RICI Revista Ibero-Americana de Ciência da Informação

ISSN 1983-5213

Metodologia de acesso a dissertações de mestrado de tradução por estrangeiros, uma abordagem preliminar

Claudio Menezes
Universidade de Brasília, Departamento de Línguas Estrangeiras Aplicadas, Brasília, DF, Brasil
claudiomenezes@unb.br

Dulce Maria Baptista
Universidade de Brasília, Faculdade de Ciência da Informação, Brasília, DF, Brasil
baptistadm368@gmail.com

ARTIGOS

Resumo: O artigo aborda a primeira etapa de uma proposta metodológica para facilitar o acesso de estrangeiros a dissertações de mestrado escritas em português disponíveis em repositórios digitais. Essa etapa consiste na criação e formatação para tratamento computacional dos *corpora* de dissertações, na seleção de um sumarizador automatico da língua portuguesa, no processamento das sumarizações automáticas e na comparação dos sumários assim obtidos com os resumos das dissertações. São também apresentadas e analisadas as métricas obtidas através do programa ROUGE (Recall-Oriented Understudy for Gisting Evaluation). É mencionada igualmente a etapa final do trabalho (que fará uso de tecnologias da língua tais como tradução auxiliada por computador, alinhamento de textos), a partir do *corpus* ora construído, a ser usado para completar o desenvolvimento da metodologia proposta, a qual poderá ser testada em salas de referência de bibliotecas.
Palavras-chave: Dissertações; Métricas; Recall-Oriented Understudy for Gisting Evaluation; Sumarização automatica; Tradução auxiliada por computador.

A methodology for access translating Master dissertation by foreigners, a preliminary approach
Abstract: This paper approaches the first part of a methodological proposal aiming at facilitating the access by foreigners to Master dissertations written in Brazilian Portuguese available in digital repositories. This initial step consists in building up and formatting a dissertations *corpora*, in selecting a computer automatic Portuguese summarizer, in processing the automatic summarization and in comparing the automatic summaries with the abstract of the M, Sc, dissertations (baseline corpus). Some metrics calculated by the computer programm ROUGE (Recall-Oriented Understudy for Gisting Evaluation) are also presented, and an analysis of results is presented. It is also mentioned the final step of this research (that will make use of language technologies such as translation aided by computer, text alignment) to complete the development of the proposed methodology, which may be tested in library reference rooms.
Keywords: Automatic Summarization; Metrics; Recall-Oriented Understudy for Gisting Evaluation; Thesis; Translation.

Una metodoloogía para el acceso a tesis de maestría por extranjeros, un abordaje preliminar
Resumen: Este artículo aborda la primera etapa de una propuesta metodológica para facilitar el acceso de extranjeros a Trabajos de Fin de Maestría (TFM) escritos en portugués y disponibles en repositorios digitales. Esta etapa consiste en la compilación y formateo de los corpus de TFM para su tratamiento computacional, selección de un resumidor automatico de la lengua portuguesa, procesamiento de los

resúmenes automáticos y comparación de los resúmenes generados con los resúmenes de los TFM. También se presentan y analizan las métricas obtenidas con el programa ROUGE (Recall-Oriented Understudy for Gisting Evaluation). Asimismo, se menciona la etapa final del trabajo, que, con la ayuda de tecnologías lingüísticas como programas de traducción asistida y de alineación de textos, utilizará el corpus construido para completar el desarrollo de la metodología propuesta, la cual podrá ser probada en salas de referencia de bibliotecas.
Palabras clave: Maestría; Métricas; Recall-Oriented Understudy for Gisting Evaluation; Resumidor automático; Trabajos de Fin de Maestría; Traducción.

