Universidade de Brasília
Instituto de Letras – IL
Departamento de Línguas Estrangeiras e Tradução – LET
Línguas Estrangeiras Aplicadas ao Multilinguismo e à Sociedade da Informação

LARISSA DE OLIVEIRA RIBEIRO

SUMARIZAÇÃO AUTOMÁTICA: UMA FERRAMENTA PARA O TRATAMENTO AUTOMÁTICO DAS LÍNGUAS NATURAIS

Brasília, 2016

LARISSA DE OLIVEIRA RIBEIRO

SUMARIZAÇÃO AUTOMÁTICA: UMA FERRAMENTA PARA O TRATAMENTO AUTOMÁTICO DAS LÍNGUAS NATURAIS

Trabalho de Conclusão de Curso apresentado à Universidade de Brasília como exigência parcial para obtenção de título de bacharela em Línguas Estrangeiras Aplicadas ao Multilinguismo e à Sociedade da Informação.

Orientador: Me. Francisco Cláudio Menezes

Brasília

2016

LARISSA DE OLIVEIRA RIBEIRO

SUMARIZAÇÃO AUTOMÁTICA: UMA FERRAMENTA PARA O TRATAMENTO AUTOMÁTICO DAS LÍNGUAS NATURAIS

Trabalho de Conclusão de Curso submetido à banca examinadora abaixo identificada, como requisito parcial para obtenção do grau de bacharela em Línguas Estrangeiras Aplicadas ao Multilinguismo e à Sociedade da Informação.

Brasília,_____/_____/_____

BANCA EXAMINADORA

_____

Prof. Me. Francisco Cláudio Menezes
Depto. De Línguas Estrangeiras
e Tradução – LET/UnB

_____

Prof. Dra. Fernanda Alencar Pereira
Depto. De Línguas Estrangeiras e
Tradução – LET/UnB

_____

Prof. Dr. Cláudio Corrêa e Castro Gonçalves
Depto. De Línguas Estrangeiras
e Tradução – LET/UnB

AGRADECIMENTOS

Agradeço primeiramente a Deus por ser meu alicerce e por me ajudar em minha caminhada, agradeço aos meus familiares por sempre me apoiarem de forma inquestionável, agradeço ao meu orientador, Francisco Cláudio, por ter confiado à mim esse trabalho e pelo incentivo e apoio em cada etapa, aos meus professores por todo o conhecimento transmitido durante esses anos de graduação e aos meus amigos que sempre estiveram ao meu lado me incentivando a continuar.

SUMARIZAÇÃO AUTOMÁTICA: UMA FERRAMENTA PARA O TRATAMENTO AUTOMÁTICO DAS LÍNGUAS NATURAIS

Larissa de Oliveira Ribeiro[2]

Resumo: Esta pesquisa explora a funcionalidade e aplicações da sumarização automática. Dois sistemas de sumarização foram escolhidos para permear o assunto, sendo o primeiro o Gist Summarizer, primeiro sistema de sumarização disponível para a língua portuguesa e o segundo, o sistema de sumarização automática SweSum, primeiro sumarizador automático para a língua sueca. O estudo desses sistemas explora o processo de sumarização automática e a estrutura interna dos sistemas de sumarização automática,como eles realizam a sumarização e como esses sumários podem auxiliar usuários web. Após esse estudo, sugere-se certas adequações a esses sistemas que podem auxiliar na sua usabilidade, particularmente na web. Por último, analisa-se o processo de sumarização automática realizado pelos sistemas de sumarização citados, com o intuito de sugerir um software auxiliar que filtra sessões de textos acadêmicos, para que se otimize esse processo e os sumários sejam disponibilizados com maior precisão e personalizados de acordo com o interesse do usuário.

Palavras-chave: Sumarização; Sumarização automática; Processamento de Linguagem Natural; Usabilidade;

---

[1] Trabalho de Conclusão de Curso orientado por Francisco Cláudio Menezes, professor e orientador do Departamento de Línguas Estrangeiras e Tradução - LET/UnB. E-mail:claudiomenezes@unb.br

[2] Graduanda do Bacharelado em Línguas Estrangeiras Aplicadas ao Multilinguismo e à Sociedade da Informação – LEA-MSI - , da Universidade de Brasília – UnB. E-mail:larissa16ribeiro46@gmail.com.

