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

Aplicação de métodos clássicos de Sumarização Automática no contexto multidocumento multilíngue: primeiras aproximações

Fabricio Elder da Silva Tosta
Ariani Di Felippo
Thiago Alexandre Salgueiro Pardo

NILC-TR-12-02

Dezembro, 2012

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

Resumo

Na Sumarização Automática Multidocumento Multilíngue (SAMM), parte-se de uma coleção de textos em diferentes línguas para se produzir um sumário em uma das línguas dos textos de entrada. Tendo em vista a escassez de trabalhos sobre a SAMM envolvendo o português, investigam-se métodos superficiais clássicos de SA no cenário multidocumento multilíngue, os quais se caracterizam pelo tratamento simples dos fenômenos linguísticos e pelo baixo custo na produção dos sumários. Neste relatório, descrevem-se as tarefas de revisão da literatura sobre SAMM, construção de um *corpus* multidocumento multilíngue e delimitação dos métodos clássicos a serem investigados. A pesquisa ora descrita foi realizada em uma iniciação científica que compreendeu o período de 01/10/2011 a 29/02/2012, sob orientação da Profa. Dra. Ariani Di Felippo e coorientação do Prof. Dr. Thiago Pardo.

O trabalho relatado contou com financiamento da FAPESP (Proc. 2011/07617-8).

1. Introdução

No Processamento Automático das Línguas Naturais (PLN), busca-se desenvolver sistemas computacionais capazes de realizar tarefas linguísticas bastante específicas, como a correção ortográfica e gramatical, a tradução, etc. Na subárea do PLN denominada Sumarização Automática (SA), busca-se automatizar a produção de sumários (ou resumos) a partir de documentos, ou seja, a geração de versões condensadas de um ou mais textos. Os sistemas que realizam tal tarefa são denominados sumarizadores automáticos (MANI, 2001). Com a grande quantidade de texto disponível na *web*, a SA tornou-se uma aplicação computacional importante para gerenciar o enorme volume de informações (SPARCK JONES, 2007).

Quanto ao número de textos-fonte, a SA pode ser *monodocumento* ou *multidocumento*. Na SA monodocumento, produz-se um sumário de um único texto-fonte. Na SA multidocumento (SAM), produz-se um sumário a partir de uma coleção de textos-fonte em uma única língua que abordam um mesmo tópico (MCKEOWN, RADEV, 1995). Quanto à quantidade de línguas, a SA pode ser *monolíngue* ou *multilíngue*. Na SA monolíngue, parte-se de uma coleção de textos que versam sobre um mesmo assunto (em inglês, *cluster*) em uma língua *x* e produz-se um sumário na língua *x*. Na SA multilíngue, parte-se de um *cluster* composto por textos sobre um mesmo assunto em duas ou mais línguas (L1 e L2) e, a partir deles, produz-se um sumário em uma das línguas dos textos-fonte (L1 ou L2) (MANI, 2001).

Enquanto a SAM em português do Brasil (PB) vem ocupando lugar de destaque nas pesquisas sobre o processamento automático do PB (p.ex.: CAMARGO *et al.*, 2011, JORGE, PARDO, 2011, CARDOSO *et al.*, 2011a, 2011b), não se tem conhecimento de trabalhos que focalizam a produção de um sumário em PB a partir de uma coleção composta por textos em PB e em línguas estrangeiras, apesar de a crescente circulação de informações em diversas línguas. A *sumarização multidocumento multilíngue* (SAMM) pode ser exemplificada por trabalhos como o de Evans *et al.* (2005) em que, a partir de coleções de textos em inglês e traduções em inglês de textos na língua árabe, sumários em inglês são produzidos pela aplicação de métodos superficiais e profundos de seleção de conteúdo e de métodos de similaridade textual.

Neste trabalho, o objetivo é investigar métodos superficiais clássicos de SA na SAMM envolvendo o PB, em particular, textos jornalísticos. Tais métodos caracterizam-se pelo tratamento mais simples dos fenômenos linguísticos e pelo baixo custo na produção dos sumários. Tendo em vista o ineditismo, este trabalho contribui para o avanço das pesquisas sobre SAMM, além de propiciar ao aluno o exercício do método científico no PLN.

Neste relatório, estão descritas as atividades realizadas durante os primeiros 5 meses do projeto, que englobou o período de 01/10/2011 a 29/02/2012.

Este relatório está estruturalmente organizado nas seguintes seções. Na Seção 2, descrevem-se as atividades programadas e desenvolvidas até o momento e tecem-se as primeiras análises a respeito das mesmas. Na Seção 4, são apresentadas algumas considerações finais sobre o trabalho ora relatado neste relatório final.

2. Atividades realizadas

2.1. Revisão da literatura

Na revisão bibliográfica, abordaram-se: (i) conceitos básicos sobre sumarização automática, (ii) conceitos básicos sobre SAMM e os principais trabalhos relacionados e (iii) métodos superficiais de sumarização, em especial, de seleção de conteúdo.

2.1.1. A sumarização automática: conceitos básicos relevantes para o projeto

A sumarização é uma atividade bastante comum. Na modalidade escrita, tem-se, por exemplo, notícias de jornal e as sinopses de filmes. Os sumários produzidos a partir de textos são úteis porque podem ser indexadores, permitindo que o leitor descubra o assunto do texto-fonte correspondente, ou podem ser suficientemente informativos a ponto de permitirem que o leitor dispense a leitura do texto de origem (MARTINS *et al.*, 2001). Os sumários também são úteis em várias tarefas de PLN: (i) recuperação de informação, (ii) categorização de textos, etc.

Diante da utilidade dos sumários, do crescimento de informação disponível (principalmente, via *web*) e dos avanços na área de PLN, é de grande interesse a automação do processo de sumarização, foco da subárea do PLN denominada Sumarização Automática (SA). Nela, busca-se produzir automaticamente sumários a partir de um ou mais textos-fonte, sendo um "sumário" entendido como a versão mais curta de um texto ou mais textos.

De acordo com a função, os sumários podem ser informativos, indicativos ou críticos (MANI, MAYBURY, 1999). Os informativos contêm as informações principais de um texto-fonte de forma coerente e coesa ao ponto de dispensar a leitura do texto-fonte. Os indicativos apenas dizem do que o texto-fonte trata. Os críticos apresentam a informação principal do texto-fonte e avaliações sobre ele. Quanto à forma, os sistemas de SA podem produzir extratos ou *abstracts* (SPARCK JONES, 1993). Os extratos são sumários compostos por trechos inalterados do(s) texto(s)-fonte. Os *abstracts* apresentam partes reescritas do(s) texto(s)-fonte.

Quanto ao número de textos-fonte, a SA pode ser monodocumento, a partir de um único texto, e multidocumento, a partir de uma coleção de textos (MCKEOWN, RADEV, 1995).

Quanto à abordagem, Mani (2001) destaca que a SA pode ser superficial ou profunda em função da quantidade e do nível de conhecimento linguístico envolvidos na sumarização. Na superficial, utiliza-se pouco ou nenhum conhecimento linguístico para selecionar as sentenças que irão compor os sumários; o conhecimento utilizado é o empírico/estatístico. Por exemplo, uma abordagem que produz um sumário a partir da seleção e justaposição das sentenças do texto-fonte que apresentam as palavras mais frequentes do texto é classificada como "superficial". Sumarizadores superficiais costumam produzir extratos. A abordagem profunda caracteriza-se pela utilização de conhecimento linguístico morfológico, sintático, semântico e/ou pragmático-discursivo na tarefa de seleção de conteúdo para construir os sumários. Assim, os sumarizadores profundos podem gerar extratos e *abstracts*. As abordagens superficiais e profundas podem ser mescladas, originando abordagens híbridas.