1 Introdução

As dissertações produzidas nos mestrados do Departamento de Línguas Estrangeiras e Tradução (LET), do Instituto de Letras da UnB, são escritas em português, com *abstract* até pouco tempo exclusivamente em inglês[1], mesmo quando se trata de trabalhos sobre outras línguas, como é o caso do espanhol e do francês, Por julgar que o público para tais trabalhos encontra-se também e talvez principalmente entre pesquisadores cuja língua materna não seja a inglesa, imaginamos ser de interesse abordar essa problemática através do uso de sumarização automática, em uma primeira etapa, e de tecnologias da língua, na etapa seguinte, compondo assim uma metodologia integrada. Leitores de língua inglesa também podem se beneficiar de tais tecnologias, particularmente no tratamento de grandes volumes de dados. Indaga-se, portanto, sobre a viabilidade de uso das tecnologias de sumarização automática (e posteriormente de tecnologias da língua, na 2ª etapa da metodologia) para propiciar uma melhor acessibilidade a tais dissertações sob a forma digital.

Nesse contexto iniciamos uma pesquisa cujo objetivo geral será facilitar o acesso de falantes de francês e espanhol aos conteúdos das dissertações de mestrado do Departamento de Língua Estrangeira e Tradução, da Universidade de Brasília.

Os objetivos específicos dessa pesquisa foram assim definidos:

a) Avaliar o funcionamento da sumarização automática para criação de sumários em português utilizando um dos sumarizadores disponíveis (iSummarize, GistSumm, TF-ISF-Summ, NeuralSumm, ClassSumm, SuPor e outros sumarizadores disponíveis);

b) Comparar o resumo obtido pela sumarização automática com o resumo disponível no Repositório Institucional da Universidade de Brasília (RIUnB) por meio de indicadores de precisão, revocação, f-medida e análise de variância;

c) Preparar um *corpus* a ser usado futuramente para testar o uso da tradução auxiliada por computador na criação de sumários em francês, a partir dos sumários automáticos em português e dos resumos das dissertações disponíveis no RIUnB.

O presente artigo trata da primeira parte dessa pesquisa que compreende as seguintes etapas:

[1] As dissertações mais recentes apresentam o *abstract* em francês ou espanhol, em função da sua temática,

1) criação do *corpus* de texto integral e de resumos (com as dissertações de mestrado de tradução em português);
2) escolha de um software de sumarização automática do português;
3) processamento dos arquivos e criação automatizada dos sumários;
4) cálculo dos indicadores (precisão, revocação e medida-F),
5) análise dos indicadores
6) considerações finais
7) referências,

2 Criação dos *corpora*

Definimos como *corpus* para este trabalho as dissertações registradas no *Repositório Institucional de Teses e Dissertações da UnB* (RIUnB) desde a fundação do Programa de Pós-Graduação em Estudos de Tradução da UnB (POSTRAD) até o mês de junho de 2014, correspondendo a um total de 12 trabalhos. Feita a escolha, os arquivos tiveram que ser preparados para a sumarização, por meio da eliminação das seções não pertinentes para esse processamento (apresentação, agradecimentos, referências bibliográficas, etc. Desse modo, foram criados novos arquivos, excluindo todas as seções não diretamente relacionadas à sumarização, mantendo-se exclusivamente no *corpus* o texto das dissertações com interesse semântico (ou seja, sem considerar elementos pré e pós-textuais).

Igualmente, foi efetivada a criação de um novo *corpus* com os resumos de cada dissertação, os quais serão utilizados como sumários de referência (*baseline corpus*).

Vale realçar que esta etapa operacional da preparação dos arquivos está relacionada com o uso de programas de avaliação de sumários, como será explicado em seguida.

3 Escolha de *software* de sumarização

O indiscutível crescimento do volume de dados digitais na sociedade da informação já alcançou números que superam a casa dos *petabytes* (10^{15} *bytes*) e exabytes (10^{18} *bytes*) tornando difícil processá-los com o uso de ferramentas ou aplicações de processamento de dados tradicionais. Vive-se numa sociedade na qual há necessidade de localizar e processar informação da forma mais rápida possível e é nesse contexto que surgiram as técnicas de sumarização automática de textos, cujo início data da década de 1950 com os trabalhos de Luhn (1958), na empresa IBM.