AUTOMATIC SUMMARIZATION: A TOOL FOR THE AUTOMATIC TREATMENT OF NATURAL LANGUAGES

Summary: This research explores the functionality and applications of automatic summarization. Two summarization systems were chosen to permeate the subject, the first being the Gist Summarizer, the first summarization system available for the Portuguese language, and the second, the automatic summarization system SweSum, the first automatic summarizer for the Swedish language. The study of these systems explores the automatic summarization process and the internal structure of automatic summarization systems, how they perform summarization, and how these summaries can assist web users. After this study, it is suggested certain adaptations to these systems that can aid in the its usability, particularly within the web. Finally, it is analyzed the automatic summarization process performed by the summarization systems mentioned, with the purpose of suggesting an auxiliary software that filters sessions of academic texts, to optimize this process and the summaries made available with more precision and personalized according to the user's interest.

Keywords: Summarization; Automatic summarization; Natural Language Processing; Usability;

LISTA DE ILUSTRAÇÕES

Figura 1 - Etapas da SA (Pardo 2002).....20
Figura 2 - Arquitetura GistSumm 23
Figura 3 - Interface GistSumm 24
Figura 4 - Interface GistSumm 25
Figura 5 - Interface SweSum 27
Figura 6 - Campos de entrada SweSum 27
Figura 7 - Campo de entrada compressão e botão de *summarize*.....28
Figura 8 - Campos de entrada de texto do SweSum 29
Figura 9 - Texto Fonte para sumarização automática 31
Figura 10 - Texto Sumarizado pelo GistSumm 32
Figura 11 - Texto Traduzido pelo Tradutor Automático *Google Translate*.....33

LISTA DE SIGLAS

ABNT- Associação Brasileira de Normas Técnicas

BDTD – Biblioteca Digital Brasileira de Teses e Dissertações

FINEP – Financiadora de Estudos e Pesquisas

HTML – HyperText Markup Language

IA – Inteligência Artificial

IBICT – Instituto Brasileiro de Informação em Ciência e Tecnologia

NILC – Núcleo Interinstitucional de Linguística Computacional

PLN – Processamento de Linguagem Natural

RST – Rethorical Sentence Theory

SSA - Sistema de Sumarização Automática

SA - Sumarização Automática

SAM – Sumarização Automática Multidocumento

LISTA DE TABELAS

Tabela 1 - Lista de Stop Sessions 35

SUMÁRIO

INTRODUÇÃO 10

1. CONCEITOS BÁSICOS DE SUMARIZAÇÃO AUTOMÁTICA 12

1.1 Abordagem Superficial x Abordagem Profunda 13

1.1.1 Abordagem Superficial 13

1.1.2 Abordagem Profunda 15

1.2 Teoria RST 17

1.3 Sumarização Automática Monodocumento x Sumarização Automática Multidocumento.....18

1.3.1 Sumarização Automática Monodocumento 18

1.3.2 Sumarização Automática Multidocumento (SAM) 18

1.3.2.1 CSTNews 19

2. PROCESSO DE SUMARIZAÇÃO AUTOMÁTICA 20

3.GIST SUMMARIZER 21

3.1 Ambiente Virtual do Sumarizador Automático Gist Summarizer 23

4.SWESUM.....25

4.1 Ambiente Virtual do Sumarizador Automático SweSum 26

5. APLICAÇÕES DA SA 29

5.1 Textos Jornalísticos 29

5.2 Tradução Automática 29

5.3 Textos Acadêmicos 33

6. PROPOSTA DE FILTRO AUXILIAR PARA A SA 34

METODOLOGIA 37

CONSIDERAÇÕES FINAIS 40

REFERÊNCIAS BIBLIOGRÁFICAS 41

GLOSSÁRIO 43

INTRODUÇÃO

O Dicionário *Priberam da Língua Portuguesa* traz várias definições de sumário, mas como principal e condizente com o assunto apresentado no presente trabalho escolhemos a seguinte:*"Texto ou relato com os aspectos principais:epítome,recapitulação,resumo,suma"*. Dentro do estudo de sumarização automática, o termo sumário, representa o texto fonte, elaborado a partir de sentenças escolhidas segundo alguns critérios.

A sumarização automática (SA) é uma subárea do Processamento de Linguagem Natural (PLN). Trata-se da produção automática de sumários de apenas um ou vários textos-fonte. Dessa forma, o sumário é um resumo, cujo objetivo é captar a ideia principal do autor e passá-la em poucas linhas para o leitor (GUELPELI, 2012).