Idealmente, a SA é realizada em três etapas: (i) análise dos textos-fonte, em que se produz uma representação completa de seu conteúdo; (ii) transformação, em que o conteúdo completo do texto-fonte é condensado por meio da aplicação de estratégias de seleção de

conteúdo[1] e (iii) síntese, em que o conteúdo condensado é expresso em língua natural na forma de um sumário (MANI, 2001). Essas etapas devem ser guiadas pela taxa de compressão, ou seja, pelo tamanho desejado do sumário. Por exemplo, definir a taxa de compressão em 70%, implica que o sumário apresente tamanho equivalente a 30% do tamanho do texto-fonte (medido em número de palavras, em geral).

A seguir, apresentam-se detalhes sobre a sumarização multidocumento multilíngue.

2.1.2. A sumarização automática multidocumento multilíngue (SAMM)

Como mencionado, na SA multilíngue, parte-se de um *cluster* composto por textos sobre um mesmo assunto em duas ou mais línguas (L1 e L2) e, a partir deles, produz-se um sumário em uma das línguas dos textos-fonte (L1 ou L2) (MANI, 2001). Uma tarefa bastante parecida com a SAMM é a denominada *cross-language summarization* (MANI, 2001). A diferença entre elas reside exatamente no número de línguas envolvidas no processo de sumarização. Na *cross-language summarization*, a SA pode ser monodocumento ou multidocumento, mas sempre será monolíngue.

As pesquisas sobre SAMM têm sido motivadas, como dito, pelo crescente volume de informação que circula na *web* em diferentes línguas, sobretudo em textos jornalísticos. Assim, desenvolver sumarizadores que identificam a informação principal a partir de textos em diferentes línguas e geram um sumário em um das línguas-fonte tornou-se bastante relevante.

Na SAMM, os textos-fonte estão em diferentes idiomas e, por isso, algum tipo tradução é necessária. Nesse cenário, o trabalho de Evans *et al.* (2005) é um exemplo paradigmático de esforço para a construção de um sumarizador mutlidocumento multilíngue.

Em Evans *et al.* (2005), o método[2] de SAMM parte de um *corpus* bilíngue formado por textos jornalísticos em inglês e em árabe. Por meio de um tradutor automático, os textos em árabe são traduzidos para o inglês. A tradução, no entanto, nem sempre é igual. No trabalho de Chen e Lin (2000), por exemplo, em que a sumarização parte de *clusters* compostos por textos em chinês e em inglês, traduzem-se apenas os verbos, entidades nomeadas e substantivos dos textos em chinês para o inglês. Em Evans *et al.* (2005), após a tradução, aplicam-se, na etapa de transformação, várias estratégias de relevância ou saliência apenas aos textos traduzidos com o objetivo de pontuar e ranquear as sentenças desses textos. Na sequência, as sentenças traduzidas mais bem ranqueadas de um *cluster*, que teoricamente expressam o conteúdo principal de uma coleção e devem compor os sumários, são comparadas às sentenças do texto original do mesmo *cluster* por meio de várias medidas de similaridade. Ao final, as sentenças originais identificadas como as mais similares às traduzidas compõem o sumário.

A relevância ou saliência informacional das sentenças no método de Evans *et al.* é identificada, especificamente, pelo sumarizador denominado DEMS (SCHIFFMAN *et al.*, 2002), que se baseia na utilização de várias estratégias superficiais e profundas para pontuar e

[1] A seleção de conteúdo é comumente feita em duas etapas. Na primeira, as sentenças são pontuadas e ranqueadas em função de algum critério de relevância ou saliência. Na segunda etapa, ao mesmo um critério é aplicado às sentenças ranqueadas para selecionar as unidades que comporão os sumários.

[2] Por "método", entende-se a instanciação ou especificação de uma abordagem, no caso, de SA. Um método é caracterizado pela estratégia de relevância empregada para pontuar e ranquear as sentenças e pela estratégia de seleção das sentenças ranqueadas que irão compor o sumário. Tais estratégias, por sua vez, caracterizam-se pelo nível de conhecimento linguístico e recursos e ferramentas computacionais que empregam.

ranquear as sentenças de um *cluster* de entrada. No caso, as estratégias superficiais do DEMS são as seguintes: (i) localização, segundo a qual sentenças que ocorrem no meio ou no final dos textos são penalizadas em detrimento de sentenças que ocorrem no início dos documentos; (ii) data de publicação, segundo a qual sentenças provenientes de textos mais antigos são penalizadas em detrimento de sentenças provenientes de textos mais recentes, e (iii) tamanho, critério segundo o qual sentenças abaixo de um limiar inferior (p. ex.: 15 palavras) e acima de um limiar superior (p.ex.: 30 palavras) são penalizadas. As estratégias mais profundas do DEMS, por sua vez, são: (i) identificação das "*lead words*", ou seja, palavras que ocorrem nas primeiras sentenças do texto e que indicam os tópicos do mesmo; (ii) identificação dos verbos semanticamente relevantes por meio da análise das estruturas "sujeito-verbo" e "verbo-objeto"; no caso, busca-se identificar os verbos de conteúdo semântico específicos que também indicam o tópico do texto e (iii) identificação dos conceitos representativos do texto por meio da identificação de sinônimos e hiperônimos/hipônimos (SCHIFFMAN *et al*., 2002). Com base nesse conjunto de critérios, as sentenças dos textos traduzidos são ranqueadas e as mais bem pontuadas são tidas como as mais informativas da coleção para compor o sumário.

Tendo em vista que as sentenças ranqueadas são traduções, estas podem apresentar problemas de gramaticalidade, coesão, coerência, etc. Assim, no método de Evans *et al.*, medidas de similaridade são aplicadas às sentenças ranqueadas e aos textos originais em inglês com o objetivo de identificar as sentenças originais mais similares às traduzidas. Tais medidas são aplicadas por meio do Simfinder (HATZIVASSILOGLOU *et al*., 2001), uma ferramenta que identifica similaridade textual com base no compartilhamento de vários atributos lexicais e sintáticos, a saber: (i) nomes próprios, (ii) itens lexicais com mesmo radical ou morfologicamente relacionados, (iii) itens lexicais sinônimos, (v) itens lexicais com mesmo hiperônimo e (iv) núcleos sintagmáticos. A partir da medida de similaridade estabelecida em função desses inúmeros fatores linguísticos, as sentenças originais em inglês que foram classificadas como as mais similares substituem as traduzidas mais bem ranqueadas. Ao final, o sumário é composto apenas por sentenças originais em inglês.

Por meio do trabalho de Evans *et al.* (2005), vê-se que, uma vez traduzidos os textos-fonte, a seleção de conteúdo na SAMM passa a ser aplicada a coleções de textos monolíngues. Em outras palavras, a sumarização deixa de ser "multidocumento" e "multilíngue" e passar a ser apenas "multidocumento". Assim, após a tradução, é possível aplicar métodos superficiais e profundos de seleção de conteúdo que são comumente aplicados à sumarização multidocumento[3], sendo a maioria, aliás, adaptações de métodos aplicados à sumarização monodocumento, como os de Evans *et al.* (2005), por exemplo.

Como dto, os métodos profundos caracterizam-se pela utilização de conhecimento linguístico de nível morfológico, sintático, semântico e até pragmático-discursivo na tarefa de seleção de conteúdo para construir os sumários, sendo que a manipulação desse conhecimento é feita pela utilização de certas ferramentas e recursos linguístico-computacionais.