Uma conceituação bastante simples de sumário formulada por Hovy[2005] e que será usada neste trabalho, considera que um sumário é texto produzido a partir de um ou mais

texto(s) cujo tamanho não pode ser superior à metade do tamanho do original. O conceito de texto inclui documentos multimedia, documentos on-line, hipertextos, etc. Diversos tipos de sumário são considerados na literatura: indicativo, informativo (crítico), extratos e *abstracts* (HOVY, 2005). Vale mencionar também que a sumarização está associada ao gênero (Manchetes (*headlines*), Breve descrição (*outlines*), Minutas (*minutes*), Biografias (*biographies*), Abreviações (*abridgments*), Resumo de Filmes (*movie summaries*), Cronologias (*chronologies*), entre outros. Há diversas outras considerações sobre os tipos de sumário que não abordaremos neste artigo, visto que queremos explorar apenas a viabilidade de sumarizar dissertações de mestrado, como etapa inicial de uma metodologia para sua difusão entre estrangeiros. De acordo com a literatura da área, sumários podem ser classificados como informativos, indicativos ou críticos. As abordagens de sumarização, usualmente consideradas em função da quantidade e do nível de conhecimento linguístico que utilizam, são denominadas abordagem superficial e abordagem profunda (em alguns casos pode haver mesclagem de técnicas dando origem a uma abordagem híbrida).

Vale notar que nos trabalhos de sumarização automática, particularmente nos desenvolvidos pela área de linguistica computacional, os conceitos de sumário e de resumo são distintos daqueles usados em Biblioteconomia[2], Lancaster (2004) denomina redação automática de resumos a sumarização automática desenvolvida no campo da linguística computacional.

No entanto, a elaboração automatizada de sumários com a mesma qualidade de sumários elaborados por seres humanos depende de diversos fatores tais como comprimento e gênero do texto fonte, estilo de escrita e uso do léxico. A literatura sobre o assunto traz diversos critérios para a escolha das sentenças que irão figurar no sumário automatico (SA), tais como abordagens linguísticas, estatísticas e centradas na informação ou combinação de ambas. O SA será construído com a seleção de excertos do original portadores da maior ostensividade comunicacional, garantindo dessa forma uma compreensão satisfatória por parte do usuário.

A técnica mais usual de montagem de sumários automáticos é a sumarização por extração por varredura do texto usando técnicas de criação de sumários mediante a seleção de excertos do documento origina. No caso da sumarização humana (SH), se requer a habilidade de entender, interpretar, criar um *abstract* e gerar um novo documento. Na SA, o

[2] CUNHA (2008) define sumário como "apresentação concisa, feita pelo autor do texto ao final de um documento com o objetivo de indicar suas descobertas e suas conclusões, para completar a orientação do leitor que estudou o texto precedente", Para o mesmo autor, resumo é uma "representação concisa e acurada do conteúdo de um documento" e "resumo automático" é "produto da análise computadorizada de um item/documento por meio de programa apropriado",

procedimento é diferente: trata-se de classificar as sentenças do original de acordo com a sua relevância (ou com a sua ostensividade comunicativa) ou probabilidade de compor o "melhor sumário"[3].

Para a seleção do *software* de sumarização foram adotados três critérios básicos: a) disponibilidade para uso sem custo; 2) disponibilidade de funcionalidades para sumarização de textos em português; 3) adequação para sumarização de literatura científica (dissertações de mestrado e teses de doutorado).

Feitas estas considerações de ordem geral sobre a sumarização textual automática, foi realizada uma pesquisa na literatura técnica sobre diversos programas de SA em português: Summarize Thingee (iSummarize), GistSumm, TF-ISF-Summ, NeuralSumm, ClassSumm e SuPor.

O programa iSummarize,disponível em http://www,tools4noobs,com/summarize/help/ apresenta diversas funcionalidades interessantes: a) alimentação do texto fonte na janela do sumarizador ou a partir de URL; b) flexibilidade de argumentos: limite de sentenças (threshold), número de linhas do sumário, comprimento mínimo das sentenças, comprimento minimo das palavras, relevância de cada sentença, palavras mais relevantes, número de palavras mais relevantes, destaque das palavras mais relevantes e das sentenças com as palavras mais relevantes. No entanto, esse sumarizador foi descartado por não haver sido concebido para a sumarização em português, nem se destinado ao gênero de literatura científica.