Em geral,os resumos são utilizados com frequência quando se quer, por exemplo, narrar algum acontecimento, onde nunca é contado exatamente o que aconteceu, mas sim um resumo do ocorrido. É através dos resumos que nós, humanos, também podemos escolher um livro, um filme, um produto eletrodoméstico, acessar uma página web, entre outros, sendo assim utilizados como facilitadores de tomada de decisões. Além disso, os resumos podem ser usados na questão de inclusão digital, por exemplo, quando um texto em português é simplificado em determinado site, para estrangeiros com pouco domínio da língua portuguesa.

É importante definir a diferença entre resumos, sumários e sumários automáticos.Os resumos, para fins de compreensão do presente trabalho, são os textos feitos pelo próprio autor do texto-fonte, que pode ou não conter fragmentos desse texto. Os sumários, também chamados de índice,são aqueles produzidos pelo autor, contendo capítulos,números de paginas, etc.Por último, os sumários automáticos, aqui são os textos de saída produzidos pelos sistemas de sumarização automática (SSA) ou sumarizadores automáticos, poderão ser chamado também, apenas de sumário quando se referir ao texto de saída do SSA. Segundo Pardo (2008, P.1), "Os sumários comumente chamados resumos, podem ser produzidos por diversas estratégias e pelo uso de conhecimentos de naturezas diversas".

Com o avanço da tecnologia e a grande quantidade de informações disponíveis, inclusive no ambiente web, fez-se necessário o aprimoramento da técnica de sumarização. Antes, tal técnica era utilizada apenas por humanos e não por computadores, o que acarreta uma série de problemas. Os resumos para o mesmo texto sempre são distintos, porque cada indivíduo os elabora segundo a sua percepção e ordem de relevância do assunto dentro do texto. Sendo assim os resumos não são iguais e é difícil eleger qual deles seria considerado “o melhor sumário”.

A SA, veio como uma tentativa para solucionar tais problemas. Ela vem sendo explorada desde a década de 50, quando ainda era utilizado como método principal, o método de palavra chave. Por um certo período, as pesquisas a cerca de sumarizadores automáticos ficaram estagnadas, mas com o crescimento de arquivos on-line, foram retomadas.

Ainda é discutido se os textos produzidos pelos sumarizadores automáticos têm ou não a mesma qualidade que textos de produção humana, mas pode-se afirmar que eles auxiliam na redução da quantidade de tempo que seria utilizada por uma pessoa para fazer uma escolha sobre determinado assunto.De acordo com Pardo (2008), as pessoas têm menos tempo para processar as informações que são necessárias.Se fossemos contar apenas com os sumarizadores de produção humana, uma boa parte dos assuntos disponibilizados atualmente não seriam acessados, porque não seria possível criar resumos em larga escala, já que a demanda é imensa.

Atualmente, existem SSA capazes de produzir bons sumários. Pode-se citar o sumarizador da língua portuguesa, o Gist Summarizer, que foi o primeiro sistema de SA disponível para o português. Outros sumarizadores, como SweSum, produzido inicialmente para a língua sueca, também foram desenvolvidos para otimizar o trabalho de sumarização.

Os sumarizadores automáticos tanto da língua portuguesa quanto de outras línguas são de extrema importância. Esses sistemas de sumarização automática são criados com base em algumas regras que dependem do gênero do texto fonte.Com o passar do tempo e o avanço das pesquisas, essas regras ou métodos vão se aprimorando, para que assim o texto produzido seja bem aproveitado pelo usuário, servindo como representação adequada do texto fonte.

Os textos de saída, ou seja, os sumários automáticos produzidos por esses sistemas, podem funcionar como auxiliadores para pessoas que não dominam a língua principal de um determinado país, no contexto apresentado, a língua portuguesa. Com os textos em tamanho menor, o individuo que não possui o domínio da língua, conseguirá entender com muito mais facilidade as ideias essenciais em um determinado texto,dispensando a leitura de todo o conteúdo.Podem também, por exemplo, utilizar sistemas de tradução automáticos, o que permitirá a um usuário trabalhar com um texto mais curto.

Com base nas informações apresentadas, essa pesquisa tem como objetivo, analisar e descrever o funcionamento dos sumarizadores automáticos Gist Summarizer e SweSum,e os critérios para realizar a sumarização e o aproveitamento do texto de saída. Serão analisados quais pontos poderiam ser mudados para que o sistema fosse otimizado.Como resultado da pesquisa, será sugerido um modelo de filtro do texto fonte como auxílio para o funcionamento

dos sumarizadores automáticos e será mostrado como o texto de saída seria utilizado para o auxilio de não falantes da língua portuguesa.

