a partir da qual o método de DLS também é aplicado e o *synset* sugerido.

Na Figura 20, vê-se que, para a palavra *opening* do texto-fonte em inglês, o editor sugeriu *opening* na janela (b).

Figura 20 – Sugestão da unidade do texto-fonte em inglês como "possível tradução".



Caso a sugestão de *synset* seja validada, o processo de anotação semântica é finalizado por meio do botão "Selecionar *synset*", localizado na parte inferior da janela (c) do editor. Caso os anotadores optem por outro *synset,* que não o pré-selecionado pela ferramenta, deve-se clicar no *synset* adequado e, na sequência, no botão "Selecionar *synset*".

Ao final, o MulSen gera um arquivo no formato XML (do inglês, *Extensible Markup Language*), um dos mais utilizados para a tarefa de anotação de *corpus*. Na Figura 21, ilustra-se o formato gerado pelo MulSen.

Na Figura 21, vê-se que, dentre os *synsets* da WN.Pr que possuem *opening* como um de seus elementos constitutivos, os anotadores cujas identificações foram omitidas na Figura selecionaram o *synset* codificado pelo ID 7452699; a seleção é codificada pelo valor "true".

Figura 21 – Formato XML da anotação semântica gerada pelo MulSen.

```
<?xml version="1.0" encoding="UTF-8"?>
<save>
        <Annotators>
                <Annotator id="1">X</Annotator>
                <Annotator id="2">Y</Annotator>
        </Annotators>
        <Files>
                <Text language="ENGLISH" name="en_C17_txt_uft8.tagged">
                        <Token>
                                <Word>opening</Word>
                                <Tag>NN</Tag>
                                <MorphoTag>Substantivos</MorphoTag>
                                <Lemma>opening</Lemma>
                                <Type>ANNOTED</Type>
                                <Translations manual_translation="false">
                                        <Translate selected="true">opening</Translate>
                                </Translations>
                                <Synsets>
                                        <Synset selected="true">7452699</Synset>
                                        <Synset selected="false">9379111</Synset>
                                        <Synset selected="false">383390</Synset>
                                        <Synset selected="false">7329363</Synset>
                                        <Synset selected="false">338641</Synset>
                                        <Synset selected="false">14485249</Synset>
                                        <Synset selected="false">6397476</Synset>
                                        <Synset selected="false">5792010</Synset>
                                        <Synset selected="false">5249636</Synset>
                                        <Synset selected="false">3848729</Synset>
                                        <Synset selected="false">3499142</Synset>
                                        <Synset selected="false">457890</Synset>
                                        <Synset selected="false">239230</Synset>
                                </Synsets>
                        </Token>
```

Observa-se na Figura 21 há existência de blocos de informação que delimitam informações relativas aos: (i) anotadores, (ii) arquivo em questão e (iii) palavra anotada. No bloco referente aos anotadores (*Annotators*), são evidenciados os nomes dos anotadores responsáveis pela coleção em questão e, no bloco referente ao arquivo (*files*), evidenciam-se o documento o formato do arquivo (p. ex.: en_C17_txt_uft8). No bloco referente à palavra anotada (*token*), apresentam-se as seguintes informações: (i) *Word* (unidade lexical anotada); (ii) *Tag* e *MorfoTag* (etiqueta morfossintática associada à unidade lexical); (iii) *Lemma* (forma canônica da unidade lexical); (iv) *Translations*

(equivalente de tradução e a forma pela qual foi especificado, ou seja, automática ou manual); (v) *Synsets* (lista de *synsets* constituídos pelo equivalente de tradução e o *synset* selecionado dentre eles).

3.6 Os procedimentos de anotação semântica

O processo de anotação semântica dos nomes do *corpus* CM2News por meio do editor MulSen, foi feito com base em um manual composto por regras gerais e específicas. A seguir, descreve-se cada regra.

3.6.1 Regras gerais

REGRA G1: *Ler cuidadosamente os textos-fonte de cada coleção*
Essa regra estabeleceu que o primeiro procedimento fosse a leitura integral dos textos-fonte da coleção sob análise antes do processo de anotação. Para tanto, a dupla de anotadores podia utilizar o próprio editor MulSen, outro editor de texto ou mesmo os textos impressos. A leitura prévia dos textos-fonte de uma coleção visava à familiarização dos anotadores com o seu conteúdo, pois, para a identificação dos conceitos/*synsets* subjacentes aos nomes, a compreensão do conteúdo global do texto no qual ocorrem é essencial.

REGRA G2: *Iniciar a anotação preferencialmente pelo texto-fonte em inglês da coleção*
Essa regra estabeleceu que, dada uma coleção, o texto-fonte em inglês fosse o primeiro a ser anotado. Ao iniciarem a anotação pelo texto em inglês, os anotadores puderam utilizar as próprias unidades lexicais originais encontradas no texto em inglês como equivalentes de tradução das unidades em português e subsequente identificação do *synset* adequado. Dessa forma, a anotação prévia do texto em inglês facilitou a busca por equivalentes de tradução das unidades lexicais em português.

Caso a anotação começasse pelo texto em português, os anotadores podiam selecionar uma tradução que julgassem adequada para expressar o conceito subjacente à palavra em português, mas que, ao anotar o texto em inglês, percebesse que o conceito em questão é comumente expresso por outra expressão linguística em inglês que não a

inicialmente selecionada. Assim, o texto em inglês das coleções foi usado como fonte de referência para a seleção das traduções das palavras em português.

REGRA G3: *Anotar todos os nomes comuns e siglas do corpus*

Essa regra determinou que todos os nomes comuns do *corpus* CM2News fossem alvo da anotação semântica, assim como as siglas, e que nomes próprios não fossem anotados em nível semântico.

Os nomes comuns foram escolhidos para alvo da anotação semântica porque se entende que eles carregam a maior carga semântica de um texto. A inclusão das siglas na anotação justifica-se pelo fato de que estas são relevantes para a transmissão do conteúdo no gênero jornalístico, que é o do *corpus* CM2News.

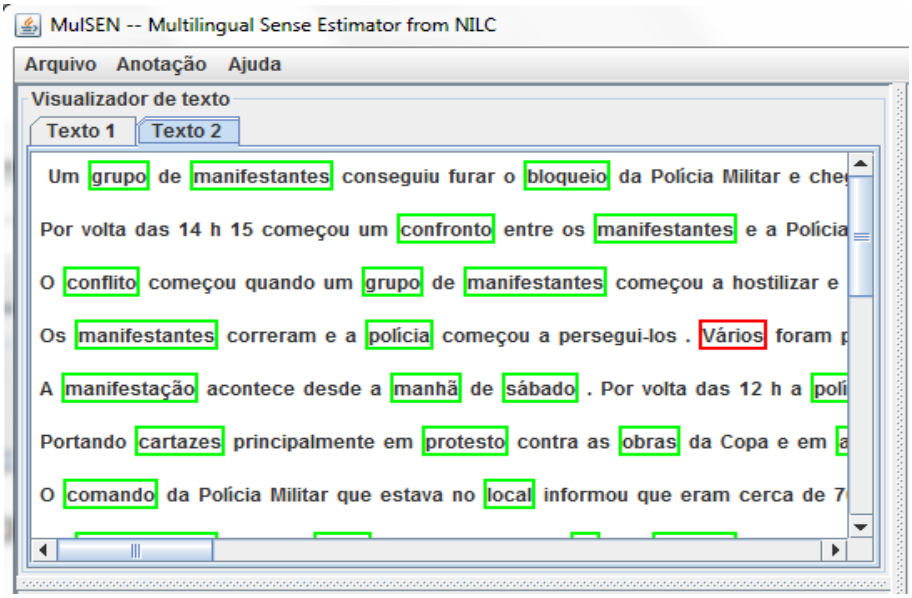
REGRA G4: *Refinar a anotação morfossintática automática*

Essa regra estabeleceu que a anotação morfossintática realizada em etapa anterior pelos etiquetadores automáticos (*taggers*) fosse revisada pelos anotadores, pois os etiquetadores automáticos não são ferramentas totalmente precisas. Especificamente, essa regra estabeleceu que os casos de silêncio (ou seja, nomes não etiquetados automaticamente pelos *taggers*) fossem identificados pelos anotadores para que o MulSen realizasse a etiquetação. Ressalta-se, sobretudo, que os *taggers* em questão não reconhecem expressões multipalavras.

REGRA G5: *Ignorar as palavras anotadas equivocadamente como nome*

Essa regra estabeleceu que os casos de ruídos (ou seja, palavras anotadas equivocadamente como nome) gerados pelos *taggers* fossem ignorados, não sendo alvo da anotação semântica subsequente. Por exemplo, na Figura 22, observa-se que o MXPost identificou erroneamente o pronome “vários” como nome, destacando-o por meio da cor vermelha. Seguindo-se a Regra 4, casos como esse foram ignorados.

Figura 22 – Ilustração dos casos de ruído gerados pelos *taggers*.



REGRA G6: *Selecionar o mesmo synset para anotar diferentes expressões linguísticas do mesmo conceito na coleção*

Com base nessa regra, os anotadores deveriam garantir a seleção do mesmo *synset* para anotar: (i) todas as ocorrências de uma palavra *x*, com o sentido *y* no mesmo texto; (ii) as ocorrências de palavras sinônimas de *x* no mesmo texto, e (iii) as ocorrências dos equivalentes de *x* no outro texto da coleção. Essa regra implica a consulta frequente a ambos os textos por meio das abas "Texto 1" e "Texto 2" durante a anotação de uma mesma coleção.

Seguindo-se essa regra, todas as ocorrências da palavra "protesto" com o sentido de "o ato de fazer uma grande manifestação pública de discordância e reprovação" no texto em português da C17 foram anotadas com o *synset* {protest} ("*the act of making a strong public expression of disagreement and disapproval*"). Ademais, "manifestação", interpretada como sinônimo de "protesto", foi anotada com o mesmo *synset,* {protest}. No texto em inglês, a palavra *protests* é o equivalente de "protesto" e "manifestação", sendo, portanto, anotada com o mesmo *synset,* {protest} ("*the act of making a strong public expression of disagreement and disapproval*").

A seguir, apresentam-se as regras específicas para a anotação de cada um dos nomes do CM2News.

3.6.2 Regras específicas

REGRA E1: *Anotar somente o nome nuclear das expressões multipalavras*
Essa regra foi estabelecida porque os etiquetadores morfossintáticos anotam somente unidades simples (ou seja, sequências de caracteres separadas por espaços em branco), não reconhecendo expressões multipalavras.

Assim, essa regra estabeleceu que os nomes etiquetados isoladamente em nível morfossintático, mas que fossem, na verdade, núcleos de expressões multipalavras, fossem anotados com *synsets* que representassem os conceitos expressos pelas expressões multipalavras, desde que houvesse tais *synsets*.

Por exemplo, com base nessa regra, o nome "gás" em (3a), que é núcleo da expressão "gás de pimenta", cuja estrutura interna é [S**N**[N]+ $_{\textbf{SPrep}}$ [Prep+N]]$_{\textbf{SN}}$[34], foi anotado com o *synset* {pepper spray} ("*a nonlethal aerosol spray made with the pepper derivative oleoresin capiscum; used to cause temporary blindness and incapacitate an attacker*") ("um spray aerosol não-letal feito com a oleorresina derivada das plantas (do gênero) *Capsicum*"), posto que este representa o conceito subjacente à expressão.

O mesmo *synset,* aliás, foi utilizado para anotar a palavra *spray* do texto em inglês da C17. Anotado isoladamente como nome, *spray* é núcleo da expressão multipalavra *pepper spray* (cuja estrutura interna é [Adj[35]+N]) e, por isso, foi anotado com o *synset* {pepper spray}, que codifica na WN.Pr o conceito subjacente a *pepper spray.*

O mesmo foi feito com "cordão" em (3b) que, apesar de ter sido etiquetado isoladamente como nome, é núcleo de "cordão de isolamento", e, por isso, foi anotado com o *synset* 1 {cordon} ("*a series of sentinels or military posts enclosing or guarding some place or thing*") ("uma série de sentinelas ou militares que cercam ou guardam algum lugar ou coisa"), que codifica o conceito subjacente à expressão multipalavra em questão.

(3) a. [...] Um dos manifestantes levou **gás** de pimenta no rosto. [..] (C17)
b. [...] A polícia montou um **cordão** de isolamento ao redor do estádio [..]. (C17)

[34] N = nome; Prep = preposição; SPrep = sintagma preposicional; SN = sintagma nominal.
[35] Adj = adjetivo.

REGRA E2: *Anotar todos os nomes de sintagmas recorrentes livres*

Essa regra determinou que todos os nomes constitutivos de sintagmas livres recorrentes (SLRs) (do inglês, *recurrent free phrases*) fossem anotados com seus respectivos *synsets* na tentativa de codificar o conceito expresso pelos SLRs. Por SLR, entende-se uma combinação de palavras que, apesar de frequente, apresentam baixo grau de estabilidade e fixação (BENTIVOGLI, PIANTA, 2003).

Seguindo-se essa regra, os nomes "foco" e "dengue", por exemplo, que constituem o SLR "foco da dengue", foram anotados separadamente com seus respectivos *synsets.*

O nome "foco" foi traduzido por *source* e, com base nesse equivalente, anotado com o *synset* {beginning, origin, root, rootage, source} ("*the place where something begins, where it springs into being*") ("lugar em que algo origina-se").