[3] Na SAM, em que dois ou mais textos sobre um mesmo assunto, provenientes de fontes distintas, são condensados em um sumário, a redundância é um fenômeno característico (JURAFSKY, MARTIN, 2009). Neste trabalho, no entanto, a redundância não é tratada, pois isso torna a sumarização muito mais complexa.

Em Evans *et al.*, por exemplo, ressalta-se a utilização de um analisador sintático[4] (em inglês, *parser*), ferramenta por meio da qual é possível identificar estruturas dos tipos "sujeito-verbo" e "verbo-objeto", e de um recurso léxico-conceitual específico, a WordNet de Princeton (WN.Pr) (FELLBAUM, 1998). A WN.Pr, em especial, é uma base léxico-conceitual desenvolvida para o inglês norte-americano em que as unidades lexicais estão organizadas em função da relação léxico-semântica da sinonímia e de várias relações semântico-conceituais, como a hiponímia e hiperonímia. Esse recurso é utilizado especificamente para (i) a identificação dos conceitos representativos do texto, um dos critérios utilizados na seleção das sentenças traduzidas para a composição dos sumários, e (ii) a identificação da similaridade entre as sentenças traduzidas e as originais em função da presença de itens lexicais sinônimos e hiperônimos/hipônimos.

Além desses, ressalta-se que, quando baseados na manipulação de conhecimento discursivo, os sumarizadores comumente utilizam um analisador discursivo (ou *parser* discursivo), ferramenta capaz de identificar relações discursivas como as previstas na teoria/modelo RST (*Rhetorical Structure Theory*) (MANN, THOMPSON, 1987) (p.ex.: causa-efeito, contraste, elaboração, etc.) entre as sentenças de um texto a ser sumarizado. A RST é amplamente utilizada para a SA monodocumento (cf. UZÊDA *et al.*, 2010) principalmente por distinguir, dada uma relação, o segmento nuclear (N) e o satélite (S). No caso, a ideia básica da SA baseada na RST é a de que os satélites são informação complementar e, portanto, não devem ser selecionados parte compor o sumário.

A utilização de recursos e ferramentas linguístico-computacionais como *wordnets* e analisadores sintáticos e discursivos caracteriza os métodos como "profundos", que comumente geram sumários melhores, principalmente quanto à informatividade e coesão/coerência. No entanto, esses recursos e ferramentas são dependentes de língua e, por isso, nem sempre existem ou estão disponíveis (LENCI *et al.*, 2002). No caso do PB, a base de dados no formato *wordnet*, a WordNet.Br, está em pleno desenvolvimento (DIAS-DA-SILVA *et al.*, 2008).

Assim, tendo em vista o baixo custo e a independência de língua, muito se tem focalizado a abordagem superficial (MANI, 2001). Em consequência da adoção da abordagem superficial, a maioria dos trabalhos produz sumários extrativos (ou extratos) e não *abstracts*, os quais apresentam problemas conhecidos de coesão e informatividade. Apesar das limitações, reconhecem-se a relevância e a utilidade dos sumários extrativos em certas aplicações.

Na literatura sobre SA monodocumento, vários métodos superficiais foram propostos. Nos métodos superficiais clássicos, a etapa de análise é bastante simples, consistindo basicamente na segmentação textual (geralmente, em sentenças) do(s) texto(s)-fonte. A etapa de transformação consiste no reconhecimento das unidades de significado do texto-fonte mais relevantes (isto é, que contêm as ideias centrais do texto) para compor o sumário por meio da aplicação de medidas estatísticas simples e, sobretudo, pelo tratamento de conhecimento linguístico básico (MANI, 2001).

A seguir, descrevem-se os métodos superficiais clássicos que são mais frequentes na literatura sobre SA.

[4] Ferramenta que reconhece a estrutura sintática de uma sentença, atribuindo funções sintáticas aos constituintes reconhecidos (CARROL, 2004).

2.1.3. Os métodos superficiais de SA

- Palavras-chave (em inglês, *keywords*): esse método é baseado no fato de o escritor usar algumas palavras-chave para expressar suas ideias principais (BLACK, JOHNSON, 1988), as quais tendem a ser recorrentes no texto. Assim, partindo da identificação das palavras-chave do texto (que podem ser as mais frequentes), selecionam-se as sentenças em que tais palavras ocorrem para compor o sumário, até que a taxa de compressão seja atingida. Uma variação desse método consiste em pontuar as sentenças do texto-fonte em função das palavras-chave nelas contidas e gerar um ranque das mesmas (BLACK, JOHNSON, 1988). No caso, a sentença de maior pontuação no ranque é, no método *keywords*, a primeira a ser selecionada para compor o sumário e, em seguida, passa-se a segunda sentença de maior pontuação. Esse processo repete-se até que todas as sentenças relevantes tenham sido selecionadas e/ou se tenha atingido a taxa de compressão desejada. Em outra variação do método *keywords*, apenas as palavras do título são consideradas "chave". Nesse caso, selecionam-se apenas as sentenças em que tais palavras ocorrem (EDMUNDSON, 1969). Outras possibilidades incluem, por exemplo, considerar apenas nomes e verbos como palavras-chave, pois as unidades lexicais dessas classes tendem a ser mais significativas.

- Frequência: esse método pontua as sentenças de um texto com base na soma das frequências de todas as suas palavras constitutivas, com exceção das *stopwords*[5]. Especificamente, calcula-se a frequência de ocorrência de todas as palavras de classe aberta em um texto. Na sequência, a pontuação de cada sentença desse mesmo texto é calculada em função da soma da frequência de cada uma de suas palavras plenas constitutivas. Para a aplicação desse método, assim como para o *keywords* e suas variações, é comum que se determine um *threshold* para seleção das sentenças, isto é, um valor mínimo de relevância aceitável (p.ex.: a média das pontuações de todas as sentenças do texto-fonte).

- Localização: esse método classifica as sentenças que ocorrem no início dos textos como as mais importantes (BAXENDALE, 1958). Esse método pauta-se na hipótese de que o conteúdo informacional em um texto está distribuído em sua estrutura de parágrafos. Os textos do tipo informativo (e gênero jornalístico), como é o caso das notícias que compõem o *subcorpus* em questão, têm a função de relatar fatos ou acontecimentos atuais, de interesse e importância para a comunidade e de fácil compreensão pelo público leitor. Para informar um fato, comumente se constroi um texto com base no método da pirâmide invertida, segundo o qual a informação é ordenada por ordem decrescente de importância. Assim, com exceção do título, uma notícia apresenta o (i) *lead*, que corresponde ao primeiro ou aos dois primeiros parágrafos do texto e expressa a informação principal a ser relatada, e (ii) o corpo do texto, que desenvolve os elementos informativos referidos no *lead* (LAGE, 2002). Por conseguinte, quanto mais inicial for uma sentença, mais importante esta é para a expressão na informatividade.

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

- Expressões indicativas: nesse método, a seleção das sentenças é feita com base na presença de certas expressões sinalizadoras (*cue phrases* ou *cue words*) (PAICE, 1981). Por exemplo, em textos científicos, expressões como "o objetivo deste trabalho é..." ou "este artigo apresenta..." e palavras como "resultados" e "conclusões" são bons guias para a seleção das sentenças a serem incluídas no sumário. Textos de diferentes gêneros e tipos podem ter diferentes expressões indicativas. Em um texto jornalístico de esportes, por exemplo, expressões como "o ganhador é" e "o placar foi de" podem guiar a seleção.