1. CONCEITOS BÁSICOS DE SUMARIZAÇÃO AUTOMÁTICA

A SA é uma subárea da PLN. O Processamento de Linguagem Natural por sua vez é um subcampo da Inteligência Artificial (IA). Esse tipo de processamento se apoiando conhecimento de computação e linguística.

Os computadores já compreendem tipos de linguagens computacionais, criadas para a interação homem-máquina. São dados comandos e instruções ao computador para que este execute uma determinada tarefa. Essas linguagens, chamadas de linguagens de programação, utilizam uma sintaxe própria, porém estas pouco se aproximam da língua natural, que é a língua utilizada pelos humanos.

A partir dessa questão, foram propostos estudos sobre métodos de aproximação da linguagem computacional à língua natural. A PLN pretende que os computadores possam interagir com as pessoas, como se fossem outros seres humanos, com capacidade cognitiva.

A sumarização automática, que está dentro desse campo, segue a função de auxiliar a interação do homem com a máquina, trazendo benefícios para o ser humano, através de seus serviços. A SA produz sumários a partir de textos escolhidos em *corpus* linguísticos, inseridos pelo usuário no computador.

Segundo Pardo (2008), os sumários produzidos pela SA podem ser classificados quanto à função,audiência e formação. Quanto à função, podem ser classificados como indicativos, informativos ou críticos. Os sumários indicativos não substituem o texto fonte, eles apenas expõem o que o texto trata.Já os sumários informativos contêm as informações mais relevantes do texto fonte de forma coerente e coesa. Esses sumários precisam ter uma boa progressão temática, serem gramaticais e legíveis,atribuindo textualidade aos textos. Por último, os sumários críticos apresentam opiniões além do conteúdo do sumário.

Quanto à audiência os sumários se classificam entre focado nos interesses do leitor e genéricos. Os sumários que são focados nos interesses dos leitores, como a própria denominação já explicita, trazem informações customizadas em relação ao conhecimento do leitor. Já os genéricos, trazem as informações mais relevantes do texto,sem se preocupar com os leitores.

Quanto à formação, os sumários se classificam em *abstracts* e extratos.Os *abstracts* são trechos do texto original, com pequenas alterações, que ocorrem porque são reescritos.O extrato por sua vez, nada mais é que um sumário com trechos idênticos ao do texto fonte, um

“copia e cola” do texto original selecionado segundo critérios de sumarização dependentes do gênero textual.

Em relação ao nível linguístico utilizado pelo sumário, temos duas abordagens: a abordagem superficial e a abordagem profunda (PARDO,2008). Se tivermos as duas abordagens em um mesmo sumário, falamos em abordagem híbrida.

É importante uma breve explicação sobre os níveis de conhecimento que podem ser utilizados na sumarização automática. Os níveis são: morfologia, sintaxe, semântica, pragmática e discurso.

> Quanto mais se sobe da morfologia em direção à pragmática e ao discurso, maior é a abstração linguístico-computacional e mais difícil se torna obter uma representação formal do nível de conhecimento e, por conseguinte, mais complexo é processar computacionalmente tal nível. (PARDO,2008, P. 4)

Na morfologia, estuda-se a formação das palavras isoladamente. É o trabalho com radicais, prefixos e sufixos, derivações e desinências, etc. Na sintaxe, são analisadas as funções das palavras dentro de uma frase, se são sujeito,predicado,complemento, objetos, etc. A semântica lida com o significado da palavra, da sentença e do texto. Por fim, tanto a pragmática quanto o discursocuidam da organização das informações dentro do texto, de forma que se mantenha a textualidade, a coesão e a coerência. Aspectos como o contexto em que o texto foi produzido e o meio em qual é veiculado, também são contemplados.

Com base nessa breve explicação, pode-se entrar no campo da abordagem superficial e profunda.

1.1Abordagem Superficial x Abordagem Profunda

1.1.1 Abordagem Superficial

A abordagem superficial ou empírica faz pouco ou quase nenhum uso do conhecimento linguístico, se resume apenas ao uso de dados morfossintáticos e sintáticos Essa abordagem costuma produzir extratos. A seguir, serão mostrados alguns métodos utilizados pela abordagem superficial para produzir os sumários (MARTINS et al, 2001), e uma breve explicação sobre cada um.