O nome "dengue", por sua vez, foi traduzido para *dengue* e anotado com o *synset* {dengue, dengue fever, dandy fever, breakbone fever} (*"an infectious disease of the tropics transmitted by mosquitoes and characterized by rash and aching head and joints"*) ("uma doença infecciosa dos trópicos transmitida por mosquitos e caracterizada por erupções cutâneas e dores de cabeça e nas articulações").

REGRA E3: *Analisar todas as traduções sugeridas pelo MulSen e os respectivos synsets*

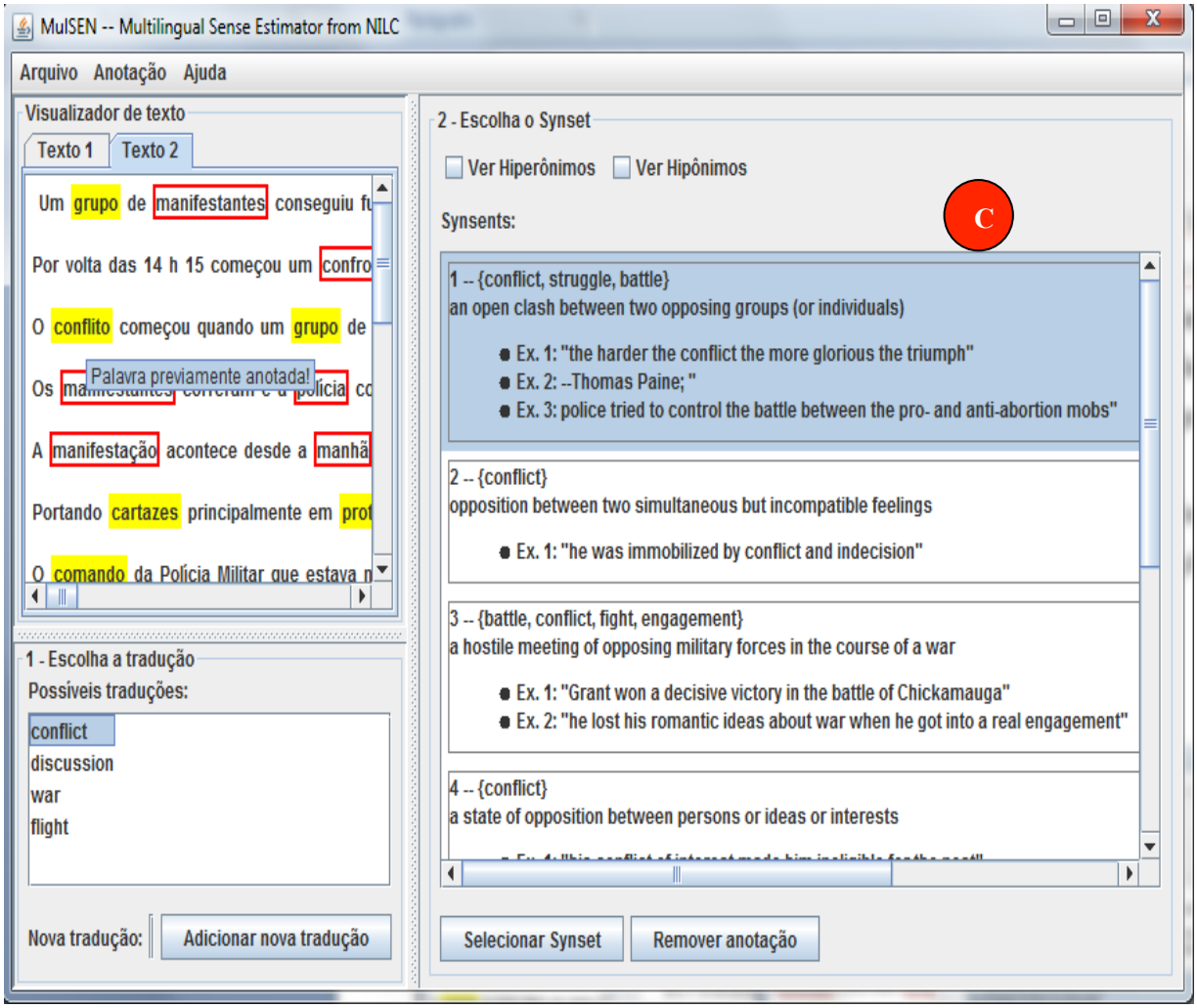
Essa regra estabeleceu que todas as traduções sugeridas pelo MulSen fossem analisadas antes da seleção definitiva do equivalente de tradução, assim como os *synsets* sugeridos para cada uma delas. Essa regra foi estabelecida com o objetivo de se selecionar a tradução mais adequada em inglês, principalmente para a anotação das palavras em português.

Especificamente, a partir da etiquetação morfossintática de uma palavra *x* em português, o editor MulSen recupera, quando disponível, os equivalentes de tradução em inglês do dicionário *online* WordReference e, aplicando o método de DLS, sugere um possível *synset* para cada equivalente de tradução.

Na Figura 23, observa-se que, para a palavra "conflito", por exemplo, o editor sugeriu 4 equivalentes na janela (b): *conflict*, *discussion*, *war* e *fight*. Segundo a Regra geral 3, todas as 4 equivalências foram analisadas para a adequada seleção da tradução e também do *synset.*

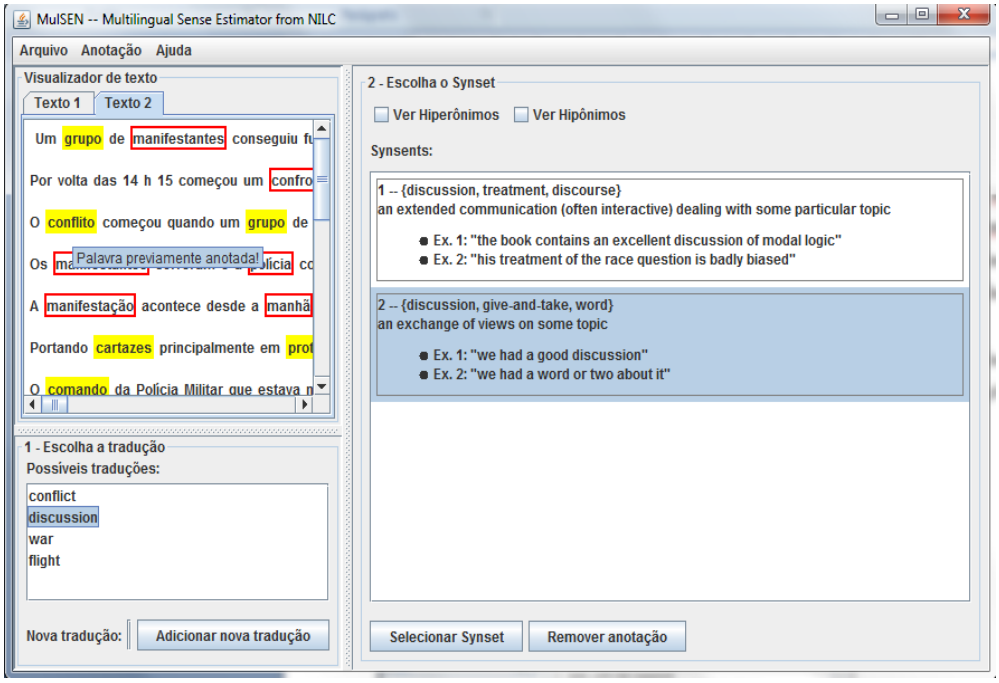
Analisando-se as sugestões, observa-se na Figura 23 que, para *conflict,* o *synset* sugerido na janela (c) foi o 1, ou seja, {conflict, struggle, tabble}, cuja glosa é "*an open clash between two opposing groups (or individuals)*" ("um confronto aberto entre dois grupos opostos (ou indivíduos)").

Figura 23 – Ilustração da tradução 1 sugerida na janela (b) do MulSen.



Na Figura 24, observa-se que, para a tradução *discussion,* o *synset* sugerido na janela (c) foi o de número 2, isto é, {discussion, give-and-take, word}, cuja glosa é "*an exchange of views on some topic*" ("uma troca de pontos de vista sobre algum tema").

Figura 24 – Ilustração da tradução 2 sugerida na janela (b) do MulSen.



Após analisar também *war* e *fight,* os anotadores selecionaram a tradução *conflit* e confirmaram a anotação de "conflito" com o *synset* {conflict, struggle, tabble} sugerido pelo editor.

REGRA E4: *Testar diferentes equivalentes antes de adicionar uma tradução ao MulSen*

Essa regra estabeleceu que diferentes traduções, quando existentes, fossem testadas antes de adicionar a expressão em inglês ao editor.

Essa regra foi estabelecida especificamente para os casos em que o editor não recuperava nenhuma possível tradução pelo acesso automático ao WordReference. Nesses casos, os anotadores deveriam incluir manualmente um equivalente no campo "Nova tradução" e adicioná-lo ao editor através do botão "Adicionar nova tradução" da janela (b). Somente a partir da inclusão de um equivalente, o editor verifica se o mesmo consta na base da WN.Pr para exibir, na sequência, os *synsets* dos quais o equivalente é elemento constitutivo.

O motivo para o estabelecimento da Regra 4 foi o fato de que, por vezes, uma palavra *y* inserida pelos anotadores não consta na WN.Pr. Isso, no entanto, não significa necessariamente que o conceito não está codificado na base, mas sim que a unidade *y* não está armazenada na base de dados.

Por conseguinte, a Regra 4 determinou que as várias formas sinônimas ou possibilidades de tradução, quando existentes, fossem testadas para que possíveis *synsets* correspondentes fossem recuperados da WN.Pr.

Vale ressaltar que a sugestão do *synset* adequado feita pelo editor só ocorria para as palavras provenientes do *WordReference* que constavam na WN.Pr. Diante da inclusão de uma tradução nova, os anotadores não contavam com a sugestão automática, sendo a escolha final do *synset* resultado exclusivo da análise dos anotadores.

Para selecionar o equivalente mais adequado, os anotadores podiam utilizar diversos recursos externos à ferramenta MulSen, como dicionários e serviços *online.* Entre os dicionários, estavam a versão *online* do *Michaelis Moderno Dicionário Inglês & Português*[36] e os diferentes dicionários disponíveis no site *Cambridge Dictionaries Online*[37]. Os serviços *online* especificados foram o *Google translate*[38] e o Linguee[39].

REGRA E5: *Selecionar os synsets mais adequados para anotar os nomes*
Essa regra estabeleceu que fosse selecionado o *synset* que representasse mais adequadamente o conceito subjacente a um nome *x.*

Especificamente a Regra 5 estabeleceu que, uma vez selecionada a tradução e analisados todos os *synsets* recuperados pelo editor, inclusive o sugerido pelo método de DLS, o *synset* mais adequado fosse escolhido para anotar a palavra em questão.

Essa regra foi formulada principalmente porque a WN.Pr, por vezes, apresenta conceitos muito próximos, cuja distinção nem sempre é simples.

Por exemplo, no texto-fonte em português da coleção C17 do CM2News, que relata "um confronto entre manifestantes e policiais horas antes do jogo de abertura da Copa das Confederações", ocorreu a palavra "apoio", cujo cotexto está descrito em (4):

(4) [...] Portando cartazes principalmente em protesto contra as obras da Copa e em **apoio** ao movimento do passe livre. [...]

Como equivalentes de tradução, o MulSen sugeriu *support, base, basis* e *foundation.* Após a análise por porte dos anotadores, o equivalente escolhido foi *support.* Dentre os 11 conceitos expressos por *support* (Quadro 4)*,* o método de DLS sugeriu o *synset* 1,

[36]Disponível em: <http://michaelis.uol.com.br/>.
[37]Disponível em: <http://dictionary.cambridge.org/>.
[38]Disponível em: <http://translate.google.com.br/>.
[39]Disponível em: <http://www.linguee.com.br/>.

{support} (“*the activity of providing or maintaining by supplying with money or necessities*”) (“a atividade de prover ou manter, pelo fornecimento de dinheiro ou outras necessidades”).

Analisando-se a sugestão e os demais conceitos, observa-se que o *synset* 1 não é o adequado e que os *synsets* 2 e 3, em especial, representam conceitos muito próximos, cuja distinção é bastante questionável. Com base na Regra 5, os anotadores selecionaram o *synset* 2; entretanto, o *synset* 3 também parece adequado.

Quadro 4 – Conceitos subjacentes a *support* e seus respectivos *synsets*.