- Relacional: nesse método, as sentenças do texto-fonte são inter-relacionadas com base nas suas palavras constitutivas (SKOROCHOD´KO, 1972). Nos métodos mais simples, as sentenças são relacionadas com base na ocorrência de formas lexicais idênticas. Assim, as sentenças inter-relacionadas pelo maior número de relações são selecionadas para compor o sumário. Ressalta-se que, dada uma cadeia de sentenças inter-relacionadas, a deleção de uma delas resultaria em um sumário menos coerente e coeso.

- Mineração de texto: o método TF-ISF (*Term Frequency – Inverse Sentence Frequency*), proposto por Larocca Neto *et al*. (2000), determina a relevância das sentenças de um texto em função das suas palavras mais representativas, que são ou não palavras-chave do texto. A medida estatística TF-ISF foi adaptada de outra medida bastante comum da área de Recuperação de Informação (*Information Retrieval*), a TF-IDF (*Term Frequency – Inverse Document Frequency*) (SALTON, BUCKLEY 1988), que determina a importância de um documento em dada coleção de documentos. Na adaptação, o foco passou a ser a importância das sentenças em um texto.

2.2. Construção do *corpus*

Para a execução desta pesquisa, foi necessário um *corpus*. Por definição, um *corpus* é um conjunto de dados linguísticos sistematizados de acordo com determinados critérios, de maneira que possa ser processado por computador com a finalidade de propiciar resultados vários e úteis para a descrição e análise (SINCLAIR, 2005). Por essa definição, um *corpus* é um artefato produzido para a pesquisa e que, por isso, a maioria de suas características é dependente dos objetivos da pesquisa. Tendo em vista o trabalho ora descrito, o *corpus* tinha de apresentar as seguintes características:

(i) Multidocumento: essa característica advém naturalmente do fato de que o foco deste trabalho está na SA multidocumento; portanto, necessitava-se de um *corpus* composto por coleções (ou *clusters*) de textos que versam sobre um mesmo assunto.

(ii) Multilíngue: além de ser multidocumento, precisava-se de um *corpus* multilíngue, ou seja, composto por coleções de textos que versam sobre um mesmo assunto em diferentes línguas; tendo em vista o processamento automático do PB e a relevância da língua inglesa e espanhola, optou-se por um *corpus* trilíngue, composto, portanto, por textos sobre um mesmo tópico em PB, inglês e espanhol.

(iii) Jornalístico: essa característica advém da facilidade de seleção e compilação de textos da *web* que versam sobre um mesmo assunto em diferentes línguas.

Diante de tais características, não foi possível encontrar na literatura um *corpus* disponível que as satisfizesse ou que pudesse ser adaptado para tanto. Assim, partiu-se para a construção de um *corpus* com base nas características descritas.

A tarefa de compilação de um *corpus* engloba dois processos: a seleção das fontes e a coleta dos textos dessas fontes. No projeto em questão, a principal fonte de coleta foi a *web* e a estratégia de coleta dos textos foi manual, ou seja, as páginas desejadas foram acessadas uma a uma e os arquivos (textos) de interesse foram salvos no computador um a um. Especificamente, foram selecionadas as versões eletrônicas dos seguintes jornais: (i) A Folha de São Paulo[6], para a coleta dos textos em PB; (ii) BBC News[7], para a coleta dos textos em inglês, e (iii) El pais[8], para os textos em espanhol. Tais fontes foram selecionadas por serem versões eletrônicas de jornais importantes quanto à circulação e qualidade das notícias. A partir de tais fontes, foram compiladas 10 coleções de textos de domínios variados, sendo que cada coleção possui 3 textos que versam sobre um mesmo assunto, sendo 1 em PB, 1 em inglês e 1 em espanhol.

Para a seleção dos domínios/tópicos, optou-se pelos critérios "atualidade" e "variedade", ou seja, buscou-se selecionar diferentes domínios e assuntos em circulação na mídia jornalística na época da coleta. No Quadro 2, os domínios a que pertencem os textos estão indicados pelo nome das seções da versão *on-line* do jornal A Folha de São Paulo nas quais os textos foram publicados. Dessa forma, vê-se que há apenas 4 domínios no *corpus* (mundo, poder, saúde e ambiente). No caso, a diversidade de domínios não foi maior devido à dificuldade de identificar notícias de outros domínios veiculadas nas três línguas em questão, pelo menos no período em que o *corpus* foi construído. Os assuntos, por sua vez, são bastante variados. Para a seleção dos textos, em especial, dois critérios foram aplicados: (i) tamanho e (ii) originalidade. Como consequência, foram selecionados, para a construção de um *cluster*, textos semelhantes quanto ao tamanho. No *cluster* 2 (C2), por exemplo, o texto em PB possui 287 palavras, o texto em inglês possui 231 e o em espanhol possui 377 palavras. Quanto à originalidade, buscou-se selecionar, para a construção de um *cluster*, textos que não fossem traduções uns dos outros. No Quadro 2, estão descritas as principais características dos textos que compõem o *corpus*, o qual, ao final, foi denominado CM3News (ou seja, *corpus* multidocumento trilíngue de textos jornalísticos).

Cluster	**Domínio**	**Assunto/Tópico**	**Documento**	**Língua**	**Publicação (data - hora)**	**Número/palavras**
C1	Mundo	Ataques em Londres	D1_C1_folha	Português	11/08/2011 – 09:11	1.721
			D2_C1_bbc	Inglês	11/08/2011 – 11:10 (GMT)	
			D3_C1_elpais	Espanhol	11/08/2011	
C2	Poder	Kit gay	D1_C2_folha	Português	25/05/2011 – 13:12	895
			D2_C2_bbc	Inglês	25/05/2011 – 21:07 (GMT)	
			D3_C2_elpais	Espanhol	26/05/2011 – 00:00 (CET)	
C3	Saúde	Intoxicação alimentar	D1_C3_folha	Português	30/05/2011 – 18:47	1.915
			D2_C3_bbc	Inglês	30/05/2011 – 5:43 (GMT)	
			D3_C3_elpais	Espanhol	29/05/2011	

[6] http://www.folha.uol.com.br/
[7] http://www.bbc.co.uk/news/
[8] http://elpais.com/

C4	Mundo	Massacre na Noruega	D1_C4_folha	Português	08/08/2011 – 14h20	1.613
			D2_C4_bbc	Inglês	02/08/2011 – 14:52 (GMT)	
			D3_C4_elpais	Espanhol	27/07/2011	
C5	Ambiente	Novo código Florestal	D1_C5_folha	Português	25/05/2011 – 00:43	2.093
			D2_C5_bbc	Inglês	25/05/2011 – 09:50 (GMT)	
			D3_C5_elpais	Espanhol	26/05/2011	
C6	Mundo	Conflito Universidade da Califórnia	D1_C6_folha	Português	20/11/2011 – 00:15	1.193
			D2_C6_bbc	Inglês	21/11/2011 – 23:26 (GMT)	
			D3_C6_elpais	Espanhol	22/11/2011 – 21:36 (CET)	
C7	Saúde	Proibição do fumo em NY	D1_C7_folha	Português	24/05/2011 – 13:38	1.224
			D2_C7_bbc	Inglês	24/05/2011 – 18:36 (HKT)	
			D3_C7_elpais	Espanhol	23/05/2011	
C8	Mundo	Terremoto na Nova Zelândia	D1_C8_folha	Português	05/03/2011 – 05:01	1.329
			D2_C8_bbc	Inglês	03/03/2011 – 04:45 (GMT)	
			D3_C8_elpais	Espanhol	01/03/2011	
C9	Mundo	Terremoto em Missouri	D1_C9_folha	Português	23/05/2011 – 08:04	1.880
			D2_C9_bbc	Inglês	23/05/2011 – 20:21 (GMT)	
			D3_C9_elpais	Espanhol	23/05/2011	
C10	Mundo	Vulcão na Islândia	D1_C10_folha	Português	24/05/2011 – 12:13	2.276
			D2_C10_bbc	Inglês	24/05/2011 – 15:51 (GMT)	
			D3_C10_elpais	Espanhol	22/05/2011	

Quadro 2: Características dos *clusters* do *corpus* CM3News.