Métodos das Palavras-chave: Esse método tem como objetivo o resumo por meio das palavras que aparecem repetidamente no texto. Pressupõe-se que essas palavras, ou palavras-chave, estão em sentenças que carregam a ideia principal do texto. Então, o sumário é feito com base nessas sentenças, e o texto é produzido na ordem em que essas sentenças aparecem

no texto-fonte. Alguns programas são feitos com base nesse método, considerando entre as palavras-chave, apenas as de classe aberta, que são aquelas que possuem um significado concreto.

Existe também a hipótese de que a maioria das palavras-chave sejam substantivos, partindo dessa ideia, é possível concluir que os substantivos são responsáveis pela ideia central do texto. Segundo Martins et al (2001):

> Na sentença: "A casa desmoronou", a palavra "casa" seria mais representativa que o verbo “desmoronou”, porque indicaria o tópico sobre o qual se fala. Caso o autor decidisse mudar o foco para o evento, propriamente dito, introduzido nesse exemplo pelo verbo “desmoronar”, seria necessário nominalizá-lo e, portanto, o substantivo continuaria sendo a forma-chave mais representativa do conteúdo textual.Assim, uma possível progressão temática mudando o foco de “casa” para “desmoronar” seria construída, por exemplo, pela sentença “O desmoronamento ocorreu devido aos ventos fortes”. (Martins et al, 2001, P 10)

Método das palavras-chave do título: Esse método utiliza apenas o titulo do texto fonte para procurar as palavras chaves e fazer um sumário com as principais informações do texto.Se o texto tiver subtítulos e sumários, estes também serão utilizados para que se extraia as palavras-chave. Parte-se do pressuposto, que um bom título é composto por palavras que trazem a ideia principal do texto. A partir disso têm-se que as sentenças que contêm as palavras-chave que aparecem no título, carregam informação principal do texto.

Método da localização: A posição de uma sentença no texto,pode estar associada a sua relevância. Por exemplo,em grande parte dos casos a primeira sentença e a última de um parágrafo, traz as informações importantes desse fragmento. Baseado nesse método, a primeira e a última sentença de cada parágrafo, devem ser incluídas no sumário.

Método das palavras sinalizadoras: Similarmente aos métodos de palavras-chave e palavras-chaves nos títulos, o método de palavras sinalizadoras, conta com um dicionário feito previamente, com as palavras mais relevantes do texto, podendo variar entre os gêneros textuais.As sentenças que contêm as palavras presentes no dicionário, recebem uma pontuação maior, aquelas que não contêm palavras que estão no dicionário, recebem pontuação negativa. Dessa maneira, as sentenças são organizadas dentro do sumário, de acordo com a pontuação recebida. Os sistemas de extração automática de frases-chave (“*Key - phrase extraction systems*”) pode ser considerado uma variante desse método.

Método relacional: Trata-se de um método no qual, é utilizado o valor semântico das sentenças.Uma sentença que possui muitas relações com outras, recebe uma pontuação

elevada e pode ser escolhida para o sumário. Essa relação semântica é percebida, por exemplo, pelos substantivos presentes em comum com outras sentenças.

Método da Frase Auto Indicativa: Este último método define a importância das sentenças pela sua estrutura. Uma frase é auto indicativa, quando a sua estrutura ocorre várias vezes no texto. Elas sempre indicam o conteúdo principal do texto, apontando para as partes importantes. Por exemplo, *O objetivo desse artigo é...,Nesse trabalho é descrito...,*etc. Com esse método é possível construir sumários indicativos, aqueles que identificam o tópico do texto, mas não é possível, produzir sumários que discorrem a cerca do assunto do texto fonte.

Esses métodos da abordagem superficial trazem resultados bons, mas os sumários da abordagem superficial, na maioria das vezes não são produzidos de uma maneira bem estruturada. Uma vez que a abordagem superficial não se atém a importância de um conteúdo para o sumário, muitas sentenças e trechos importantes acabam sendo perdidos.

1.1.2 Abordagem Profunda

A abordagem profunda, por sua vez, usa dos conhecimentos linguísticos e utiliza a estrutura da língua. Nessa abordagem, é procurada a proposição central,ou seja, as proposições complementares do texto-fonte que poderão auxiliar na compreensão do que o autor quer passar, de forma coerente e coesa.