O SuPor é considerado em alguns testes o melhor sumarizador conhecido para o português (LEITE *et al*,, 2007; RINO *et al*,, 2007). Embora seja um sumarizador enquadrado na categoria de abordagem superficial, utiliza sete métodos de sumarização adaptados para o português (palavra mais frequente, tamanho da sentença, posição da sentença no parágrafo, nomes próprios, cadeias lexicais, importância dos tópicos e mapa de relacionamentos). No presente artigo, a escolha do SuPor foi descartada porque seu uso dependeria de informações fornecidas por um engenheiro do conhecimento (LEITE, D. S.; RINO, L. H. M., 2006), o que dificultaria o desenvolvimento de um trabalho de natureza preliminar de associação de técnicas de sumarização e tradução auxiliada por computado. Além disso, não foi possível obter uma cópia do SuPor nos portais que oferecem programas de sumarização automática gratuitos.

O TF-ISF Summ (**T**erm **F**requency-**I**nverse **S**entence **F**requency-based **Summ**arizer) faz uso de uma métrica desenvolvida por Salton (1988) para classificar as sentenças que serão escolhidas para figurar no extrato. A métrica TF é a frequência do termo em um documento e a

[3] Vale realçar que o conceito de sumário em biblioeconomia (lista em sequência de capitulos e seções) é distinto do que se usa na "sumarização automática" (cujo resultado se compõe de frases do texto original),

ISF é uma função do número de sentenças nas quais esse termo aparece. A exemplo de outros sumarizadores automáticos, o TF-IST Summ efetua a sumarização em três etapas: 1) pre-processamento; 2, classificação das sentenças com base numa métrica; 3) geração do extrato. Resolvemos descartar esse software por não termos conseguido obter uma cópia, além de não haver referência na literatura sobre a sua adequação para sumarização de literatura científica.

O NeuralSumm (LAROCCA NETO *et al*, 2002) usa uma técnica de aprendizagem por máquina, em quatro etapas para produzir o sumário: a) segmentação do texto; b) extração de atributos (comprimento da sentença, posição da sentença no texto, posição da sentença no parágrafo ao qual ela pertence, presença de palavras chave na sentença, presença das palavras relevantes (*gist words*) na sentença, indicador de relevância da sentença com base na frequência das palavras, indicador de relevância da sentença com base na métrica TF-ISF e presença de palavras indicativas na sentença; c) classificação das sentenças por atributo e d) produção do sumário. A técnica de aprendizagem por máquina faz uso de um mapa construído a partir do próprio texto: o SOM, "self-organizing map".

O GistSumm (RINO *et al.*, 2004) é baseado na escolha da sentença mais importante do texto, a "sentença-gist". Na pesquisa realizada encontramos algumas variantes da aplicação da metodologia do GistSumm, mas de um modo geral o sumarizador obtem o extrato da seguinte forma: a) através de estatísticas simples a sentença-gist ou uma aproximação dela é obtida; b) por meio dessa sentença, são obtidos extratos coerentes que incluem essa sentença e outras sentenças complementares. O Relatório técnico "Sumarização Automátrica de Textos Científicos: Estudo de Caso com o Sistema GistSumm" (BALAGE FILHO *et al.,* 2007) apresenta diversas experiências utilizando esse sumarizador para literatura científica. Em tais experiências, tanto houve variação no critério de escolha da sentença-*gist*, como na seleção de sentenças para figurar no extrato. As experiências também fizeram uso de *corpora* constituidos por teses e dissertações. Por tais razões, o GistSumm foi o sumarizador escolhido para o experimento com as Dissertações do Mestrado de Tradução da UnB.

4. Processamento dos arquivos e criação automatizada dos sumários

Os originiais das dissertações de mestrado do POSTRAD estão disponíveis no Repositório Institucional da Universidade de Brasília (http://repositorio,unb,br/handle/10482/14608), em formato PDF. No entanto, tanto o formato do documento como alguns elementos componentes capítulos das dissertações são inadequados para sumarização automática, como é o caso de agradecimentos, referências, lista de tabelas, entre outros. Desse modo, foi efetuado um trabalho de reformatação e pré-processamento manual para obter os *corpora* a serem usados no processo de sumarização.