Ressalta-se que os textos do CM3News estão organizados em diferentes pastas. Os textos originais, por exemplo, nomeados segundo o padrão *documento_cluster_fonte* e salvos em formato txt (p.ex.: `D1_C1_folha.txt`), estão armazenados na pasta "Originais". Além disso, a numeração dos documentos em um *cluster* indica a língua de origem do mesmo, sendo 1 para os textos em PB, 2 para os textos em inglês e 3 para os textos em espanhol. No exemplo em questão, a nomeação (`D1_C1_folha.txt`) indica que se trata do texto em PB do *cluster* 1. Os textos foram armazenados na pasta "Originais" sem os seus respectivos títulos, posto que estes não são processados na sumarização. Os títulos dos textos foram armazenados, também em formato txt, em uma pasta específica, denomina "Títulos".

2.3. Tradução e Segmentação dos textos-fonte

A etapa de tradução automática é típica da SAMM e foi realizada após a construção do CM3News. Diante do objetivo de produzir sumários em PB a partir de *clusters* compostos por textos em PB, inglês e espanhol, essa etapa consistiu na tradução dos textos em inglês e espanhol para o PB.

Tendo em vista que, na SAMM, o foco é o processo de sumarização (ou melhor, de seleção de conteúdo), utilizam-se comumente ferramentas (ou serviços) externas de tradução. No caso, utilizou-se o serviço *online* Google Translator[9] devido ao reconhecido desempenho do

[9] http://translate.google.com/

mesmo na literatura. Na competição internacional realizada em 2005 pelo NIST (do inglês, *National Institute of Standards and Technology*), em que os sistemas estatísticos e os baseados em regras (ou simbólicos) foram comparados, o referido tradutor obteve resultados 376% mais adequados que o Systran, para, por exemplo, o par de línguas inglês-árabe. Além disso, ressalta-se que a principal motivação do surgimento da SAMM foi o enorme volume de informação redundante disponível em várias línguas especificamente na *web*. Sendo o Google Translator uma aplicação *web-based*, pareceu-nos adequado a integração do mesmo no cenário da SAMM.

Os textos traduzidos foram salvos em formato `txt` e armazenados na pasta "Traduções". Ao padrão de nomeação dos arquivos originais em inglês e espanhol, foi inserida a etiqueta T para indicar que se trata de uma tradução (p.ex.: `D2_C1_folhaT.txt`).

Uma vez traduzidos, os textos-fonte foram segmentados. A segmentação importante para várias aplicações de PLN e o nível de segmentação varia conforme a especificidade das mesmas. Esse processo pode ser no nível das palavras, sintagmas, orações, sentenças, parágrafos ou tópicos. Tendo em vista que os sumários extrativos são comumente construídos pela concatenação de sentenças dos textos-fonte, a segmentação sentencial dos textos originais e dos traduzidos do CM3News fez-se necessária.

Vale ressaltar que a segmentação dos textos originais foi feita porque todos os textos ou documentos (ou seja, D1, D2, e D3) de um *cluster* C são submetidos ao processo de sumarização, não somente os traduzidos como em Evans *et al.* (2005). Neste trabalho, não são aplicadas estratégias de similaridade textual para substituir as sentenças (traduzidas) selecionadas para compor o sumário por sentenças originais, posto que isso é bastante complexo para ser aplicado nesta primeira investigação sobre SAMM envolvendo o PB.

Assim, seguindo a tradição em SA, a segmentação dos textos originais e traduzidos foi realizada no nível sentencial. Para tanto, utilizou-se a ferramenta SENTER (em inglês, *SENtence splitTER*) (PARDO, 2006), um segmentador sentencial automático para o PB. Essa ferramenta identifica as sentenças por regras simples que se baseiam na ocorrência de sinais de pontuação[10], como o ponto final e os sinais de interrogação e exclamação e, na sequência, gera um arquivo no formato `txt` para cada texto segmentado, sendo que cada sentença é disposta em uma linha individual. A nomeação dos arquivos segmentados segue a dos arquivos originais, sendo que, a essa nomeação, é acrescida a etiqueta `seg` (p.ex.: `D1_C1_folha.txt.seg`). Os textos segmentados foram armazenados na pasta "Segmentação". Para ilustração, tem-se, no Quadro 4, um trecho segmentado do texto original do Quadro 3. No exemplo, as sentenças são delimitadas por colchetes e numeradas em sequência.

> [...] Segundo o ministro Gilberto Carvalho (Secretaria Geral), Dilma considerou o material do MEC "inadequado" e o vídeo "impróprio para seu objetivo".
> A manifestação ocorreu na esteira de uma reunião de Carvalho com a bancada evangélica da Câmara. O grupo de parlamentares chegou a ameaçar o governo com obstrução da pauta no Congresso, colaborar com assinaturas para convocar o ministro Antonio Palocci (Casa Civil) a se explicar sobre sua evolução patrimonial e propor uma CPI para investigar o MEC. [...]

Quadro 3: Texto original.

[10] A distinção entre o ponto das abreviaturas e o ponto final é feita pela consulta a uma lista de abreviaturas.

[...] [Segundo o ministro Gilberto Carvalho (Secretaria Geral), Dilma considerou o material do MEC "inadequado" e o vídeo "impróprio para seu objetivo".]**2**
[A manifestação ocorreu na esteira de uma reunião de Carvalho com a bancada evangélica da Câmara.]**3**
[O grupo de parlamentares chegou a ameaçar o governo com obstrução da pauta no Congresso, colaborar com assinaturas para convocar o ministro Antonio Palocci (Casa Civil) a se explicar sobre sua evolução patrimonial e propor uma CPI para investigar o MEC.]**4** [...]

Quadro 4: Texto segmentado.

2.4. Aplicação do método da frequência

Tendo em vista que, neste trabalho, realizam-se as primeiras investigações sobre a SAMM envolvendo PB, optou-se por investigar os métodos superficiais de mais simples aplicação e mais difundidos na literatura[11], a saber: (i) frequência e (ii) localização. A seguir, descreve-se a aplicação do método baseado na frequência.

Como já foi dito, nos métodos superficiais de SA, utiliza-se conhecimento linguístico simples ou nenhum conhecimento linguístico. No método da frequência das palavras, a seleção de conteúdo para a composição do sumário é pautada em um critério estatístico.

Especificamente, as sentenças de um texto são pontuadas em função da soma das frequências de ocorrência de suas palavras constitutivas. A partir dessa pontuação, as sentenças são ranqueadas e as de maior pontuação são selecionadas para compor o sumário até que a taxa de compressão seja atingida.