Os sumários que são produzidos a partir da abordagem profunda são os *abstracts*. Abstracts são sumários elaborados pelo autor do texto e que apresentam partes do texto original, mas ao contrário dos extratos produzidos pela abordagem superficial, eles são reescritos, ou seja, parte da estrutura é alterada, e possivelmente também seu significado.

Comparando-se a abordagem superficial com a profunda, é perceptível que a abordagem profunda se sobressai em qualidade. Esta última usa de elementos linguísticos, o que não é utilizado na abordagem superficial. A partir dos elementos linguísticos é possível analisar o texto no seu significado principal de acordo com contexto em que é apresentado e transmitir a ideia principal no sumário de uma forma coerente e coesa. Em outras palavras, não serão mais apenas trechos escolhidos “no escuro”, mas excertos do original analisados de forma lógica.

Por exemplo, métodos como o da frase autoindicativa podem não ser precisos, pois não pode se afirmar com certeza, que o conteúdo principal do texto e a ideia original estão sempre contidos nas frases que ocorrem várias vezes no texto. Essas frases podem ter somente o objetivo de dar inicio a um parágrafo.

De acordo com Rino (1996), a abordagem profunda tem como objetivo, identificar no texto fonte,qual é a sua proposição central, quais as informações complementares à proposição encontrada que podem auxiliar a transmissão da mensagem no sumário e a preservação da ideia original do texto-fonte e permitir a estruturação do sumário com base em normas de coerência e coesão, para que o resultado seja um texto no sentido linguístico.

É importante salientar que na abordagem profunda existe uma dependência da mensagem principal do texto-fonte. Todo o sumário está relacionado com essa mensagem fonte. É justamente nesse momento que começam os problemas na abordagem profunda, o que às vezes desencoraja seu uso, quando comparada com a abordagem superficial.

Essas características têm a ver com a percepção do leitor sobre clareza ou não do sumário. Nem sempre um sumário que tem pontos bem explícitos passa necessariamente clareza para o leitor. Isso vai depender do conhecimento do leitor sobre linguística, gêneros textuais e tipologia.

Gênero textual e tipologia são assuntos um pouco delicados, mas extremamente necessários quando estudamos sumarização automática. Para cada gênero e tipo textual há um processamento adequado. Por exemplo, se o texto é jornalístico, é necessário que o sumarizador automático utilize um tipo especifico de processamento para que sejam gerados sumários de boa qualidade. Segundo Martins et al (2001, P.23): “A consideração do gênero e tipo de textos na sumarização automática pode ser de grande valia para ajudar a reconhecer a estrutura textual e a especificar um modelo de processamento automático correspondente”.

Gêneros textuais podem ser definidos como a escolha para classificar o uso da língua em um determinado tipo de comunicação. Alguns exemplos de gêneros textuais são os debates, relatórios científicos, seminários, biografias, carta, artigo ou teses.

Já o tipo textual, que não deve ser confundido com gênero textual, é a forma como o texto é colocado, tem a ver com a finalidade para o qual ele foi escrito. Exemplos de tipo textual são os textos narrativos, expositivos, informativos ou analíticos.

Os gêneros textuais e os tipos devem ser considerados para que sejam criados sumários de qualidade, que possam satisfazer a busca do leitor. Para isso é necessário que o sumário seja coeso e tenha também coerência.

Coesão e coerência, apesar de sempre andarem juntas, têm significados distintos e dentro da SA, são utilizadas para diferentes finalidades. Coesão é a ligação das ideias dentro do texto, é a relação entre os parágrafos, que fazem com que eles fiquem dispostos de uma forma lógica. Dentro da SA, a coesão refere-se à abordagem superficial.

Entretanto, a coerência apesar de também ser ligada a estrutural textual, refere-se à lógica interna do texto. O assunto que é abordado no texto não pode ser modificado ao longo do desenvolvimento,deve ser objetivo e linear, para que as ideias sejam claras para o leitor. A coerência, dentro da sumarização realizada por máquinas, está relacionada à abordagem profunda.

As propriedades coesivas geralmente têm a ver com os marcadores do discurso. Entende-se por marcadores do discurso as palavras que guiam a interpretação do texto, relacionando as ideias dentro dele. Alguns exemplos de marcadores são: portanto, contudo, inclusive e por conseguinte. Os marcadores, por auxiliarem o mantimento das relações lógicas dentro do texto, estão relacionados a abordagem profunda. Por isso o nível de dificuldade de utilizar a abordagem profunda nos sumarizadores é maior. Exemplos como os dos marcadores são lógicos e intuitivos para os seres humanos, mas muito difícil de manipular dentro de um programa de computador.