	Synset	**Glosa/Frase-exemplo (Tradução da glosa)**
1	{support}	the activity of providing for or maintaining by supplying with money or necessities; "his support kept the family together"; "they gave him emotional support during difficult times" (“atividade de prover ou manter, fornecendo dinheiro ou condições essenciais à vida”)
2	{support}	aiding the cause or policy or interests of; "the president no longer had the support of his own party"; "they developed a scheme of mutual support" (“ ajuda à causa política ou aos interesses de”)
3	{support}	something providing immaterial assistance to a person or cause or interest; "the policy found little public support"; "his faith was all the support he needed"; "the team enjoyed the support of their fans" (“algo que provê assistência imaterial a uma pessoa ou causa ou interesse”)
4	{support, reinforcement, reenforcement}	a military operation (often involving new supplies of men and materiel) to strengthen a military force or aid in the performance of its mission; "they called for artillery support" (“uma operação militar (muitas vezes envolvendo novos suprimentos de homens e material) para fortalecer a força militar ou ajuda no desempenho de sua missão”)
5	{documentation, support}	documentary validation; "his documentation of the results was excellent"; "the strongest support for this view is the work of Jones" (“validação documental”)
6	{support, keep, livelihood, living, bread and butter sustenance}	the financial means whereby one lives; "each child was expected to pay for their keep"; "he applied to the state for support"; "he could no longer earn his own livelihood" (“os meios financeiros por meio das quais se vive”)
7	{support}	supporting structure that holds up or provides a foundation; "the statue stood on a marble support" (“estrutura de suporte que sustenta ou fornece uma base”)
8	{support, supporting}	the act of bearing the weight of or strengthening; "he leaned against the wall for support" (“o ato de suportar o peso ou reforçar”)

9	{accompaniment, musical accompaniment,backup,support}	a subordinate musical part; provides background for more important parts (“uma parte subsidiária da melodia”)
10	{support}	any device that bears the weight of another thing; "there was no place to attach supports for a shelf (“qualquer dispositivo que carrega o peso de outra coisa”)
11	{support, financial support, funding, backing, financial backing }	financial resources provided to make some project possible; "the foundation provided support for the experiment" (“recursos financeiros previstos para fazer algum projeto possível”)

Fonte: Adaptada de Fellbaum, (1998).

REGRA E6: *Selecionar synsets hiperônimos*

A Regra 6 estabeleceu que, diante da inexistência de um *synset* que representasse o conceito específico subjacente a uma palavra, o *synset* hiperônimo (ou seja, mais genérico) fosse selecionado.

Por exemplo, no texto em português da coleção C2 do CM2News, que relata a “suspensão da produção e distribuição de um kit anti-homofobia”, ocorreu a sigla CPI (Comissão Parlamentar de Inquérito), cujo cotexto está descrito em (5):

(4) [...]convocar o ministro Antonio Palocci a se explicar sobre sua evolução patrimonial e propor uma **CPI** para investigar [...].

Essa sigla expressa um conceito específico do domínio político no cenário brasileiro e, por isso, não está armazenado na WN.Pr.

Seguindo-se a Regra 6, a sigla foi anotada com o *synset* que representa um conceito mais genérico, necessitando da inserção de um equivalente de tradução que expressasse esse conceito mais genérico.

Quanto à sigla do exemplo em (4), ressalta-se que, para CPI, inseriu-se o equivalente *investigation* e, por meio dele, o MulSen retornou os *synsets* descritos no Quadro 5. Destes, o *synset* 1 foi selecionado pelos anotadores. A seleção do *synset* 1 pode ter sido feita com base principalmente na frase-exemplo, “*there was a congressional probe into the scandal*” (“houve uma investigação do congresso sobre o escândalo”), que ilustra o conceito em uso exatamente no contexto político.

Quadro 5 – Subjacentes a *investigation* e seus respectivos *synsets*.

	Synset	**Glosa/Frase-exemplo (Tradução da glosa)**
1	{probe, investigation}	an inquiry into unfamiliar or questionable activities. "there was a congressional probe into the scandal" ("uma investigação sobre atividades desconhecidas ou questionáveis")
2	{investigation, investigating}	the work of inquiring into something thoroughly and systematically ("o trabalho de se investigar algo de forma completa e sistemática")

Fonte: Adaptada de Fellbaum, (1998).

REGRA E7: *Anotar o núcleo de expressões metafóricas*

A Regra 7 estabeleceu que apenas os nomes nucleares em expressões metafóricas fossem anotados com o *synset* correspondente ao conceito da expressão.

Por exemplo, no texto em português da coleção C5 do CM2News, que engloba notícias sobre a "aprovação, na Câmara dos Deputados, do texto-base da reforma do Código Florestal", ocorreu o nome "feixe" que, apesar de etiquetado em isolado, é núcleo da expressão "feixe de lenha", cujo cotexto está descrito em (6).

(6) [...] Como relator, não aguento mais amarrar e desamarrar esse **feixe** de lenha e carregá-lo por mais tempo [...].

De acordo com o cotexto da C5 em que a expressão ocorreu, os anotadores interpretaram que "feixe de lenha" foi empregado em sentido metafórico, referindo-se ao "texto final da reforma florestal".

Seguindo-se a Regra 7, apenas "feixe" foi anotado. Para tanto, essa palavra foi traduzida para *text*, por meio dela, selecionou-se o *synset* {text, textual matter} (*"the words of something written*") ("as palavras de algo escrito").

Os dados gerados pela anotação em questão serviram de base para a aplicação dos métodos de SAMM deste trabalho (item 4).

3.7 Criação de sumários de referência

Para viabilizar a avaliação automática da informatividade dos sumários gerados por ambos os métodos, criaram-se sumários de referência para o *corpus* CM2News.[40]

Para tanto, organizou-se uma equipe de 13 linguistas computacionais membros do grupo de SA do NILC, que se reuniu em um encontro presencial, com duração de 90 minutos. Os 13 linguistas computacionais seguiram um protocolo simples para a elaboração dos sumários, o qual visou a delimitação e uniformização da tarefa.

O protocolo em questão continha os seguintes passos: (a) leitura de ambos os texto da coleção; (b) elaboração de um sumário abstrativo referente à coleção em questão com tamanho equivalente a 30% (em número de palavras) do maior texto da coleção, já que a taxa de compressão foi estipulada em 70%.

Os sumários foram criados (i) de forma abstrativa, ou seja, foram elaborados livremente pelos participantes, contendo livre reescrita dos textos-fonte e (ii) de forma a serem informativos, ou seja, os sumários foram elaborados com o objetivo de representarem os textos-fonte, contendo as informações mais relevantes (informativas) da coleção, a ponto de substituir a leitura da mesma.

Cada participante recebeu 1 ou 2 coleções do CM2News, composta por 1 texto em inglês e 1 em português. Ao final, criou-se um sumário de referência para cada uma das 20 coleções do CM2News.

[40] Sumáro de referência é um sumário considerado ideal, normalmente produzido por humano.

4 MÉTODOS EXTRATIVOS DE SAMM INVESTIGADOS

4.1 Descrição e aplicação dos métodos

Neste mestrado, investigou-se a aplicação de conhecimento léxico-conceitual na SAMM envolvendo português de tal forma que a etapa de TA dos textos-fonte não fosse realizada.

Além das hipóteses sobre a aplicação de conhecimento profundo do tipo léxico-conceitual, este trabalho foi motivado pelo fato de que o único trabalho sobre SAMM envolvendo o português utiliza conhecimento superficial.

Dessa forma, foram desenvolvidos 2 métodos de SAM baseados em conhecimento léxico-conceitual. Tratam-se de métodos extrativos de SAMM, capazes de sumarizar uma coleção composta por textos em português e inglês, selecionando conteúdo com base em conhecimento linguístico de nível léxico-conceitual, para gerar um sumário em português.

No Método 1, as sentenças originais em inglês e em português são pontuadas e ranqueadas em função da frequência, na coleção, de seus conceitos lexicalizados. Nesse método, apenas as sentenças mais bem pontuadas em português são selecionadas para compor o sumário até que a taxa de compressão seja atingida.

O Método 1 pauta-se na hipótese de que um sumário composto exclusivamente por sentenças originais provenientes do texto-fonte em português reflete as informações mais relevantes de toda uma coleção bilíngue de textos-fonte, pois as sentenças foram pontuadas e ranqueadas levando-se em consideração a frequência de ocorrência dos conceitos em toda a coleção, ou seja, nos textos em português e em inglês.

No Método 2, assim como no Método 1, as sentenças originais em inglês e em português são pontuadas e ranqueadas em função da frequência de seus conceitos constitutivos na coleção. Nesse método, no entanto, as sentenças são selecionadas até que a taxa de compressão seja atingida independentemente de sua língua de origem; nos casos em que sentenças em inglês são selecionadas, estas são traduzidas de forma automática para o português.

A proposição do Método 2 recai sobre a hipótese de que um sumário composto por sentenças originais em português e por sentenças traduzidas para o português apresenta menos problemas de TA do que um sumário produzido a partir de uma

coleção composta por textos integralmente traduzidos para o português e originais nessa mesma língua (como em Tosta et al (2013)).

A aplicação dos métodos às 20 coleções do CM2News foi feita de forma automática, ou seja, por meio de uma rotina computacional. A partir dos arquivos no formato exemplificado na Figura 21 (p. 71), a rotina computacional realizou 2 tarefas necessárias à aplicação dos métodos de SAMM, a saber: (i) segmentação em nível sentencial dos textos-fonte anotados e (ii) pontuação e ranqueamento das sentenças em função da frequência de ocorrência dos conceitos nominais anotados na coleção (item 4.2).

4.2 Ranqueamento e criação dos sumários

Para a criação do ranque, calculou-se de forma automática a frequência de cada *synset* em cada uma das coleções. A frequência de ocorrência do *synset* em uma coleção C equivale ao número de vezes que o *synset* em questão foi anotado na coleção. Na Figura 25, observa-se, por exemplo, que a pontuação do *synset* de número **<7210553>,** indexado às palavras "*protesters"* e "manifestantes", é "**(16)"** e que a pontuação do *synset* de número **<74522699>**, indexado às palavras "*opening*" e "abertura", é de "**(2)**". A soma das pontuações dos *synsets* de uma sentença S resulta na pontuação final dessa sentença, que representa a sua importância. No caso, a soma das frequências dos *synsets* da sentença "*Brazil's opening Confederations Cup match was affected by protesters that feft 39 people injured"* da Figura 25 resultou na pontuação **"28"**. A sentença em português, por sua vez, obteve a pontuação **"51"**.

Figura 25 – Unidades lexicais e frequência de *synsets:*

Brazil's **<9379111>(4)** **opening<74522699>(2)** Confederations Cup **match<7470671>(5)** was affected by **protesters<10002760>(16)** that feft 39 **people<7942152>(1)** injured. =28

.....

Um **grupo<31264>(6)** de **manifestantes<10002760>(16)** conseguiu furar o bloqueio<8376948>(2) da Polícia Militar e chegar ao **estádio<4295881>(14)** Mané Guarrincha neste **sábado<15164570>(4)**, **horas<15227846>(2)** antes do **jogo<7470671(5)** de **abertura<7452699>(2)** da Copa das Confederações. =51

.....

Após o cálculo da importância de cada sentença em uma coleção C, construiu-se o ranque desses sentenças, a partir do qual os diferentes métodos de seleção de conteúdo se basearam. As sentenças com maior pontuação ocupam o topo do ranque. No Quadro 6, apresenta-se o ranque da coleção 17.

Quadro 6 – Exemplo de ranque de sentenças (Cluster 17).

Colocação	**Sentenças de C17**	**Pontuação**
1º	Um grupo de manifestantes conseguiu furar o bloqueio da Polícia Militar e chegar ao estádio Mané Garrincha neste sábado, horas antes do jogo de abertura da Copa das Confederações.	51
2º	O Governo do Distrito Federal divulgou uma nome que informa que os manifestantes que se concentram nas imediações do estádio Mané Garrincha estão contidos e são acompanhados pelas forças policiais.	45
3º	BBC Sport's South American football correspondent Tim Vickery told BBC World Service: Brazilian society was explicitly told in 2007 that all of the money spent on stadiums would be private money.	43
4º	A tropa de choque e a cavalaria da PM cercou os manifestantes que andavam em grupo de um lado e de outro nas entradas do estádio.	43
5º	Up to 1,000 Brazilians demonstrated outside the country's national stadium to vent their anger at the amount of money the country is spending on staging next year's World Cup.	41
6º	A polícia montou um cordão de isolamento em volta do estádio, o que aumentou a confusão com os torcedores que chegavam e que procuravam informações sobre os portões de entrada.	39
7º	Police used tear gas and pepper spray to control protestors before the match, in which Brazil beat Japan 3 - 0.	35
8º	More than 90% of the money being spent on football stadiums is public money.	33
9º	O conflito começou quando um grupo de manifestantes começou a hostilizar e atirar objetos nos torcedores que estavam na fila.	32
10º	Portando cartazes principalmente em protesto contra as obras da Copa e em apoio ao movimento do passe livre em São Paulo, eles conseguiram chegar próximo a um dos portões de entrada do estádio.	31
11º	Demonstrators held up posters reading: We don't need the World Cup and We need money for hospitals and education.	30
12º	Brazil's opening Confederations Cup match was affected by protests that left 39 people injured.	28
13º	O acordo entre os organizadores e o comando da PM era que a manifestação seria feita apenas na rodoviária de Brasília e não subiria para o estádio, mas que acabou não sendo respeitado.	26