A criação automatizada dos sumários foi processada usando o GistSumm e compondo um *corpus* a ser usado no cálculo das métricas, como se descreve a seguir. A escolha da taxa de compressão foi feita de forma a atender a definição de sumário de Hovy (2003), no valor de 80%.

5. Cálculo dos indicadores

As métricas usuais para avaliar a qualidade de sumários apresentadas pela literatura da área são a precisão, revocação, medida=F, assim definidas:

Precisão (P) = Correct / (Correct + Wrong), isto é, o número de sentenças do sumário automático presentes no sumário de referência em relação ao número de sentenças do sumário de referência ("baseline").

Revocação ou Cobertura (C) = Correct / (Correct + Missed), isto é, o nº de sentenças do sumário automático presentes no sumário de referência ("baseline") em relação ao número de sentenças do sumário automático.

Medida-F = 2*(Precisão*Revocação)/(Precisão + Revocação) (balanço métrico entre a Precisão e a Cobertura).

(calculados a partir de um texto original de entrada e 2 extratos construídos pelo computador e por uma pessoa), onde:

Correct = nº de sentenças comuns aos 2 extratos.

Wrong = nº de sentenças do sumário automatizado ausentes do sumário elaborado pelo ser humano.

Missed = nº de sentenças extraídas pelo humano mas ausentes do sumário automatizado.

A medida-F indica a qualidade de cada sumário automático. À medida que se aproxima de 1, a medida-F evidencia que foram obtidos sumários de melhor qualidade e valores próximos de zero se referem a sumários de baixa qualidade em relação ao texto de referência.

Para obter os valores das métricas foi utilizado o programa ROUGE, pacote de avaliação baseado na co-ocorrência de n-gramas entre textos que se deseja comparar. Essa ferramenta permite analisar os sumários automáticos com relação aos resumos desenvolvidos pelos autores de cada dissertação de mestrado, considerados como sumários ideais. Os resultados são apresentados a seguir.

6 Análise dos Indicadores

O experimento de cálculo dos indicadores foi efetuado mediante uso do programa "Recall-Oriented Understudy for Gisting Evaluation" (ROUGE), Elaborado por Lin (2004), esse

programa permite medir a qualidade de um sumário (em geral, um SA) por meio de uma comparação com um sumário (ideal) criado pelos autores. Para cálcular tais métricas, o ROUGE conta o número de superposições dos textos sob comparação, tais como n-gramas, sequência de palavras e pares de palavras dos sumários. Há diversas outras considerações sobre o uso desse tipo de programa que, no entanto, escapam ao escopo deste artigo.

O resultado do processamento do programa ROUGE para os sumários do *corpus* sob estudo encontra-se na tabela 1.

Tabela 1 - Avaliação dos Sumários Automáticos das Dissertações de Mestrado do Programa POSTRAD, da UnB

N, da Dissertação	Precisão (P)	Cobertura (R)	Medida F
1	0,87540	0,02295	0,04473
2	0,79600	0,03379	**0,06483**
3	0,79116	0,02730	0,05278
4	0,79600	0,01729	0,03384
5	**0,59424**	0,01582	0,03082
6	0,75188	0,03067	0,05894
7	0,72826	0,01331	0,02614
8	0,89744	0,02354	0,04588
9	0,81215	0,01832	0,03583
10	0,74915	**0,04138**	0,07843
11	**0,90773**	0,04131	0,07902
12	0,79825	**0,01294**	**0,02547**
Média Aritmética	0,79147	0,024885	0,04806

Observa-se na tabela que:

a) a precisão P varia entre 0,90773 e 0,59424, A media aritmética das 12 observações é de 0,79147;

b) a cobertura R varia entre 0,01294 e 0,04138, A media aritmética das 12 observações é de 0,024885; e

c) a medida F varia entre 0,02547 e 0,06483, A media aritmética das 12 observações é de 0,04806,

Os dados acima, particularmente quanto ao indicador P cuja média aritmética foi de

aproximadamente 80%, nos permitem considerar aceitável o *corpus* de sumários automáticos para a etapa seguinte da metodologia: a tradução auxiliada por computador.

7 Considerações finais

Nesta primeira etapa, foi realizada a criação dos *corpora*, a escolha e uso de programas de sumarização e cálculo de métricas e a análise dos indicadores de qualidade dos sumários automáticos, como descrito neste artigo.