Para dar continuidade ao conceito de abordagem profunda, é necessário conhecer um pouco da teoria RST.

1.2 Teoria RST

A teoria RST foi desenvolvida para interpretação e análise linguística dos textos. É importante o seu conhecimento, pois é a teoria com mais influência no processamento de línguas naturais e consequentemente na sumarização automática (MARTINS et al,2001).

RST (*Rhetorical Structure Theory*) ou no português, Teoria da Estruturação Retórica, define algumas relações retóricas de coerência que representam as relações existentes dentro de um determinado texto. Essas relações apresentam certo nível de dificuldade, uma vez que elas remetem ao significado do texto e portanto dependem da interpretação da língua.

Essa teoria trabalha com o conceito de núcleo e satélite. Núcleo é a informação principal do texto, é a mensagem que o autor quer passar ao leitor. Satélite é aquela informação adicional, que serve como suporte para que a interpretação seja feita pelo leitor de forma clara.

Cada relação precisa de uma quantidade de informações, que irão defini-la dentro do processo. Essas informações contêm campos chamados de restrição e efeito, respectivamente. Os campos são: restrição sobre o núcleo, restrição sobre o satélite, a combinação do núcleo e o satélite e por último o efeito, que é a definição do efeito que a relação causa no leitor.

Essas relações RST, podem ser consideradas como fracas ou fortes. Quando uma relação é considerada fraca, o trecho que contém essa relação pode ser excluído do texto original, ajudando assim a confecção do sumário automático.

Essas exclusões devem ser bem analisadas. Se forem feitas a qualquer modo, podem trazer incoerência para o sumário.

> O corte de relações RST inteiras e/ou a interpretação de núcleos das relações,somente pode implicar a perda da coerência do texto, com a geração de parágrafos muito curtos, a perda ou o excesso de pontuação e a existência de expressões referencias ou marcadores de discurso cujos referentes se perdem no texto resultante. (Martins et al ,2001,p 30)

1.3 Sumarização Automática Monodocumento x Sumarização Automática Multidocumento

1.3.1 Sumarização Automática Monodocumento

A SA pode ser classificada quanto a quantidade de textos que participam do processo, em monodocumento ou multidocumento. Como pode-se inferir, monodocumento refere-se a apenas um texto e multidocumento a vários textos-fontes.

A SA monodocumento utiliza apenas um texto-fonte no processamento. O sumário é feito a partir de um único texto e é utilizado para informar o leitor acerca de um só artigo, tese, notícia, etc.

1.3.2 Sumarização Automática Multidocumento (SAM)

A Sumarização Automática Multidocumento (SAM) trabalha com vários textos-fonte ao mesmo tempo. A sumarização multidocumento é a mais utilizada atualmente, já que a quantidade de informações disponível na internet aumentou consideravelmente.É um pouco mais trabalhosa que a sumarização monodocumento, por questões de coerência e coesão, que nesse contexto, tornam-se mais complicadas de serem colocadas em prática.Segundo Dias et al (2014, P.1) “a Sumarização Automática Multidocumento (SAM) é a tarefa de se produzir sumários (comumente chamados resumos) de maneira automática a partir de um conjunto de textos-fonte sobre um mesmo assunto.”

Alguns problemas com a SA foram surgindo ao longo do uso dessa técnica. Por serem vários textos, é necessário identificar o que é informação importante e qual informação não é relevante para o sumário, é necessário também identificar se será preciso seguir uma ordem temporal de acordo com os textos-fonte e acima de tudo, manter a coerência e coesão do

texto, lembrando que, esses textos podem ser escritos por pessoas diferentes e por isso podem ter uma estrutura diferente entre eles.

1.3.2.1 CSTNews

No Brasil, foi criado um *corpus* para poder subsidiar as pesquisas em SAM. O CSTNews foi o primeiro *corpus* composto de textos e sumários para ser analisado e acrescentar dados as pesquisam na área de SAM.