14º	Grupos de índios, punks, pessoas pedindo dinheiro para a saúde e de professores reclamando melhores salários chegaram depois para reforçar o protesto.	26
15º	Um manifestante foi ferido com um tiro de borracha na perna, caiu no chão e machucou o rosto.	26
16º	The Confederations Cup is a dress rehearsal for the World Cup next year and the stadium in Brasilia used for the inaugural match was one of the most expensive of the six built, costing around £ 380 m.	25
17º	Por fim, o governo do Distrito Federal informa que, neste momento, os torcedores ingressam na arena sem qualquer tumulto.	25
18º	Uma equipe da Rede Globo de Brasília foi hostilizada pelos manifestantes que xingaram os jornalistas e também alguns torcedores.	25
19º	Ele justificou o uso de bombas de efeito moral como técnica para tentar conter a evolução dos manifestantes.	25
20º	Vickery added: Of the 12 stadiums for the World Cup, some of them are in the heartlands of Brazilian football will be very well used indeed.	24
21º	Um dos manifestantes levou gás de pimenta no rosto.	24
22º	Brasilia does not have a team in either of the first two divisions of Brazilian football and doesn't really have much of a local football tradition at all.	23
23º	Segundo ele, a maior dificuldade em negociar com os manifestantes é que são vários grupos heterogêneos.	23
24º	Os manifestantes jogaram livros da Constituição ao pé dos policiais e muitos distribuíam flores.	23
25º	Os manifestantes correram e a polícia começou a persegui-los.	23
26º	Dessa forma, os torcedores que se encaminham ao estádio Mané Garrincha têm o acesso garantido normalmente.	22
27º	Em nota, o governo do Distrito Federal falou em 200 integrantes.	21
28º	O temor da polícia é com a grande quantidade de torcedores que chegavam para a cerimônia de abertura da Copa das Confederações.	19
29º	Durante a tentativa de fuga, um manifestante chegou a ser atropelado por uma moto da PM.	19
30º	Ainda de acordo com a nota, a Polícia Militar fez uso progressivo da força desde as primeiras horas da manifestação, garantido atuação pacífica e o controle absoluto do movimento.	18
31º	Por volta das 14 h 15 começou um confronto entre os manifestantes e a Polícia Militar.	18
32º	A polícia jogou bombas de efeito moral e atirou com balas de borracha.	18
33º	O texto informa ainda que não será permitida a perturbação da ordem pública e nem qualquer tipo de ameaça à realização do jogo e ao público participante dessa grande festa para o Distrito Federal.	16

34º	Por volta das 12 h a polícia jogou bombas de efeito moral e gás de pimenta para tentar contê-los.	16
35º	Em torno do estádio, são 1.700.	14
36º	O tenente-coronel da PM Zilfrank Antero, comandante da Comunicação, disse que até o início da tarde não havia sido registrado nenhum incidente nem prisões.	9
37º	A manifestação acontece desde a manhã de sábado.	9
38º	There were also reports rubber bullets were used and 30 arrests were made.	8
39º	A cada movimento do grupo, eles eram seguidos.	8
40º	Isadora Rodrigues, 22, estudante de artes cênicas da UnB, jogava bolhas de sabão nos policiais.	7
41º	O comando da Polícia Militar que estava no local informou que eram cerca de 700 pessoas.	7
42º	Ao todo, 3.200 policiais fazem a segurança do evento em Brasília.	6
43º	Houve tumulto e confusão.	5
44º	It hasn't worked out that way at all.	4
45º	Torcedores ficaram no meio da confusão.	4
46º	I'm a little bit dubious about that one.	3
47º	The idea seems to be that it is going to be viable with pop concerts.	2
48º	As únicas bandeiras de partido político era do PSTU.	2
49º	But there are four - Brasilia plus three others - that you really wonder where the long-term viability will be.	1
50º	Essa é a nossa arma, disse.	1
51º	Vários foram presos.	0

Após construído o ranque sentencial, o próximo passo é o de seleção das sentenças para o sumário, processo que é descrito nos algoritmos dos Métodos 1 e 2, que são apresentados na próxima Seção.

4.2.1 Método 1: seleção com base na frequência de conceitos e na língua do usuário

No Método 1, a seleção de conteúdo é feita com base na língua do usuário que, no caso, é o português. Dessa forma, a partir do ranque das sentenças de ambos os textos-fonte de uma coleção C, apenas as sentenças em português mais bem pontuadas são selecionadas para compor o extrato informativo e genérico correspondente à coleção C, até que a taxa de compressão seja atingida.

Na Figura 26, descreve-se o algoritmo do Método 1.

Figura 26 – Algoritmo do Método 1.

MÉTODO 1	
ANÁLISE	1. Analisar cada um dos textos da coleção em nível léxico-conceitual, ou seja, anotar os substantivos (nomes comuns) com os conceitos/synsets da WordNet de Princeton.
TRANSFORMAÇÃO	2. Calcular a taxa de compressão em 70% (calculado a partir do número de palavras do maior texto da coleção) 3. Pontuar as sentenças em função da frequência de ocorrência dos *synsets*/conceitos na coleção 4. Ranquear as sentenças em função da pontuação dos conceitos 5. Selecionar a 1ª sentença do ranque que seja proveniente do texto em português 6. Caso a taxa de compressão não tenha sido atingida: 6.a. Selecionar a próxima sentença em português do ranque 6.b. Verificar a redundância da sentença em questão com a já selecionada 6.c. Eleger a sentença somente se não for redundante 7. Repetir o passo 6 até que a taxa de compressão seja atingida
SÍNTESE	8. Justapor as sentenças na ordem em que foram selecionadas 9. Ordenar os segmentos/sentenças pela ordem de ocorrência nos textos-fonte.

Como mencionado, o Método 1 produziu sumários compostos exclusivamente por sentenças originais em português que reflitam as informações mais relevantes de toda a coleção, pois as sentenças são ranqueadas em função dos conceitos que também ocorreram nos textos em inglês.

Para garantir que as sentenças selecionadas do ranque não fossem similares entre si, aplicou-se um fator de redundância, calculado por meio da medida *word overlap,* a mesma utilizada em Tosta et al. (2013). Caso a medida *word overlap* entre uma sentença já selecionada e uma sentença do ranque candidata a compor o sumário tivesse sido superior a um limiar determinado empiricamente (do inglês, *threshold*), a sentença candidata não era selecionada.

Quanto à taxa de compressão, especificou-se o valor de 70%, a mesma utilizada para a geração dos sumários automáticos e humanos do CSTNews (CARDOSO et al, 2011), *corpus* multidocumento de referência para as pesquisas com o português no cenário da SAM. Assim, os sumários produzidos pelo Método 1 apresentam tamanho equivalente a 30% do maior texto-fonte da coleção (medido em número de palavras).

A produção dos sumários foi realizada pela justaposição das sentenças na ordem em que ocorreram nos textos-fonte, sob a hipótese de que tal ordenação pudesse melhorar a coesão/estrutura textual. O truncamento ou parada na seleção das sentenças deu-se a partir do valor que se distanciou menos da taxa de compressão. Dessa forma, caso a última sentença selecionada ultrapassasse 15 palavras da taxa de compressão, por exemplo, e a sua deleção resultasse em uma diferença de 5 palavras a menos do valor da taxa de compressão, a opção foi por manter o sumário com tamanho inferior à taxa de compressão, optando, assim, pela opção mais próxima ao limiar.

4.2.2 Método 2: seleção com base na frequência de conceitos

No Método 2, a seleção de conteúdo é feita com base exclusivamente no ranque das sentenças em função dos conceitos, sem preferência à língua de origem.

Assim, a partir do ranque das sentenças de uma coleção C, a seleção automática de conteúdo no Método 2 é apresentada pelo algoritmo da Figura 27.

Figura 27 – Algoritmo do Método 2.

MÉTODO 2	
ANÁLISE	1. Analisar cada um dos textos da coleção em nível léxico-conceitual, ou seja, anotar os substantivos (nomes comuns) com os conceitos/synsets da WordNet de Princeton.
TRANSFORMAÇÃO	2. Calcular a taxa de compressão em 70% (calculado a partir do número de palavras do maior texto da coleção) 3. Pontuar as sentenças em função da frequência de ocorrência dos *synsets*/conceitos na coleção 4. Ranquear as sentenças em função da pontuação dos conceitos 5. Selecionar a 1ª sentença do ranque 6. Caso a taxa de compressão não tenha sido atingida: 6.a. Selecionar a próxima sentença do ranque 6.b.Traduzir a sentença selecionada para o português, caso seja do inglês 6.c. Verificar a redundância da sentença em questão com a já selecionada 6.d. Eleger a sentença somente se não for redundante 7. Repetir o passo 6 até que a taxa de compressão seja atingida
SÍNTESE	8. Justapor as sentenças na ordem em que foram selecionadas 9. Ordenar os segmentos/sentenças pela ordem de ocorrência nos textos- fonte. Em caso de empate entre sentenças originais e traduzidas, selecionar a menor.

Assim como estipulado para o Método 1, aplicou-se um fator de redundância no Método 2 para garantir que as sentenças selecionadas do ranque não fossem similares entre si. Nesse caso, verificou-se a similaridade ou redundância de uma sentença S com todas as sentenças da coleção.

A partir do ranque das sentenças de ambos os textos-fonte de uma coleção C, a primeira sentença mais bem pontuada, seja ela em português ou em inglês, foi selecionada para compor o extrato informativo e genérico correspondente à coleção C e, assim, sucessivamente, até que a taxa de compressão fosse atingida. Caso sentenças em inglês fossem selecionadas, as mesmas seriam traduzidas de forma automática por meio do serviço grátis *online* de tradução, *Bing translator*.[41]

Quanto à taxa de compressão, utilizou-se o mesmo valor aplicado à geração dos sumários segundo o Método 1 (ou seja, 70%).

Assim como no Método 1, a seleção de conteúdo foi realizada de forma automática.

[41] Disponível em: <http://www.bing.com/translator>.

5 AVALIAÇÃO DOS MÉTODOS

Nesta investigação, a avaliação adotada foi a intrínseca. Em especial, avaliaram-se os métodos quanto à qualidade linguística e informatividade dos sumários. A avaliação intrínseca da qualidade linguística foi realizada manualmente, e a da informatividade de forma automática. A opção pela avaliação intrínseca deu-se pela necessidade de avaliação da qualidade linguística dos sumários gerados.

5.1 Avaliação instrínseca da qualidade linguística

A avaliação da qualidade linguística adotada para este trabalho seguiu os parâmetros propostos pela DUC´2007. Especificamente, os sumários foram avaliados de forma manual quanto aos cinco critérios propostos pela DUC ("gramaticalidade", "não-redundância", "clareza referencial", "foco" e "estrutura e coerência") (ver item 2.4.1).

Para viabilizar a avaliação em questão, as 20 coleções do *corpus* foram divididas em 5 grupos de 4 coleções cada. Cada um dos grupos continha os sumários gerados pelos Métodos 1 e 2, totalizando 8 sumários automáticos. Cada um dos 5 grupos de sumários foi analisado por 3 juízes diferentes. No total, essa avaliação contou com a contribuição de 15 linguistas computacionais.

Para a realização do procedimento, desenvolveram-se formulários *online* por meio dos quais os juízes leram, avaliaram e os submeteram a julgamento individualmente. Os formulários continham: (i) a descrição da tarefa, (ii) a exemplificação dos critérios de qualidade e regras de pontuação, e (iii) os sumários a serem avaliados. Cada sumário avaliado era seguido dos cinco critérios linguísticos mencionados, sendo que, para cada critério, o juiz deveria atribuir uma nota de "1" a "5", de acordo com a escala apresentada no Quadro 7.

Quadro 7 – Pontuações e níveis para a avaliação da qualidade linguística.

Pontuação	Nível
1	Péssimo
2	Ruim
3	Regular
4	Bom
5	Excelente

Como regra, cada critério deveria ser avaliado individualmente.

5.1.1 Resultados da avaliação da qualidade linguística

Na Tabela 2, observam-se as médias de cada uma das pontuações atribuídas pelos juízes quanto ao critério de gramaticalidade:

Tabela 2 – Pontuações dos métodos: critério de "gramaticalidade"

	Gramaticalidade										
	Péssimo (1)		Ruim (2)		Regular (3)		Bom (4)		Excelente (5)		Média
Método 1	**1**	**1,9%**	**5**	**9,6%**	**3**	**5,7%**	**15**	**28%**	**28**	**53%**	4,3 (**bom**)
Método 2	4	7,6%	11	21,1%	12	23,7%	13	25%	12	23%	3,5 (**regular**)

Na média, a gramaticalidade dos sumários gerados pelo Método 1 obteve nível "bom", uma vez que ela recebeu a pontuação média de 4,3. A gramaticalidade dos sumários gerados pelo Método 2 foi definida, em média, como "regular", pois sua média foi de 3,5. Apesar do Método 1 não ter sido considerado, em média, "excelente", esse método recebeu 28 vezes a pontuação "excelente" dos juízes, contra 12 do Método 2. Dessa forma, percebe-se que o método 1 gera sumários extrativos com menos problemas de ortografia, pontuação e sintaxe. A seleção de sentenças traduzidas automaticamente pelo Método 2 pode ter influenciado suas médias inferiores para o critério em questão.

Quanto ao critério de não-redundância, o Método 1 também mostra certa superioridade em relação ao Método 2 (Tabela 3).

Tabela 3 – Pontuações dos métodos quanto a "não-redundância"

	Não-redundância										
	Péssimo (1)		Ruim (2)		Regular (3)		Bom (4)		Excelente (5)		**Média**
Método 1	0	**0%**	3	**5,7%**	6	**11,5%**	17	**32,6%**	25	**48,1%**	4,3 (**bom**)
Método 2	0	**0%**	11	**21,1%**	15	**28,3%**	18	**28,3%**	8	**40%**	3,4 (**regular**)

O Método 1 recebeu a atribuição "excelente" 25 vezes, enquanto que o Método 2 recebeu a mesma atribuição 8 vezes. Considerando a pontuação "ruim", o Método 2 apresenta uma média de 21% contra 5,7% do Método 2. Esses resultados que mostram a superioridade do Método 1 não são surpreendentes, já que, no Método 1, as sentenças

selecionadas para o sumário são provenientes de um mesmo texto-fonte, o que normalmente resulta em um sumário menos redundante.