Para a pontuação das sentenças dos *clusters* do CM3News em função do método da frequência, foram utilizadas algumas funcionalidades do GistSumm (PARDO, 2002). Esse sistema é um sumarizador extrativo monodocumento desenvolvido para o PB que foi estendido para realizar, entre outras tarefas, a sumarização multidocumento (PARDO, 2005). Essa ferramenta seleciona as sentenças para compor os sumários com base na identificação da ideia central do texto. Na SAM, o GistSumm produz um ranque das sentenças de um *cluster* em função da frequência de suas palavras (no *cluster*) e assume a sentença de maior pontuação como a sentença *gist,* ou seja, a que possui a ideia central do *cluster.* Essa sentença é selecionada para iniciar o sumário. A seleção das demais sentenças para compor o sumário é feita, por exemplo, com base no compartilhamento de palavras com a sentença *gist.*

Para a realização deste trabalho, os textos do CM3News foram submetidos ao GistSumm apenas para a pontuação das sentenças com base na frequência. A funcionalidade de seleção de conteúdo da ferramenta, baseada na identificação da sentença *gist*, não foi aplicada. Entretanto, futuramente, os textos do CM3News poderão ser efetivamente sumarizados pelo GistSumm e os resultados desse processamento poderão ser comparados aos resultados deste trabalho, posto que o método do GistSumm é diferente dos métodos aqui investigados.

[11] A escolha pelos métodos superficiais justifica-se também pelo fato de que os resultados de trabalhos futuros sobre a investigação de métodos mais profundos na SAMM poderão ser comparados aos resultados deste primeiro trabalho envolvendo o PB.

a) O pré-processamento dos textos

A partir de um *cluster*, fornecido ao GistSumm como dados de entrada, o sistema realiza 3 processos que precedem à pontuação e ranqueamento das sentenças. Tais processos são:

(i) *Case folding*: processo automático em que todas as letras de um texto, por exemplo, são colocadas em uma mesma caixa (minúscula ou maiúscula) (DIAS-DA-SILVA *et al.*, 2007); no caso deste trabalho, o *case folding* consistiu em trocar todas as letras maiúsculas dos textos por minúsculas; p.ex.: na sentença "Uma mulher foi levada ao hospital na Polônia na segunda-feira." (D2_C3_bbcT.txt), a letra "U" do artigo "Uma" foi substituída por "u" e a letras "P" de Polônia foi substituída por "p";

(ii) Radicalização (em inglês, *stemming):* processo automático em que as palavras de um texto sob processamento são reduzidas aos radicais (SPARCK-JONES, WILLET, 1997); por radical, entende-se a forma mínima que está presente em todas formas de uma mesma palavra ou a parte da palavra comum às variações de flexão;

(iii) Remoção de *stopwords*: processo automático em que as *stopwords*, ou seja, as palavras funcionais (p.ex.: preposições, artigos, conjunções, etc.), são removidas do texto por serem muito frequentes e vazias de significado; para tanto, o sistema faz uso de uma *stoplist*, em que constam as *stopwords* a serem ignoradas caso ocorram no texto sob análise.

No Quadro 5, ilustram-se os resultados ou dados de saída produzidos por cada um dos processo realizados pelo GistSumm que antecedem a pontuação e o ranqueamento das sentenças. Para tanto, a *sentença "Uma mulher foi levada ao hospital na Polônia na segunda-feira"* foi considerada o dado de entrada do sistema.

Processos	**Exemplos**
Case folding	uma mulher foi levada ao hospital na polônia na segunda-feira
Radicalização	um mulh ser[12] lev a o hospita em a polôn em a segunda-feir
Remoção de *stopwords*	mulh ser lev hospita polôn segunda-feir

Quadro 5: Exemplos de dados de saída do pré-processamento do GistSumm.

Após o pré-processamento dos textos de um *cluster*, o GistSumm realiza efetivamente as tarefas de pontuação e ranqueamento das sentenças. A seguir, descreve-se com mais detalhes o cálculo da frequência para a pontuação das sentenças.

b) A pontuação das sentenças

Como mencionado, a pontuação das sentenças de um *cluster*, composto por 1 texto segmentado em PB e 2 textos traduzidos e segmentados, é feita pela soma da frequência de ocorrência de suas palavras constitutivas no *cluster*. Ressalta-se, neste ponto, que, apesar de calcular a frequência de todas as palavras de uma sentença, incluindo-se as *stopwords*, estas não são consideradas no cálculo da pontuação da sentença.

[12] No caso dos verbos com alomorfia de radical, p.ex. o verbo "ser", o sistema opta por reduzir as palavras do paradigma flexional à forma canônica. No caso, vê-se que os sistema reduziu "foi" a "ser".

A seguir, exemplifica-se o processo de pontuação excluindo-se as *stopwords.* No caso da sentença *"Uma mulher foi levada ao hospital na Polônia na segunda-feira"*, o sistema identificou que os elementos restantes dos três processos do Quadro 5 possuem as seguintes frequência no *cluster* em questão: mulh=6, ser=29, lev=3, hospita=1, polôn=1 e segunda-feir=2. Somando-se a frequência de todos esses elementos (6+29+3+1+1+2), a sentença sob análise recebe a pontuação 42. Assim como ilustrado, calcula-se a frequência das palavras de todas as sentenças do *cluster* e a pontuação das sentenças. Ao final, as sentenças são ranqueadas de acordo com sua pontuação.

No Quadro 6, estão descritas as 20 sentenças mais bem pontuadas do *cluster* C3 em função do método de saliência ou relevância baseado na frequência. Em outras palavras, diz-se que, quanto à frequência de ocorrência das palavras no *cluster,* as 20 sentenças do Quadro 6 são as mais relevantes ou salientes do C3.

Ranque	**Pontuação**	**Sentenças do *cluster* (texto-fonte)**
1ª	242	Isto foi dito hoje em Lleida Secretário de Estado para a Água Rural, Josep PUXEU, que em conferência de imprensa denunciou a situação de emergência "barreiras" que alguns países e operadoras, como a Áustria, estão colocando as exportações frutas e vegetais espanhol como resultado de informações "não testados" feita a partir de Alemanha sobre esse surto, informou à Agência Efe. **(D3_C3_elpaisT.txt)**
2ª	235	Comerciantes austríaca, a pedido das autoridades de saúde começaram a recordar os tomates, pepinos e beringelas espanhol para evitar serem consumidos aqueles que contêm as bactérias prejudiciais E_ coli e está se espalhando entre as pessoas e provocar uma epidemia, como a Alemanha, onde há 10 mortos e mais de mil infectados. **(D3_C3_elpaisT.txt)**
3ª	227	Os alemães foram alertados para não comer pepinos até que os testes identificar a fonte de um surto de E. coli mortal que autoridades locais dizem que já matou 13 pessoas. **(D2_C3_bbcT.txt)**
4ª	202	Acredita-se que a fonte da infecção sejam pepinos contaminados que foram importados da Espanha e depois enviados da Alemanha para outros países europeus, mas isso ainda terá que ser confirmado por testes. **(D1_C3_folha.txt)**
5ª	195	Como mídia alemã informou o número de pessoas infectadas aumentou para 1.200, ministro da Saúde, Daniel Bahr estava se preparando para realizar conversações de emergência com Consumidores Ministro Ilse Aigner e representantes do Estado regionais para discutir o surto, disseram autoridades. **(D2_C3_bbcT.txt)**
6ª	193	Até o momento, 1,2 mil casos confirmados ou suspeitos de infecções pela E_ coli foram registrados na Alemanha, e centenas de pessoas também foram infectadas na Suécia, Dinamarca, Holanda e Reino Unido. **(D1_C3_folha.txt)**
7ª	192	Os três tipos de vegetais que as autoridades austríacas decidiram retirar-se dois grossistas alemã, que comprou no mercado central de Hamburgo (norte da Alemanha) e foram os que alertaram os seus parceiros austríacos e retirou esses produtos do mercado após ter sido informado pelo Ministério da Saúde alemão. **(D3_C3_elpaisT.txt)**
8ª	180	A decisão russa foi anunciada após o pedido do governo alemão para que a população não coma pepinos até que os cientistas consigam identificar a origem da bactéria que já