A pesquisa sobre a metodologia ainda requer novas elaborações e aperfeiçoamentos. Seria interessante, por exemplo, explorar a idéia da criação de sumários automatizados com o mesmo número de parágrafos dos resumos elaborados pelos autores das dissertações. Outra necessidade será buscar um processo de automatização na criação dos arquivos, visto que a exclusão das sessões sem relevância para a sumarização foi feita por um processo manual. Como tais funcionalidades não estão disponíveis nas ferramentas utilizadas, poderão ser objeto de trabalhos a serem empreendidos na continuação da pesquisa ora iniciada.

Quando à etapa seguinte – uso da tradução auxiliada por computador e de outras tecnologias da língua – há uma série de considerações ainda a serem estudadas. O estabelecimento de critérios para escolha de um tradudor automatico adequado, a elaboração de interfaces entre os arquivos das diversas etapas da metodologia, a construção de indicadores da qualidade da tradução automatica são algumas das questões a serem examinadas no prosseguimento desta pesquisa.

Desse modo, e cumpridas todas as atividades previstas, será possível completar a construção da metodologia proposta, a qual tem como objetivo viabilizar o acesso às dissertações em português por estrangeiros de forma simples, rápida e eficaz e que poderá ser testada em salas de referência de bibliotecas.

Referências

BALAGE FILHO, P. P.; PARDO, T. A. S.; NUNES, M. G. V. Summarizing Scientific Texts: Experiments with Extractive Summarizers. In: INTERNATIONAL CONFERENCE ON INTELLIGENT SYSTEMS DESIGN AND APPLICATIONS – ISDA, 7th, Rio de Janeiro, Brazil, October, 22-24, 2007. **Proceedings.** Rio de Janeiro, 2007. P. 520-524.

CUNHA, M. B. da; CAVALCANTI, C. R. de O. **Dicionário de biblioteconomia e arquivologia.** Brasília: Briquet de Lemos / Livros, 2008. ISBN 978-85-85637-35-4

HOVY, E. Text Summarization. **The Oxford Handbook of Computational Linguistics,** January 2005, Chapter 32, p. 583-98.

LAROCCA NETO, J.; FREITAS, A. A.; KAESTNER, C. A. A. **Automatic text summarization using a machine learning approach.** In: BRAZILIAN SYMPOSIUM ON ARTIFICIAL INTELLIGENCE, 16. **Lecture Notes in Artificial Intelligence**, v. 2057, p. 205-215, 2002.

LANCASTER, F. W. **Indexação e resumos**: teoria e prática. 2ª ed. Brasília, DF: Briquet de Lemos Livros, 2004. ISBN 85-85637-24-2

LEITE, D. S.; RINO, L. H. M.; PARDO, T. A. S.; NUNES, M. G. V. Extractive Automatic Summarization: Does more linguistic knowledge make a difference? In: BIEMANN, C. I.; MATVEEVA, R. Mihalcea; RADEV, D. (eds.). **Proceedings of the HLT/NAACL Workshop on TextGraphs-2: Graph-Based Algorithms for Natural Language Processing.** Rochester, NY: 2007.

LIN, Chin-Yew. ROUGE: A Package for Automatic Evaluation of Summaries. In: WORKSHOP ON TEXT SUMMARIZATION BRANCHES OUT, Post-Conference Workshop of ACL 2004, Barcelona Spain. **Proceedings**. Barcelona, 2004.

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

RINO, L. H. M.; PARDO, T. A. S.; SILLA JR, C. N.; KAESTNER, C. A. A.; POMBO, M. **A** Comparison of Automatic Summarizers of Texts in Brazilian Portuguese. Advances in Artificial Intelligence – SBIA 2004. **Lecture Notes in Computer Science**, v. 3171, p. 235-244, 2004.

SALTON, G.; BUCKLEY, C. Term-weigthing approaches in automatic text retrieval. **Information Processing and Management,** v. 24, p**.** 513-523, 1988.

Recebido/Recibido/Received: 2015-10-05
Aceitado/Aceptado/Accepted: 2017-03-21