O critério de clareza referencial apresentado pela Tabela 4 demonstra, em média, a maior discrepância encontrada entre os Métodos 1 e 2. Essa variação deu-se quanto à frequência da pontuação "excelente" (Tabela 4).

Tabela 4 – Pontuações dos métodos quanto à "clareza refencial"

	Clareza Referencial										
	Péssimo (1)		Ruim (2)		Regular (3)		Bom (4)		Excelente (5)		**Média**
Método 1	1	**1,9%**	4	**7,6%**	15	**28,8%**	17	**32,6%**	14	**40,3%**	3,7 (**bom**)
Método 2	5	**9,6%**	8	**15,3%**	16	**30,7%**	18	**34,6%**	4	**7,6%**	3,3 (**regular**)

O Método 1 apresenta 40,3% da sua pontuação como "excelente" contra apenas 7,6% do Método 2. Ao se considerar a pontuação "péssimo", o Método 2 apresenta uma frequência de 9,6% , enquanto o Método 1 apresenta a média de 1,9%. Tal resultado também não é surpreendente, pois o Método 1, além de selecionar as sentenças de um único fonte-fonte, organiza a justaposição destas no sumário na mesma ordem em que ocorrem no texto-fonte, sob a hipótese de que tal odernação contribui para a clareza referencial do sumário.

Quanto ao "foco", o Método 1 ainda supera com uma grande margem a atribuição de pontuações "excelente", 40% contra 13,4% do Método 2, como se observa na Tabela 5.

Tabela 5 – Pontuações dos Métodos: critério de "foco"

	Foco										
	Péssimo (1)		Ruim (2)		Regular (3)		Bom (4)		Excelente (5)		**Média**
Método 1	0	**0%**	0	**0%**	8	**15,3%**	23	**44,2%**	21	**40,3%**	4,1 (**bom**)
Método 2	2	**3,8%**	5	**9,6%**	13	**25%**	25	**40%**	7	**13,4%**	3,5 (**regular**)

Ademais, o Método 1 apresenta porcentagem de 0% na ocorrência de pontuações "péssimo" e "ruim", enquanto que o Método 2 apresenta uma margem considerável de 3,8% e 9,6 % respectivamente, nessas mesmas pontuações.

O último critério de qualidade linguística analisado, "estrutura e coerência", é apresentado pela Tabela 6.

Tabela 6 – Pontuações dos Métodos: critério de “estrutura e coerência ”

	Estrutura e Coerência										
	Péssimo (1)		Ruim (2)		Regular (3)		Bom (4)		Excelente (5)		**Média**
Método 1	0	**0%**	8	**15,3%**	18	**34,6%**	21	**40,3%**	5	**9,6%**	3,4 (**bom**)
Método 2	9	**17,3%**	13	**25%**	20	**38,4%**	8	**15,3%**	2	**3,8%**	2,6 (**ruim**)

Para o critério em questão, observa-se que nenhum dos métodos apresentou porcentagens significativas quanto à atribuição de pontuações “excelente”. No entanto, para a pontuação “bom”, observa-se que o Método 1 supera com sua média de 40,3% os 15,3% do Método 2. A superioridade do Método 1 para o critério em questão acentua-se ao observarmos as médias das pontuações “péssimo” e “ruim”. Enquanto o Método 1 apresenta 0% de pontuações “péssimo” e 15,3% de atribuições “ruim”, o Método 2 apresenta respectivamente as médias 17,3% e 25%. Mais uma vez, o resultado em questão não é revelador, pois ao selecionar sentenças de 2 textos-fonte distintos, o Método 2 tende a produzir sumários menos coerentes e coesos, ou até mesmo contraditórios, fenômeno este considerado comum no contexto multidocumento.

Por fim, as médias entre os resultados finais obtidos pela avaliação da qualidade linguística podem ser visualizadas detalhadamente pela tabela 7.

Tabela 7 – Média dos resultados da avaliação linguística dos Métodos 1 e 2

Critérios	**Método 1**	**Método 2**
Gramaticalidade	4,3	3,5
Não-redundância	4,3	3,4
Clareza referencial	3,7	3,3
Foco temático	4,1	3,5
Estrututura e Coerência	3,4	2,6

Pelos resultados obtidos, percebeu-se que, nos 5 critérios investigados, o Método 1 superou a qualidade linguística do Método 2. Tal resultado indica que a opção por eleger do ranque somente as sentenças do texto em português melhora significativamente a qualidade linguística do sumário.

Nas Figuras 28 e 29, apresentam-se os sumários produzidos a partir dos ranques exemplificados para o Método 1 e 2, respectivamente.

Figura 28 – Exemplo de Sumário do Método 1 (coleção 17)

Um grupo de manifestantes conseguiu furar o bloqueio da Polícia Militar e chegar ao estádio Mané Garrincha neste sábado, horas antes do jogo de abertura da Copa das Confederações.
O Governo do Distrito Federal divulgou uma nota que informa que os manifestantes que se concentram nas imediações do estádio Mané Garrincha estão contidos e são acompanhados pelas forças policiais.
A tropa de choque e a cavalaria da PM cercou os manifestantes que andavam em grupo de um lado e de outro nas entradas do estádio.
A polícia montou um cordão de isolamento em volta do estádio, o que aumentou a confusão com os torcedores que chegavam e que procuravam informações sobre os portões de entrada.
O conflito começou quando um grupo de manifestantes começou a hostilizar e atirar objetos nos torcedores que estavam na fila.
Portando cartazes principalmente em protesto contra as obras da Copa e em apoio ao movimento do passe livre em São Paulo, eles conseguiram chegar próximo a um dos portões de entrada do estádio.
Um manifestante foi ferido com um tiro de borracha na perna, caiu no chão e machucou o rosto.

Figura 29 – Exemplo de sumário do Método 2 (coleção 17)

Um grupo de manifestantes conseguiu furar o bloqueio da Polícia Militar e chegar ao estádio Mané Garrincha neste sábado, horas antes do jogo entre Brasil e Japão pela abertura da Copa das Confederações.
O Governo do Distrito Federal divulgou uma nota que informa que os manifestantes que se concentram nas imediações do estádio Mané Garrincha estão contidos e são acompanhados pelas forças policiais.
A tropa de choque e a cavalaria da PM cercou os manifestantes que andavam em grupo de um lado e de outro nas entradas do estádio.
Correspondente do BBC Sport, futebol sul-americano Tim Vickery disse a BBC World Service: sociedade brasileira foi explicitamente disse em 2007 que todo o dinheiro gasto com estádios seria dinheiro privado.
Até 1.000 brasileiros demonstraram frente ao Estádio Nacional do país para desabafar sua raiva com a quantidade de dinheiro, o país está gastando na preparação da Copa do mundo do próximo ano.
A polícia montou um cordão de isolamento em volta do estádio, o que aumentou a confusão com os torcedores que chegavam e que procuravam informações sobre os portões de entrada.
Polícia usou gás lacrimogêneo e spray de pimenta para controlar os manifestantes antes do jogo, em que o Brasil venceu Japão 3-0.

Ao observar os sumários resultantes de ambos os métodos, verficou-se clamente a interferência da qualidade da TA. No sumário referente ao Método 2, tal interferência resulta em uma qualidade linguística perceptivelmente inferior, não só ao se tratar do parâmetro de gramaticalidade, mas também no que diz respeito à estruturação e coerência textuais, que são claramente superiores no sumário do Método 1. Esse

resultado pode ser explicado pelo fato de que o sumário do Método 1 é composto apenas por sentenças do texto em português, que, além de não apresentarem conteúdo agramatical, apresentam menos redundância, se comparado ao contexto multidocumento.

Em contrapartida, o sumário exemplificado do Método 2 apresenta, mesmo que com gramaticalidade inferior, informações não presentes no sumário do Método 1, como por exemplo uma possível causa para o manifesto: "*Até 1.000 brasileiros demonstraram frente ao Estádio Nacional do país para desabafar sua raiva com a quantidade de dinheiro, o país está gastando na preparação da Copa do mundo do próximo ano.*"

Considerando que o único trabalho de SAMM utilizando o português que se tinha conhecimento era o de Tosta et al. (2012, 2013), compararam-se os resultados obtidos pelos referidos autores aos gerados pelos Métodos 1 e 2 propostos neste trabalho. Para tanto, considerou-se como parâmetro o método de Tosta et al. (2012, 2013) que obteve o melhor resultado, a saber: o método de localização com tratamento de redundância.

Na Tabela 8 a seguir, observa-se a comparação entre as médias das pontuações dos métodos deste trabalho (Métodos 1 e 2) e as do melhor método de Tosta et al. (2012, 2013).

Tabela 8 – Comparação com o melhor método de Tosta et al. (2013).

Critérios	**Método 1**	**Método 2**	**Método de Tosta et al., 2012, 2013**
Gramaticalidade	**4,3**	3,5	3
Não-redundância	**4,3**	3,4	3
Clareza referencial	**3,7**	3,3	3,2
Foco temático	**4,1**	3,5	4
Estrutura e Coerência	**3,4**	2,6	2,8

Ao se comparar o método de Tosta et al. (2013) com o Método 2, percebe-se que o Método 2, por um lado, supera o método de Tosta et al. em 3 dos 5 critérios avaliados (gramaticalidade, não-redundância e clareza referencial) e ainda, por outro lado, o

Método 2 é superado pelo de Tosta et al. em 2 dos critérios (estrutura e coerência e foco).

Ao serem comparados os resultados do método de Tosta et al. (2013) com o Método 1, observa-se que o Método 1 supera o melhor método de Tosta et al. em todos os critérios avaliados, a saber: gramaticalidade, não-redundância, clareza referencial, foco temático e estrutura/coerência. Essa superioridade pode ser justificada pelo fato de que o Método 1 seleciona apenas sentenças em português (original), as quais, portanto, são provenientes de um único texto-fonte, o que possibilita driblar problemas referentes aos fenômenos multilíngue (p.ex. agramaticalidade) e multidocumento (p.ex. redundância e contradição).

5.2 Procedimento de avaliação da informatividade

Para a realização da avaliação automática da informatividade, utilizou-se o pacote de medidas ROUGE (LIN; HOVY, 2003). Neste trabalho, utilizaram-se duas medidas: (i) ROUGE-1, que calcula a informatividade pela sobreposição de unigramas entre o sumário automático e o de referência, e (ii) ROUGE-2, que se baseiam na sobreposição de bigramas entre o sumário automático e o de referência. Essa escolha pautou-se no fato de que a ocorrência de unigramas e bigramas são mais frequentes nas línguas. Em especial, calcularam-se a ROUGE-1 e a ROUGE-2 dos Métodos 1 e 2 propostos neste trabalho. Em seguida, apresentam-se os resultados obtidos.

5.2.1 Resultados da avaliação de informatividade

Os resultados da ROUGE para o Método 1 e Método 2 são apresentados nas Tabelas 9 e 10, respectivamente.

Tabela 9 – Resultado da ROUGE: Método 1.

Coleção	ROUGE-1	ROUGE-1	ROUGE-1	ROUGE-2	ROUGE-2	ROUGE-2
	Cobertura	**Precisão**	**Medida-F**	**Cobertura**	**Precisão**	**Medida-F**
C1	0,27941	0,26027	0,2695	0,0963	0,08966	0,09286
C2	0,28846	0,29412	0,29126	0,09804	0,1	0,09901
C3	0,40909	0,37762	0,39273	0,09924	0,09155	0,09524
C4	0,35849	0,4086	0,38191	0,19048	0,21739	0,20305
C5	0,45763	0,40909	0,432	0,30769	0,27481	0,29032

C6	0,40625	0,45614	0,42975	0,28571	0,32143	0,30252
C7	0,34	0,33663	0,33831	0,12121	0,12	0,1206
C8	0,35417	0,32381	0,33831	0,16842	0,15385	0,16081
C9	0,34677	0,34959	0,34817	0,13821	0,13934	0,13877
C10	0,36552	0,35333	0,35932	0,17361	0,16779	0,17065
C11	0,54945	0,52083	0,53476	0,3	0,28421	0,29189
C12	0,5	0,56044	0,5285	0,30693	0,34444	0,3246
C13	0,25217	0,21481	0,232	0,0614	0,05224	0,05645
C14	0,4	0,38596	0,39285	0,2963	0,28571	0,29091
C15	0,33835	0,31034	0,32374	0,08333	0,07639	0,07971
C16	0,28235	0,26966	0,27586	0,05952	0,05682	0,05814
C17	0,48077	0,48077	0,48077	0,29126	0,29126	0,29126
C18	0,24793	0,22388	0,23529	0,03333	0,03008	0,03162
C19	0,37647	0,38095	0,3787	0,10714	0,10843	0,10778
C20	0,43956	0,48193	0,45977	0,27778	0,30488	0,2907
TOTAL	**0,373642**	**0,3699385**	**0,371175**	**0,174795**	**0,175514**	**0,1748445**

Tabela 10 – Resultados da ROUGE: Método 2.