		matou 14 pessoas no país. **(D1_C3_folha.txt)**
9ª	180	O Centro Europeu para a Prevenção e Controle de Doenças (ECDC, na sigla em inglês), com sede na Suécia, disse que o surto de SHU é "um dos maiores que já foram registrados no mundo e o maior já registrado na Alemanha". **(D1_C3_folha.txt)**
10ª	175	As autoridades espanholas disseram que a Europa não deve ser tão rápido para culpar produzir espanhol, e disseram que iriam buscar uma resposta da UE por perdas e danos incorridos pela reivindicação. **(D2_C3_bbcT.txt)**
11ª	161	Duas estufas espanholas identificadas como fontes para o surto foram fechadas e estão atualmente sob investigação para ver se o surto se originou lá ou em outros lugares, disse um porta-voz da UE. **(D2_C3_bbcT.txt)**
12ª	160	Acreditamos que estes produtos não são mais disponíveis para venda, embora não podemos excluir isso, disse a emissora pública austríaca ORF Rendi-Wagner, que apelou aos consumidores que "se os sintomas, como diarréia nos próximos dias, deve correu para o médico. " **(D3_C3_elpaisT.txt)**
13ª	158	A Rússia proibiu nesta segunda-feira a importação de alguns tipos de vegetais comestíveis da Alemanha e da Espanha, incluindo pepinos e tomates, por causa de um surto de infecções provocado por uma variedade da bactéria. **(D1_C3_folha.txt)**
14ª	155	Até o momento, suspeita-se que pepinos orgânicos espanhóis exportados para outros países europeus pela Alemanha ou diretamente pela Espanha, tenham desencadeado o surto infeccioso. A epidemia na Alemanha, centrada no norte, está em curso e já 10 pessoas, todas da região, que morreram, nove dos quais são mulheres, enquanto o número de infectados ou suspeitos de estarem mais de mil neste epidemia que, segundo estimativas oficiais, ainda não atingiu seu pico. **(D1_C3_folha.txt)**
15ª	152	A doença não é contagiosa, mas diretamente ele pode ser transferido entre as pessoas, se uma pessoa infectada prepara comida para os outros. **(D2_C3_bbcT.txt)**
16ª	151	O chefe da Eppendorf de Hamburgo Universidade de Clinic, Joerg Debatin, disse que mais mortes eram esperados, como 30 pessoas infectadas com HUS tinha perdido a função renal. **(D2_C3_bbcT.txt)**
17ª	148	Cientistas suspeitam que pepinos, tomates e alface possam ter espalhado a bactéria, mas até agora apenas amostras de pepinos analisados em Hamburgo, na Alemanha, tiveram a contaminação constatada. **(D1_C3_folha.txt)**
18ª	147	Suspeita caiu sobre pepinos orgânicos importados da Espanha pela Alemanha, mas, em seguida, re-exportados para outros países europeus, ou exportados directamente pela Espanha. **(D2_C3_bbcT.txt)**
19ª	143	Todos, mas uma dessas mortes foram registradas no norte da Alemanha, mas teme que o surto foi se espalhando aumentaram quando uma mulher de 91 anos morreu no estado ocidental da Renânia do Norte-Vestfália. **(D2_C3_bbcT.txt)**
20ª	141	Por seu turno, o Governo espanhol não descarta responsabilizadas pelas "enormes danos" que a "especulação" pelas autoridades alemãs sobre a origem dos pepinos contaminados estão trazendo sector agrícola e alimentar. **(D3_C3_elpaisT.txt)**
...	...	...
69ª	42	Uma mulher foi levada ao hospital na Polônia na segunda-feira. **(D2_C3_bbcT.txt)**

Quadro 6: Exemplo de ranqueamento das sentenças de um *cluster* em função da frequência.

Além das 20 sentenças mais bem pontuadas, aponta-se, no Quadro 6, que a sentença utilizada como exemplo neste relatório ("*Uma mulher foi levada ao hospital na Polônia na segunda-feira*"), a qual obteve 42 pontos no cálculo da frequência de suas palavras constitutivas, ocupa a posição de número 69, de um total de 75 sentenças no *cluster* C3. A pontuação das sentenças em função da frequência já foi calculada para os 10 *clusters* que compõem o *corpus* CM3News.

Apesar da seleção das sentenças para compor o sumário a partir do ranqueamento descrito não ter sido efetivamente feito, algumas observações, no entanto, podem ser tecidas. Tomando–se uma taxa de compressão de 70%, o sumário do *cluster* C3 deve apresentar tamanho equivalente a 30% do tamanho dos textos-fonte. No caso, C3 possui 1.915 palavras e seu sumário, por conseguinte, deve conter 574 palavras. Apenas com base no raqueamento, o sumário englobaria as primeiras 14 sentenças justapostas, em um total de 546 palavras. Além disso, outros critérios de seleção podem ser considerados, por exemplo, a língua de origem. Por exemplo, caso haja sentenças com pontuações idênticas ou muito próximas, dar-se-á preferência à original em PB. Isso é o que acontece com as 5ª, 6ª, 7ª sentenças, que obtiveram, respectivamente, as pontuações 195, 193 e 192. Segundo o critério de seleção da língua de origem, escolher-se-á a sentença na 6ª posição (193), posto que esta é proveniente do texto D1_C3_folha.txt. Esse critério, no entanto, nem sempre é aplicável. Por exemplo, as sentenças nas 8ª e 9ª colocações, obtiveram a mesma pontuação, 180, e ambas são provenientes do texto original em PB. Nesse caso, outro critério de seleção precisa ser aplicado. Uma possibilidade é a consideração do tamanho da sentença. Segundo esse critério, dar-se-á privilégio à sentença de menor tamanho para compor o sumário, no caso, a da 8ª posição, que possui 35 palavras enquanto e da 9ª posição possui 42.

Por fim, ressalta-se que, para a aplicação do método da frequência, não se está tratando o fenômeno da redundância, típico da SAM. Assim, sabe-se, de antemão, que os sumários produzidos com base nos métodos superficiais de relevância (isto é, frequência, palavra-chave, localização, expressões indicativas, relacional e mineração de texto) e nas estratégias de seleção aqui mencionadas (ranqueamento, língua de origem e tamanho) irão apresentar excesso de informação repetida. Além disso, por construir os sumários com base simplesmente na justaposição das sentenças do ranque, independentemente do método de relevância e de seleção adotados, sabe-se também que os mesmos apresentarão problemas de ordenação das sentenças, pois cada texto-fonte possui uma ordem própria de apresentação de idéias/ fatos (LIMA, PARDO, 2011).

Entretanto, o não tratamento desses dois fatores, redundância e ordenação, e, por conseguinte, de coesão/coerência, justifica-se pelo fato de este trabalho concretizar as primeiras investigações sobre SAMM envolvendo o PB.

A seguir, apresentam-se as tarefas futuras.

3. Etapas futuras

- **Tarefa 5:** <u>Aplicação do método da localização/ Pontuação das sentenças</u>
  Consiste na aplicação do outro método superficial mais difundido da literatura, o da localização. Especificamente, as sentenças de uma coleção seriam também pontuadas e raqueadas em função de sua posição nos textos-fonte.