Coleção	ROUGE-1	ROUGE-1	ROUGE-1	ROUGE-2	ROUGE-2	ROUGE-2
	Cobertura	**Precisão**	**Medida-F**	**Cobertura**	**Precisão**	**Medida-F**
C1	0,28676	0,24528	0,2644	0,11111	0,09494	0,10239
C2	0,28846	0,29412	0,29126	0,09804	0,1	0,09901
C3	0,37879	0,34722	0,36232	0,09924	0,09091	0,09489
C4	0,39623	0,40385	0,4	0,19048	0,19417	0,19231
C5	0,47458	0,44444	0,45902	0,30769	0,288	0,29752
C6	0,4375	0,38889	0,41177	0,28571	0,25352	0,26865
C7	0,26	0,22034	0,23853	0,10101	0,08547	0,09259
C8	0,4375	0,38182	0,40777	0,2	0,17431	0,18627
C9	0,33065	0,33065	0,33065	0,08943	0,08943	0,08943
C10	0,36552	0,32317	0,34304	0,20139	0,17791	0,18892
C11	0,41758	0,35185	0,38191	0,16667	0,14019	0,15229
C12	0,5	0,56044	0,5285	0,30693	0,34444	0,3246
C13	0,23478	0,19014	0,21012	0,02632	0,02128	0,02353
C14	0,32727	0,29032	0,30769	0,24074	0,21311	0,22608
C15	0,2406	0,22378	0,23189	0,03788	0,03521	0,0365
C16	0,23529	0,20833	0,22099	0,04762	0,04211	0,0447
C17	0,47115	0,39837	0,43171	0,2233	0,18852	0,20444
C18	0,26446	0,25	0,25703	0,01667	0,01575	0,0162
C19	0,37647	0,35955	0,36782	0,09524	0,09091	0,09302
C20	0,3956	0,36364	0,37895	0,26667	0,2449	0,25532
TOTAL	**0,3559595**	**0,32881**	**0,3412685**	**0,155607**	**0,144254**	**0,149433**

A partir da comparação do desempenho dos Métodos 1 e 2 para os resultados da ROUGE 1, percebe-se que o Método 1 obtém médias levemente maiores do que o

Método 2 em todos os resultados, tanto na medida de cobertura (0,37 contra 0,35) quanto de precisão (0,36 contra 0,32). Com relação à média, realizada pela medida-f, o Método 1 obteve um desempenho de 0,37, enquanto o Método 2 de 0,34.

Na ROUGE 2, o Método 1 também apresenta médias maiores do que o Método 2 em todas as medidas analisadas. Na medida de cobertura, o Método 1 apresenta uma pontuação de 0,17 contra 0,15 do Método 2. Quanto à medida de precisão, o Método 2 tem a pontuação de 0,17, contra 0,14 do Método 2. A medida-f do Método 1, por sua vez, obteve o valor de 1,17, enquanto o Método 2 obteve um valor aproximado de 0,15.

Na Tabela 11 pode-se visualizar, com mais clareza, a comparação entre os resultados da ROUGE dos dois métodos.

Tabela 11 – Comparação entre os resultados da ROUGE dos métodos 1 e 2.

MÉDIA	ROUGE-1	ROUGE-1	ROUGE-1	ROUGE-2	ROUGE-2	ROUGE-2
	Cobertura	**Precisão**	**Medida-F**	**Cobertura**	**Precisão**	**Medida-F**
MÉTODO 1	0,373642	0,3699385	0,371175	0,174795	0,175514	0,1748445
MÉTODO 2	0,3559595	0,32881	0,3412685	0,155607	0,144254	0,149433

Observa-se pela Tabela 11 que o Método 1 apresenta, para os resultados da avaliação de informatividade, maiores pontuações em todos os quesitos investigados, a saber: cobertura, precisão e medida-f, em ambas as medidas (ROUGE-1 e 2).

Ressalta-se que os resultados mencionados das médias para cobertura, precisão e medida-f foram submetidos ao teste estatístico de Wilconoxon para amostras pareadas, e os resultados mostraram 95% de confiança, o que comprova que há, realmente, diferença estatisticamente significativa entre o Método 1 e o Método 2.

Em suma, os resultados dizem que o Método 1 supera o Método 2 também no quesito informatividade. Uma hipótese provável para tal resultado pode ser pautada pelo fato dos sumários de referência terem sido criados por falantes do português, apesar da construção dos sumários de referência ter sido baseada na leitura de coleções bilíngues. Nesse sentido, a língua nativa dos elaboradores dos sumários de referência pode ter influenciado nas escolhas lexicais destes a partir dos textos-fonte em português.

6 CONSIDERAÇÕES FINAIS

A pesquisa ora apresentada aplicou o conhecimento léxico-conceitual na tarefa de SAMM, evitando-se assim a tradução integral dos textos-fonte e contribuindo para o avanço da SAMM envolvendo o português, que até então se resumia a um único trabalho, no qual foram investigados métodos *baseline* (TOSTA et al., 2013).

6.1 Verificação das hipóteses

Salienta-se que, a realização deste trabalho confirma suas hipóteses "fracas" (1 e 2) e as "fortes" (3 a 6).

A hipótese 1, de que é possível realizar a SAMM sem traduzir integralmente os textos-fonte ou os sumários com base em conhecimento léxico-conceitual, foi confirmada, uma vez que a explicitação do significado lexical de cada unidade lexical por meio de um rol único de conceitos permite criar uma espécie de interlíngua, pela qual unidades lexicais de línguas diferentes são codificadas em um mesmo *synset*, que representa determinado conceito, sem que haja a necessidade de tradução prévia e integral dos textos-fonte.

A hipótese 2, de que o conhecimento léxico-conceitual reflete os fenômenos multidocumento (redundância, complementariedade e contradição), foi confirmada, na medida em que o cálculo de *synsets* mais frequentes de cada coleção foi utilizado como critério de pontuação das sentenças, sendo realizado a partir do fenômeno da redundância entre os conceitos presentes nos textos-fonte. O fato dos sumários gerados terem sido informativos em relação aos textos-fonte comprova que a aplicação desse tipo de conhecimento é eficaz para mapear a redundância.

A hipótese 3, de que a ocorrência de conceitos em múltiplos textos reflete a informação principal, foi também confirmada, pois ambos os métodos obtiveram bons resultados na avaliação da informatividade.

A hipótese 4, de que as divergências lexicais entre as línguas não teriam impacto na SAMM devido à representação conceitual, obteve, por sua vez, confirmação. No caso, essa hipótese se confirmou porque os *synsets* da WN.Pr funcionaram como uma espécie de interlíngua, permitindo resolver divergências lexicais entre as línguas, o que, consequentemente, possibilitou que tais divergências não impactassam a SAMM.

A hipótese 5, de que um sumário composto exclusivamente por sentenças originais em português refletiria as informações mais relevantes da coleção, também foi confirmada. Isso se deve ao fato de que tais sentenças, mesmo em uma única língua, foram selecionadas a partir de um ranque construído em função da ocorrência dos conceitos nos textos em português e inglês. A confirmação dessa hipótese se verifica no fato de que a informatividade dos sumários gerados pelo Método 1 (que é composto somente por sentenças do texto-fonte em português) é em média mais alta do que a informatividade dos sumários gerados pelo Método 2.

Nos resultados da ROUGE, o método 1 gerou sumários mais informativos do que o Método 2, o que comprova que é possível capturar as informações mais relevantes da coleção sem incluir nos sumários sentenças proveniente do texto em inglês, posto que, no levantamento dos conceitos mais frequentes da coleção, os textos-fonte de ambas as línguas foram considerados.

Por fim, a hipótese 6, segundo a qual um sumário composto por sentenças originais em português e por sentenças traduzidas individualmente para o português apresenta menos problemas de TA do que um sumário produzido a partir de uma coleção composta por textos integralmente traduzidos para o português e originais nessa mesma língua, também foi confirmada. Especificamente, essa confirmação pode ser verificada na comparação da qualidade linguística dos sumários do Método 2 frente à qualidade linguística dos sumários gerados pelo melhor método de Tosta et al., (2012), (2013).

O resultado mostra que um sumário composto por sentenças originais em português e por sentenças traduzidas individualmente para o português (Método 2) apresenta menos problemas de TA do que um sumário produzido a partir de uma coleção composta por textos integralmente traduzidos para o português e originais nessa mesma língua (abordagem *early translation* de Tosta et al., (2012)).

Essa afirmação pode ser evidenciada principalmente pela gramaticalidade dos sumários produzidos pelo Método 2 deste trabalho, que se mostrou superior em comparação à gramaticalidade de Tosta et al. (2012, 2013). Essa confirmação não é surpreendente, visto que, quanto maior a quantidade de texto a ser traduzido de uma só vez, maior a probabilidade de ocorrência de erros de TA (WAN; LI; XIAO, 2010). Consequentemente, o método cuja abordagem realiza a tradução de sentenças individuais em detrimento do texto integral obteve melhores resultados.

Diante dos experimentos realizados, discutem-se na próxima Seção algumas contribuições e limitação deste trabalho e, sobretudo, propõem-se alguns trabalhos futuros relacionados a esta pesquisa.

6.2 Contribuições

A primeira contribuição deste trabalho refere-se à construção do CM2News, *corpus* bilíngue com 20 coleções, cada uma delas composta por 1 notícia em português e 1 em inglês sobre o mesmo assunto.

A segunda contribuição consiste na anotação semântica dos nomes comuns do CM2News, o qual, uma vez enriquecido com essa anotação e outras, pode se tornar referência para as pesquisas sobre SAMM, especialmente às que envolvem o português.

A terceira contribuição advém da elaboração de sumários de referência para todas as coleções do CM2News.

A realização deste trabalho possibilitou a elaboração de um manual sistematizado de regras gerais e específicas para anotação léxico-conceitual, que poderá ser utilizado em outros trabalhos relacionados a este.

Outra contribuição foi o desenvolvimento do editor MulSen, que permite agilizar o processo de anotação léxico-conceitual no contexto multilíngue inglês-português.

Por fim, salientam-se os resultados promissores para a aplicação de conhecimento léxico-conceitual na SAMM e SAM, bem como sua aplicação para driblar problemas de TA no contexto da SAMM.

6.3 Limitação

Reconhece-se como uma limitação deste trabalho a criação dos sumários de referência somente por linguistas computacionais que são falantes nativos do português. Uma vez falantes do português, tais pesquisadores podem ter produzido sumários de referência fortemente embasados em material linguístico advindos dos textos-fonte em português, o que pode oferecer um impacto na avaliação automática com base na ROUGE, já que esta se baseia na sobreposição de unidades lexicais entre os sumários de referência e os automáticos.

No caso, os bons resutados obtidos pelos sumários automáticos compostos apenas por sentenças em português podem ter sido influenciados, em parte, pelo fato de os sumários de referência serem compostos por material linguístico advindo preferencialmente dos textos-fonte em português.

6.4 Trabalhos futuros

Tendo em vista a exploração da SAMM com a aplicação de conhecimento léxico-conceitual envolvendo o português, sugerem-se os seguintes trabalhos futuros:

- Anotar palavras das classes dos adjetivos e verbos, visto que também veiculam conteúdo textual relevante.
- Refinar a tarefa de avaliação, considerando a elaboração de sumários automáticos e de referência com taxas de compressão diferentes das consideradas neste trabalho, sob a hipótese de que sumários menores possuem menos problemas linguísticos.
- Considerar uma pontuação especial para *synsets* hipônimos/hiperônimos, visto que, neste trabalho, a pontuação da frequência/peso dos *sysnsets* mais importantes considera apenas a sobreposição de *synsets* idênticos. Nesse sentido, os conceitos das unidades lexicais "cão" e "animal", por exemplo, receberiam uma pontuação específica por estabelecerem relação de hiponímia/hiperonímia, também importante para a informatividade.
- Tratar a redundância a partir do *overlap* de conceitos, visto que, nesta investigação, a redundância entre as sentenças foi calculada a partir do *overlap* de lemas de palavras de classe aberta. Tal proposta apoia-se na hipótese de que o *overlap* de conceitos pode refinar a tarefa, uma vez que algumas unidades lexicais com lemas distintos podem apresentar o mesmo conceito e esse tipo de redundância não é capturada por *overlap* de lemas.
- Analisar sistematicamente os problemas de TA mais recorrentes dos sumários gerados pelo Método 2 e identificar, dentre eles, quais tiveram maior impacto quanto à avaliação de qualidade linguística dos sumários.
- Realizar a construção de sumários de referência com a participação de juízes falantes nativos do inglês e proficientes em português. Tal procedimento apoia-

se na hipótese de que a língua nativa dos juízes (criadores dos sumários de referência) pode interferir nos resultados da avaliação de informatividade.

REFERÊNCIAS

AFANTENOS, S. D.; DOURA, I.; KAPELLOU, E.; KARKALETSIS, V. Exploiting cross-document relations for multi-document evolving summarization. In: METHODS AND APPLICATIONS OF ARTIFICIAL INTELLIGENCE/ HELENIC CONFERENCE ON AI, 3., 2004, Samos, Greece. **Proceedings…** Samos, 2004. p. 410-419.

AFANTENOS, S.D.; KARKALETSIS, V.; STAMATOPOULOS, P.; HALATSIS, C. Using synchronic and diachronic relations for summarizing multiple documents describing evolving events. **Journal of Intelligent Information Systems**, v. 30, n. 3, p. 183-226, 2008.

AGIRRE, E.; EDMONDS, P.G. **Word sense disambiguation: Algorithms and applications**. Springer Science-Business Media, 2006.

AKABANE, A.T.; PARDO, T.A.S.; RINO, L.H.M. Explorando medidas de redes complexas para sumarização multidocumento. In:STIL STUDENT WORKSHOP ON INFORMATION AND HUMAN LANGUAGE TECHNOLOGY, 2,2011, Cuiabá/MT. **Proceedings…**Cuiabá, 2011, p.1-3.

AKABANE, A.T.; PARDO, T.A.S.; RINO, L.H.M. Explorando medidas de redes complexas para sumarização multidocumento. In: STIL STUDENT WORKSHOP ON INFORMATION AND HUMAN LANGUAGE TECHNOLOGY, 2., 2011, Cuiabá. **Proceedings…** Cuiabá: UFMT, 2011. p. 1-3.

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

BARZILAY, R.; MCKEOWN, K.; ELHADAD, M. Information fusion in the context of multi-document summarization. In:ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 37., 1999, Maryland. **Proceedings…** Maryland, 1999. p. 550-557.

BAXENDALE, P. B. Machine-made index for technical literature-an experiment. **IBM Journal of Research and Development**, no. 2, p.354-361, 1958.

BENTIVOGLI, L.; PIANTA, E. Beyond lexical units: enriching wordNets with *phraset*s. In: EUROPEAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 3, 2003, Budapest, Hungary. **Proceedings...** Budapest, 2003. p. 67-70.

BOSSARD, A.; RODRIGUES, C. Combining a Multi-Document Update Summarization System –CBSEAS– with a Genetic Algorithm. **Systems and Technologies**, v.8, p. 71-87, 2011.

BOURDIN, F. A.; HUET, S.; TORRES-MORENO, JUAN-MANUEL. Graph-based approach to cross-language multi-document summarization. In: CONFERENCE ON

INTELLIGENT TEXT PROCESSING AND COMPUTATIONAL LINGUISTICS, 12., 2011, Tokyo, Japan. **Proceedings…** Tokyo: CICLing , 2011. p.113-118.

CARBONELL, J.; GOLDSTEIN, J. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 21., 1998, Melbourne. **Proceedings…** Melbourne, 1998. p. 335-336.

CARDOSO, P.C.F.; MAZIERO, E.G.; CASTRO JORGE, M.L.R.; SENO, E.M.R.; DI-FELIPPO, A.; RINO, L.H.M.; NUNES, M.G.V.; PARDO, T.A.S. CSTNews - A Discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese. In: RST BRAZILIAN MEETING, 3., 2011, Cuiabá. **Proceedings…** Cuiabá: UFMT, 2011. p. 88-105.

CASTRO JORGE, M.L., PARDO, T.A.S. A Generative approach for multi-document summarization using the noisy channel model. In: RST BRAZILIAN MEETING, 3., 2011, Cuiabá. **Proceedings**… Cuiabá: UFMT, 2011. p. 75-87.

CASTRO JORGE, M.L.R.; PARDO, T.A.S. Experiments with CST-based Multi-document summarization. In:ACL WORKSHOP TEXTGRAPHS-5: GRAPH-BASED METHODS FOR NATURAL LANGUAGE PROCESSING, 2010, Uppsala/Sweden. **Proceedings…**Uppsala, 2010, p. 74-82.

CELIKYILMAZ, A.; HAKKANI-TUR, D. Discovery of topically coherent sentences for extractive summarization. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 49., 2011, Portland. **Proceedings…** Portland, 2011. p. 491-499.

CLARKE, J.; LAPATA, M. Discourse constraints for document compression. **Computational Linguistics,** v. 36, n. 3, p. 411-441, 2010.

CONROY, J. M.; O'LEARY, D. P. Text summarization via Hidden Markov models. In: ANNUAL INTERNATIONAL ACM SIGIR, 24, 2001, New York/ USA. **Proceedings…** New York, 2001. p. 406-407.

COWIE, J., MAHESH, K., NIRENBURG, S., AND ZAJAZ, R. MINDS - multilingual interactive document summarization.In: AAAI SPRING SYMPOSIUM ON INTELLIGENT TEXT SUMMARIZATION, 1998, Menlo Park, CA. **Proceedings…**, Menlo, 1998. p.131-132.

CREMMINS, E.T. **The art of abstracting**. Arlington, Virginia: Information Resources Press, 1996.

CRUSE, D. **Lexical semantics.** Cambridge: Cambridge University Press, 1986.

DI-FELIPPO, A. **Delimitação e alinhamento de conceitos lexicalizados no Inglês Norte-americano e no Português Brasileiro**. 2008. Tese (Doutorado em linguística) - Faculdade de Ciências e Letras, Universidade Estadual Paulista/Unesp,Araraquara, 1998.

DIAS-DA-SILVA, B. C. **A face tecnológica dos estudos da linguagem**: o processamento automático das línguas naturais.1996.Tese (Doutorado em Letras) - Faculdade de Ciências e Letras, Universidade Estadual Paulista/Unesp, Araraquara, 1996. 272p.

DOLZ, J.; SCHNEUWLY, B. **Gêneros orais e escritos na escola**. Campinas, SP: Mercado de Letras, 2004. 278 p.

DOS SANTOS AZEVEDO, Francisco Ferreira. **Dicionário analógico da língua portuguesa:(idéias afins)**. Coordenada Editora, 1974.

EDMUNDSON, H. P. New Methods in automatic extracting. **Journal of the ACM**, Vol. 16, p. 264-285, 1969.

ENDRES-NIGGEMEYER, B. **Summarization Information**. Berlin: Springer, 1998.

EVANS, D.K.; KLAVANS, J.L.; MCKEOWN, K.R. Columbia NewsBlaster: multilingual news summarization on the web.In: NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTACIONAL LINGUISTICS: HUMAN LANGUAGE TECHNOLOGIES, 2004, Boston. **Proceedings…** Boston, 2004. p.1-4.

EVANS, D.K.; KLAVANS, J.L.; MCKEOWN, K.R. **Similarity-based multilingual multi-document summarization**.New York: Columbia University, 2005. (Technical Report CUCS-014-05).

FELLBAUM, C (Ed.). **Wordnet: an electronic lexical database** (Language, speech and communication). Massachusetts: **MIT Press**, 1998.

FUNG, P.; NGAI, G. One story, one flow: Hidden markov models for multilingual multidocument summarization. **ACM Transactions on Speech and Language Processing**, v.3, n.2, p-1-16, 2006.

GANTZ, J.; REINSEL, D. **Extracting Value from Chaos**. International Data Corporation iView, 2011.

GEY, F. C.; et al. New directions in multilingual information acces. **SIGIR Workshop Report**. Vol 40, No2, Dez, 2006.

GRUBER, T. Toward principles for the design of ontologies used for knowledge sharing. **International Journal Human-Computer Studies,** v. 43, n. 5-6, p. 907-928, 1995.

GUPTA, V; LEHAL, G. S. A survey of text summarization extractive techniques. **Journal of Emerging Technologies in Web Intelligence**, v. 2, n. 3, p. 258-268, 2010.

HAGHIGHI, A.; VANDERWENDEV, L. Exploring content models for multi-document summarization. In: HUMAN LANGUAGE TECHNOLOGIES: THE 2009 ANNUAL CONFERENCE OF NORTH AMERICAN CHAPTER OF THE

ASSOCIATION FOR COMPUTACIONAL LINGUISTICS, 2009, Boulder, Colorado. **Proceedings…** Boulder, 2009. p.362-370.

HATZIVASSILOGLOU, J. L.; KLAVANS J.L.; HOLCOMBE, M. Simfinder: a flexible clustering tool for summarization. In: NAACL AUTOMATIC SUMMARIZATION WORKSHOP, 2001. Pittsburgh, PA, USA. **Proceedings**… Pittsburgh, 2001. p.9.

HENNIG, L., UMBRATH, W., WETZKER, R. An ontology-based approach to text summarization. In: WORKSHOP ON NATURAL LANGUAGE PROCESSING AND ONTOLOGY ENGINEERING (NLPOE 2008), 3, Toronto, 2008. **Proceedings…**Toronto, Canada, 2008. p. 291-294.

JURAFSKY, D; MARTIN, J. **Speech and Language Processing**: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. New Jersey: Prentice Hall, 2007. 1024p.

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

LAGE, N. **A reportagem**: teoria e técnica de entrevista e pesquisa jornalística. Rio de Janeiro: Record, 2004.

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

LESK, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In: The 5th Annual International Conference on Systems Documentation, *1986*, New York, NY, USA. **Proceedings…** New York, NY, 1986, p. 24–26.

LI, L., WANG, D., SHEN, C., LI, T. Ontology-enriched multi-document summarization in disaster management. In: ACM SPECIAL INTEREST GROUP ON INFORMATION RETRIEVAL (SIGIR), 2010, Geneva. **Proceedings…** Geneva, Switzerland, 2010. p. 819-820.

LIN, C-Y.; HOVY, E.H. Automatic evaluation of summaries using n-gram cooccurrence statistics. In: THE 2003 CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS ON HUMAN LANGUAGE TECHNOLOG, 2003, Edmonton, Canada. **Proceedings…** Edmonton, 2003.p.71-78.

LIN, C.; HOVY, E.H**.** From Single to Multi-document Summarization:A prototype system and its evaluation. In: ANNIVERSARY MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (ACL-02), 40., 2002, Philadelphia, Pennsylvania. **Proceedings…** Philadelphia, 2002. p. 7-12.

LITVAK, M.; LAST, M.; FRIEDMAN, M. A New approach to improving multilingual summarization using a genetic algorithm.In: THE ANNUAL MEETING OF THE

ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 48., 2010, Stroudsburg, PA, USA**. Proceedings…**Stroudsburg, 2010. p. 927-936.

LITVAK, M.; LIPMAN, H.; GUR, A. B.; e LAST, M. Towards multi-lingual summarization: a comparative analysis of sentence extraction methods on English and Hebrew *corpora*. In: INTERNATIONAL WORKSHOP ON CROSS LINGUAL INFORMATION ACCESS/COLING, 4, 2010, Beijing, China. **Proceedings**... Beijing, 2010.p.61-69.

LOUIS, A; NENKOVA, A. Automatically Assessing Machine Summary Content Without a Gold Standard. **Computational Linguistic,** v. 39, p. 267-300, 2013.

LOUIS, A.; JOSHI, A.; NENKOVA, A. Discourse indicators for content selection in summarization. In: ANNUAL MEETING OF THE SPECIAL INTEREST GROUP ON DISCOURSE AND DIALOGUE, 11, 2010. **Proceedings…** 2010. p. 147–156.

LUHN, H. P. The automatic creation of literature abstracts. **IBM Journal of Research and Development***,* v. 2, p. 159-165, 1958.

LYONS, J. **Introdução à linguística geral.** Supervisão de tradução de Isaac Nicolau Salum. São Paulo: Editora Nacional/Edusp, 1979.

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

MANI, I.; BLOEDORN, E. Multi-document summarization by graph search and matching. In: NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE (AAAI), 14., 1997, Rhode Island. **Proceedings…** Rhode Island, 1997, p. 622-628.

MANI, I.; MAYBURY, M. T.**Advances in automatic text summarization**. Massachusetts: The MIT Press, 1999.

MANN, W.C.; THOMPSON, S.A. **Rhetorical Structure Theory**: a theory of text organization. 1987. (Technical Report ISI/RS-87-190).

MARCU, D. **The theory and practice of discourse Parsing and Summarization**. Cambridge, Massachusetts: The MIT Press, 2000.

MARTINS, D. B. J.; Caseli, H. M. (2013).**Anotação manual de erros de tradução automática em textos traduzidos de inglês para português do Brasil.** São Carlos:ICMC-USP,2013. 24 p.(Série de Relatórios do NILC. NILC-TR-13-02).

MAZIERO, E. G.; JORGE, M. L. C.; PARDO, T. A. S. Identifying multi-document relations. In: INTERNATIONAL WORKSHOP ON NATURAL LANGUAGE PROCESSING AND COGNITIVE SCIENCE**,** 7., 2010, Funchal, Madeira. Proceedings... Funchal, 2010. p. 60-69.

MAZIERO, E.G.**Identificação automática de relações multidocumento**.2012. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação (ICMC) - Universidade de São

Paulo, São Carlos, 2012.

MCKEOWN, K., RADEV, D.R. Generating summaries of multiple news articles. In: ANNUAL INTERNATIONAL ACM-SIGIR, 18, 1995. **Proceedings…**1995. p. 74-82.

MCKEOWN, K.; KLAVANS, J.; HATZIVASSILOGLOU, V.; BARZILAY,R; ESKIN, E. Towards multi-document summarization by reformulation: Progress and prospects.. In: NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE, 16., 1999, Florida.**Proceedings...**1999. p. 453–460.

MCKEOWN, K.; PASSONNEAU, R.; ELSON, D.; NENKOVA, A.; HIRSCHBERG J. Do summaries help? A task-based evaluation of multi-document summarization. In: ANNUAL INTERNATIONAL ACM-SIGIR, 28, 2005, Salvador. **Proceedings…** Salvador, 2005. p. 210-217.