- **Tarefa 6:** Seleção das unidades textuais/ Produção dos sumários
  Consiste na eleição das sentenças mais bem ranqueadas segundo a localização para compor os sumários até que a taxa de compressão seja atingida. Além desse critério, a seleção das sentenças para compor o sumário poderá ser feita em função, por exemplo, da língua de origem ou do tamanho da sentença, etc.
- **Tarefa 7:** Avaliação dos métodos básicos investigados
  Consiste no julgamento humano dos sumários produzidos pela concatenação das sentenças selecionadas na Tarefa 6 quanto a: (i) informatividade, (ii) coerência e coesão, (iii) inadequações linguísticas, (iv) problemas de tradução, (v) falta de adaptações culturais, etc.
- **Tarefa 8:** Escrita de relatórios e artigos científicos

4. Considerações finais

Quanto às tarefas do projeto em si, ressalta-se que as dificuldades encontradas foram quanto à familiarização aos conceitos da SAMM e sistematização dos dados para descrição e análise. Tais dificuldades foram contornadas por meio da consulta à bibliografia de referência e das reuniões periódicas com a orientadora e co-orientador.

Por fim, ressalta-se que este trabalho tem objetivo final de criar uma *baseline* sobre a SAMM envolvendo o PB, ou seja, um conjunto dos métodos mais simples da literatura com os quais demais trabalhos possam ser comparados.

5. Referências bibliográficas

BAXENDALE, P. B. Machine-made index for technical literature – an experiment. IBM Journal of Research and Development, 2, p. 354-361, 1958.

BLACK, W. J.; JOHNSON, F. C. **Expert systems for information management**: a practical evaluation of two rule-based automatic abstraction techniques. Manchester: Department of Computation. University of Manchester Institute of Science and Technology. Vol.1. N. 3, 1988.

CALDAS, JR. *et al*. Evaluation of a stemming algorithm for the Portuguese language (in Portuguese). In: CONGRESS OF LOGIC APPLIED TO TECHNOLOGY, 2, 2001, São Paulo. **Proceedings…** São Paulo, p. 267-274, 2001.

CAMARGO, R.T.; DI FELIPPO, A.; PARDO, T.A.S. Em direção à caracterização de sumários humanos multidocumento. In: WORKSHOP ON PORTUGUESE DESCRIPTION, 2, 2011, Cuiabá. **Proceedings…** Cuiabá/MT, Brasil, p. 47-54, 2011.

CARDOSO, P. C. F.; PARDO, T. A. S.; NUNES, M. G. V. Métodos para Sumarização Automática Multidocumento usando modelos semântico-discursivos. In: RST BRAZILIAN MEETING, 3, 2011, Cuiabá. **Proceedings…** Cuiabá/MT, Brasil, p. 59-74, 2011.

___.; MAZIERO, E.G.; JORGE, M.L.C.; SENO, E.M.R.; DI FELIPPO, A.; RINO, L.H.M.; NUNES, M.G.V.; PARDO, T.A.S. 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/MT, Brasil, p. 88-105, 2011.

CARROL, J. Parsing. In: MITKOV, R. (Ed.). **The Oxford handbook of computational linguistics**. Oxford, New York: Oxford University Express, 2004, cap. 12, p. 233-248.

CHEN, H.; LIN, C. A multilingual news summarizer. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, 18, 2000, Stroudsburg. **Proceedings…** Stroudsburg, PA, USA, p. 159-165, 2000.

DIAS-DA-SILVA, B. C. *et al.* Introdução ao Processamento das Línguas Naturais e algumas aplicações. **Série de Relatórios Técnicos NILC-TR-07-10**, São Carlos-SP, 2007, 121p.

DIAS-DA-SILVA, B.C.; DI FELIPPO, A.; NUNES, M.G.V. The automatic mapping of Princeton Wordnet lexical-conceptual relations onto the Brazilian Portuguese Wordnet database. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 6, 2008, Marrakech. **Proceedings...** Marrakech, Morocco, 2008.

EDMUNDSON, H. New methods in automatic extracting. **Journal of the ACM,** 16(2), p. 264-285, 1969.

EVANS, D. K. *et al.* Similarity-based multilingual multi-document summarization, **Technical Report CUCS-014-05**, Columbia University,2005. 8p.

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

HATZIVASSILOGLOU, J. L. *et al.* **Simfinder**: a flexible clustering tool for summarization. In: NAACL'01 AUTOMATIC SUMMARIZATIONWORKSHOP, 2001. Pittsburgh, PA, USA. **Proceedings…** Pittsburgh, 2001, 9p.

JORGE, M.L.C.; PARDO, T.A.S. A generative learning approach for Multi-document Summarization using semantic-discursive information. In: BRAZILIAN SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY (STIL), 8, 2011, Cuiabá. **Proceedings…** Cuiabá/MT, Brasil, p. 224-228, 2011.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing**: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. Prentice-Hall: New Jersey, 2009.

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

LAROCCA NETO, J. *et al.* **Generating text summaries through the relative importance of topics**. In: MONARD, M.C.; SICHMAN, J.S. (Eds.), Lecture Notes in Artificial Intelligence, no. 1952. Springer-Verlag, 2000, p. 300-309.

LENCI *et al.* Multilingual summarization by integrating linguistic resources in the MLIS-MUSI Project. In: LREC, 3, 2002, Las Palmas, Canary Islands, Spain. **Proceedings…**. Las Palmas, p. 1464–1471, 2002.

LIMA, J.B.P.; PARDO, T.A.S. Ordenação de sentenças em sumários multidocumento. In: SIMPÓSIO INTERNACIONAL DE INICIAÇÃO CIENTÍFICA DA UNIVERSIDADE DE SÃO PAULO (SIICUSP), 19, 2011, São Carlos. **Anais…**São Carlos/SP, p.1-1, 2011.

LITVAK, M. *et al*, 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, p.61-69, 2010.

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

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

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

MARTINS, C. B. *et al.* Introdução à Sumarização Automática. **Rel. Técnico RT-DC 002/2001**, Departamento de Computação, UFSCar, São Carlos. Abril, 2001. 38p.

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

PAICE, C. D. The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases. In: ANNUAL ACM-SIGIR, 3, 1980, Kent, UK. **Proceedings…** Kent, Butterworth & Co., 1981, p. 171-191.

SALTON, G.; BUCKLEY, C. Term-weighting approaches in automatic text retrieval. **Information Processing & Management,** 24(5), 1988, p. 513–523.

SCHIFFMAN, B. *et al.* Experiments in multidocument summarization. In: **International Conference on Human Language Technology Research**, 2, 2002, San Francisco, CA, USA. **Proceedings…** San Francisco, 2002.

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

SKOROKHOD´KO, E.F. Adaptive method of automatic abstracting and indexing. In: IFIP CONGRESS, 1971, Ljubjana, Yugoslavia. **Proceedings…** Amsterdam. North Holland, 1972, p. 1179-1182.

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

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

____.; GALLIERS, J. Evaluating Natural Language Processing systems: an analysis and review. Berlin: Springer-Verlag, 1996.

PARDO, T. A. S. GistSumm: um sumarizador automático baseado na idéia principal de textos. **Série de Relatórios do NILC. NILC-TR-02-13**. São Carlos-SP, 25p, 2002.

___. GistSumm – GIST SUMMarizer: extensões e novas funcionalidades. **Série de Relatórios do NILC. NILC-TR-05-05**. São Carlos-SP, 8p, 2005.

PARDO, T.A.S. SENTER: um segmentador sentencial automático para o Português do Brasil. **Série de Relatórios do NILC. NILC-TR-06-01**. São Carlos-SP, 6p, 2006.

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**, 6(4), 2010, p. 1-20.