MIHALCEA, R.; TARAU, P. An algorithm for language independent single and multiple document Summarization. In: INTERNATIONAL JOINT CONFERENCE ON NATURAL LANGUAGE PROCESSING (IJCNLP), 2005, Korea. **Proceedings…** Korea, 2005. p.19-21.

MILLER, G. A., FELLBAUM, C. Semantic networks of English. **Cognition.** v. 41. p. 197 – 229, 1991.

NENKOVA, A. Discourse Factors in multi-document summarization. In: ANNUAL AAAI/SIGART DOCTORAL CONSORTIUM, 10., 2005a, Pittsburgh. **Proceedings…** Pittsburgh, 2005. p. 1654-1655.

NILES, I.; PEASE, A. Origins of the IEEE standard upper ontology. In: WORKSHOP ON THE IEEE STANDARD UPPER ONTOLOGY –IJCAI, 2001, Seattle, Washington. **Proceedings...** Seattle, Washington, 2001. p.4-10.

NÓBREGA, F. A. A. E PARDO, T. A. S. (2012). Explorando métodos de desambiguação lexical de sentidos de uso geral para o português. In: Encontro Nacional de Inteligência Artificial, 2012, Curitiba, Paraná. **Proceedings…** Curitiba, 2012.

NÓBREGA, F. A. A. E PARDO, T. A. S. (2012). Explorando Métodos de Uso Geral para Desambiguação Lexical de Sentidos para a Língua Portuguesa. In: the 9th Brazilian Symposium in Information and Human Language Technology, 2013, Fortaleza. **Proceedings…** Fortaleza, 2013. p. 138-147.

NÓBREGA, F.A.A. **Desambiguação lexical de sentidos para o português por meio de uma abordagem multilíngue mono e multidocumento.** 2013.Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e da Computação (ICMC), Universidade de São Paulo, São Carlos, 2013.

O'DONNELL, M. Variable-length on-line document generation. In: EUROPEAN WORKSHOP ON NATURAL LANGUAGE GENERATION, 6., 1997a, Duisburg. **Proceedings…** Duisburg, 1997.

ORĂSAN, C.; CHIOREAN, O.A. Evaluation of a cross-lingual Romanian-English multi-document summariser.In: LANGUAGE RESOURCES AND EVALUATION CONFERENCE, 6., 2008, Marrakesh.**Proceedings…** Marrakesh, Morocco, 2008. p.6.

ORĂSAN, C. Automatic summarization in the informational age. In: INTERNATIONAL CONFERENCE ON RECENT ADVANCES IN NATURAL LANGUAGE PROCESSING, 7., 2009, Borovets,**Proceedings...** Borovets, Bulgaria: Association on Computational Linguistics, 2009.

OTTERBACHER, J.C.; RADEV, D.R.; LUO, A. Revisions that improve cohesion in multi-document summaries: a preliminary study. In WORKSHOP ON AUTOMATIC SUMMARIZATION, 2002, Stroudsburg, PA. **Proceedings…** Stroudsburg, 2002, p 27-36.

PARDO, T.A.S. **GistSumm** – GIST SUMMarizer: extensões e novas funcionalidades. São Carlos:ICMC-USP,2005. 8p. (Série de Relatórios do NILC. NILC-TR-05-05).

PARDO, T.A.S., RINO, L.H.M. DMSumm: review and assessment. In: RANCHHOD, E., MAMEDE (Eds.). **Advances in Natural Language Processing** (Lecture Notes in Artificial Intelligence 2389). Germany: Springer-Verlag, 2002. p. 263-273.

PARDO, T.A.S.; ALEIXO, P. **CSTNews**: um córpus de textos jornalísticos anotados segundo a teoria discursiva multidocumento CST (*Cross-document Structure Theory*).São Carlos: NILC-ICMC, 2008. (Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional).

PARDO, T. A. S.; RINO, L. H. M., NUNES, M. G. V. GistSumm: a summarization tool based on a new extractive method. In: MAMEDE, N.J., BAPTISTA, J., TRANCOSO, I., NUNES, M.G.V. (Eds.). **Workshop on Computational Processing of the Portuguese Language- Written and Spoken** (PROPOR) (Lecture Notes in Artificial Intelligence 2721), Faro/Portugal, 2003. p. 210-218.

PITLER, E.; LOUIS, A.; NENKOVA, A. Automatic evaluation of linguistic quality in multi-document summarization. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (ACL), 48., 2010, Uppsala,Sweden.**Proceedings...**Uppsala, 2010. p. 544-554.

RADEV, D. ; ALLISON T. ; BLAIR-GOLDENSOHN, S. ; BLITZER, J. ; CELEBI, A.; DIMITROV, S.; DRABEK, E.; HAKIM, A. ; LAM, W.; , LIU, D.; OTTERBACHER, J.; QI, H.; SAGGION, H.; TEUFEL S.; TOPPER, M; , WINKEL, A.; ZHANG, Z.MEAD - a platform for multi-document multilingual text summarization. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC),4., 2004, Lisbon, Portugal.**Proceedings…** Lisbon, 2004.

RADEV, D.R. A common theory of information fusion from multiple text sources, step one: Cross-document structure. In:ACL SIGDIAL WORKSHOP ON DISCOURSE AND DIALOGUE, 1, 2000, Hong Kong. **Proceedings…**Hong Kong, 2000. p. 74-83.

RADEV, D.R.; MCKEOWN, K. Generating natural language summaries from multiple on-line sources. **Computational Linguistics**, v. 24, n. 3, p. 469-500, 1998.

RATNAPARKHI, A. A maximum entropy model for part-of-speech tagging. In:CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 1996,Philadelphia, PA. **Proceedings**… Philadelphia, 1996. p. 133-142.

RIBALDO, R.; RINO, L.H.M.; PARDO, T.A.S. Graph-based Methods for Multi-document Summarization: Exploring Relationship Maps, Complex Networks and Discourse Information. In:INTERNATIONAL CONFERENCE ON COMPUTATIONAL PROCESSING OF PORTUGUESE, 10., 2012, Coimbra. **Proceedings...**Coimbra: Universidade de Coimbra, 2012. p. 260-271.

RINO, L. H. M.; PARDO, T. A. S.; SILLA Jr., C. N.; KAESTNER, C. A., POMBO, M. A comparison of automatic summarization systems for Brazilian Portuguese texts. In: BRAZILIAN SYMPOSIUM ON ARTIFICIAL INTELLIGENCE (SBIA) (Lecture Notes in Artificial Intelligence 3171), 17, 2004, São Luis. **Proceedings…** São Luis, 2004. p. 235-244.

SAGGION, H. Multilingual multi-document summarization tools and evaluation. In: INTERNATIONAL CONFERECE ON LANGUAGE RESOURCES AND EVALUATION (LREC), 4., 2006.**Proceedings**...Lisbon, 2006. p. 1312-1317.

SAGGION, H. SAGGION, H.; RADEV, D.; TEUFEL, S.; LAM, W.; STRASSEL, S.M. Developing Infrastructure for evaluation of single and multi-document summarization systems in a cross-lingual environment. In:INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 3., 2002, Canary Islands, Spain.**Proceedings**… Las Palmas, 2002. p.747-754.

SAGGION, H.; LAPALME, G. Concept identification and presentation in the context of technical text summarization. In: NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTACIONAL LINGUISTICS- WORKSHOP ON APPLIED INTERLINGUAS: PRACTICAL APPLICATIONS OF INTERLINGUAL APPROACHES TO NLP, 2.,2000. **Proceeedings...** Stroudsburg, Pennsylvania, 2000. p.1-10.

SALTON, G.; SINGHAL, A.; MITRA, M.; BUCKLEY C. Automatic text structuring and summarization. **Information Processing & Management**, v. 33, n. 2, p. 193-207, 1997.

SCHIFFMAN, B.; NENKOVA, A.; MCKEOWN, A. Experiments in multi-document summarization. In: INTERNATIONAL CONFERENCE ON HUMAN LANGUAGE TECHNOLOGY RESEARCH, 2., 2002, San Francisco, CA, USA. **Proceedings**… San Francisco, 2002.p.52-58.

SCHILDER, F.; KONDADADI, R. FastSum: fast and accurate query-based multi-document summarization. In: MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (ACL), 46, 2008, Columbus, Ohio. **Proceedings…**Columbus, 2008. p. 205–208.

SCHMID, H. Probablistic part-of-speech tagging using decision trees. In: INTERNATIONAL CONFERENCE ON NEW METHODS IN LANGUAGE PROCESSING, Manchester, UK.**Proceedings…** Manchester, 1994. p. 44-49.

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.

SPARCK JONES, K. **Automatic summarising**: a review and discussion of the state of the art. Cambridge: University of Cambridge, 2007. (Technical Report UCAM-CL-TR-679).

SPARCK JONES, K. Discourse modeling for automatic summarisation. **Tech. Report No. 290**. University of Cambridge. UK, February, 1993.

SPARCK JONES, K. **Discourse modeling for Automatic Summarisation**.Cambridge: University of Cambridge-UK,1993. (Technical Report No. 290).

SPARCK-JONES, K. Automatic summarizing: factors and directions. In:MANI, I.; MAYBURY, M. T.(Eds.). **Advances in automatic text summarization**. Massachusetts:MIT Press, 1999. p.1-14.

SPARCK-JONES, K.; GALLIERS, J.R. **Evaluating natural language processing systems**: an analysis and review. Spring-Verlag HeidelBerlag, 1996.

STEINBERGER, J.; TURCHI, M. Machine translation for multilingual summary content evaluation. In: WORKSHOP ON EVALUATION METRICS AND SYSTEM COMPARISON FOR AUTOMATIC SUMMARIZATION- ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 2012. Jeju Island. **Proceeding**s... Jeju Island, Korea, 2012, p.19-27.

SUANMALI, L.; SALIM, N.; BINWAHLAN, M.S. Fuzzy genetic summarization based text summarization. In: INTERNATIONAL CONFERENCE ON DEPENDABLE AUTOMATIC AND SECURE COMPUTING, 9., 2001, Sydney.**Proceedings...** Sydney: IEEE, 2011. p. 1184-1191.

SVORE, K.; VANDERWENDE, L.; BURGES, C. Enhancing single-document summarization by combining RankNet and third-party sources. In:JOINT CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING AND COMPUTATIONAL NATURAL LANGUAGE LEARNING, 2007. **Proceedings…** 2007. p. 448-457.

TOSTA, F. E. S.; DI-FELIPPO, A.; PARDO, T.A. S. Estudo de métodos clássicos de sumarização automática no cenário multidocumento multilíngue.In: WORKSHOP DE INICIAÇÃO CIENTÍFICA EM TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (TILIC), 4., 2013, Fortaleza. **Proceedings...** Fortaleza**,** 2013. p.34-36.

TOSTA, F. E. S.; DI-FELIPPO, A.; PARDO, T.A. S. **Investigação de métodos clássicos de sumarização automática no cenário multidocumento multilíngue**:

primeiras aproximações. São Carlos:ICMC-USP, 2012. 18p. (Série de Relatórios do NILC. NILC-TR-12-02).

UZÊDA, V.R.; PARDO, T.A.S.; NUNES, M.G.V. A comprehensive comparative evaluation of RST-based summarization methods. **ACM Transactions on Speech and Language Processing**, v. 4, n. 6, p. 1-20, 2010.

VAN-HALTEREN, H.; TEUFEL, S**.** Examining the consensus between human summaries: initial experiments with factoid analysis. In: HLT-NAACL DUC WORKSHOP, 2003, Edmonton. **Proceedings**… Edmonton, 2003.p. 57-64.

VOSSEN, P. Introduction to EuroWordNet. **Computers and the Humanities,** v. 32, p. 73- 89, 1998.

WAN, X; LI, H.; XIAO,J. Cross-Language document summarization based onMachine translation quality prediction.In: ANNUAL MEETING OF ASSOCIATION FOR COMPUTACIONAL LINGUISTICS (ACL), 48., 2010, Uppsala, Sweeden. **Proceedings…** Uppsala, Sweeden, 2010.p. 917-926.

WAN, X. An exploration of document impact on graph-based multi-document summarization. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 2008, Honolulu. **Proceedings…** Honolulu, 2008. p. 755-762.

WANG, L.F. e YANG, C.C. The impact analysis of language differences on an automatic multilingual text summarization system. **Journal of the American Society for Information Science and Technology,**v.57, n.9, p.684-696, 2006.

WHITE, J.; DOYON, J.; TALBOTT, S. Task tolerance of MT output in integrated text processes. In: ANLP-NAACL 2000 WORKSHOP: EMBEDDED MT SYSTEMS WORKSHOP, 2000. Seattle, WA. **Proceedings…** Seattle, 2000. p. 9-16

WU, C-W.; LIU, C-L. Ontology-based Text Summarization for Business News Articles.**Computers and Their Applications**, v. 2003, p. 389-392, 2003.

ZHANG, Z.; BLAIR-GOLDENSOHN, S.; RADEV, D.R. Towards CST-enhanced summarization. In: AAAI CONFERENCE, 2002. Edmonton, Alberta. **Proceedings…** Edmonton, 2002. p. 439-445.

ZHANG, Z.; OTTERBACHER, J.; RADEV, D.R. Learning cross-document structural relationships using boosting. In: ACM CIKM, 2003, New Orleans, Louisiana. **Proceedings…** New Orleans, 2003. p. 124-130.