Esse *corpus* atualmente, segundo Pardo (2008, P.1), é composto de 50 coleções de textos, sendo que cada coleção dispõe de 4 textos de diferentes fontes sobre um mesmo assunto. Esses textos são todos jornalísticos, retirados de diferentes jornais em circulação no Brasil. O gênero jornalístico foi escolhido pela facilidade que se têm de encontrá-los no ambiente web e também devido a sua linguagem mais acessível, utilizada no dia-a-dia pelos leitores.

Cada uma das coleções de textos é composta por um sumário abstrativo e um extrativo, e são acompanhados por anotações linguísticas tanto sobre o sumário quanto sobre o texto-fonte. Entre essas anotações, estão as segmentações dos textos-fonte em subtópicos, aspectos informativos dos sumários abstrativos, substantivos e verbos mais frequentes, etc.

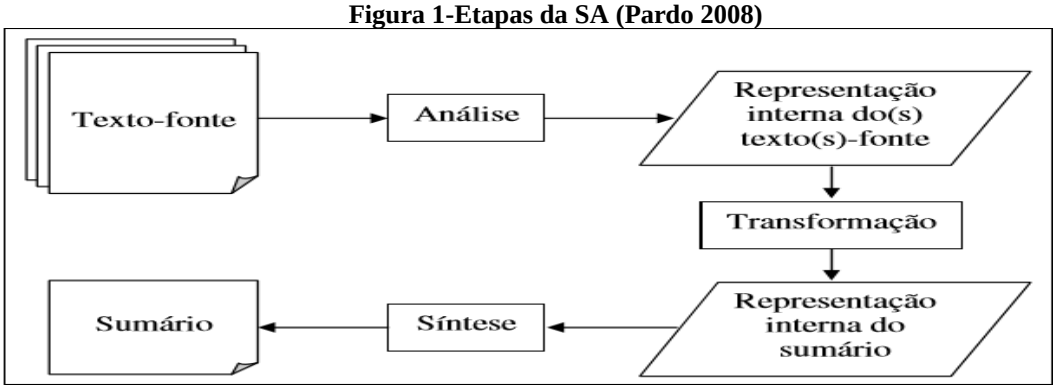
Uma pesquisa apresentada por Dias (2014), mostra sumários feitos por 20 pesquisadores de PLN, sendo que cada um produziu 25 sumários entre extratos e *abstracts*. No resultado, pôde-se perceber que os *abstracts* foram similares em tamanho aos extratos e que também houve uma sentença ao menos em comum entre os sumários e que essa sentença geralmente era a primeira de um dos textos-fonte que compõe a coleção. Assim, Dias et al (2014), inferiram que a informação principal do texto-fonte estava contida no primeiro parágrafo.

Com esses sumários produzidos por humanos a partir de relações lógicas, será possível otimizar os sumarizadores automáticos multidocumentos. Segundo Dias (2014), os dados coletados revelam quais são as principais informações que serão necessárias para o leitor entender a mensagem principal que o autor do texto-fonte pretendia passar.

2. PROCESSO DE SUMARIZAÇÃO AUTOMÁTICA

O processo de sumarização automática é composto por 3 etapas: análise, transformação e síntese (PARDO,2008).

Figura 1-Etapas da SA (Pardo 2008)



Fonte: Disponível em: PARDO,2008,P.6.

Na etapa de análise, um ou vários textos são processados e uma representação interna do(s) texto(s)-fonte com o conteúdo dos textos é produzida. Na segunda etapa, a de transformação, é realizado o processo de sumarização com base na representação interna do(s) texto(s)-fonte gerada na etapa de análise, é criada então a representação interna do sumário, que contém o conteúdo mais relevante para ser exposto. E a terceira e última etapa, a síntese, apenas reproduz em língua natural a representação interna do sumário, criada na etapa de transformação.

Todo o processo de SA, é guiado pela taxa de compressão. Taxa de compressão é a taxa do texto fonte que deve ser retirada, sendo o restante o que deve compor o sumário. Por exemplo, se temos uma taxa de compressão de 60%, o sumário deve ter o equivalente a 40% do texto-fonte.

As etapas do processamento de SA são independentes, sendo assim para cada método escolhido para realizar a sumarização, as etapas podem ser adaptadas. Em alguns casos, análise e síntese podem ser suprimidas do processo e em outros, dependendo da abordagem, elas são essenciais, assim como na abordagem profunda.

Vale ressaltar, que as etapas encontradas no processo de sumarização automática são similares ao processo de Geração Automática de Textos, uma área dentro de PLN. O processo de SA, pode ser muitas vezes tratado como um processo de geração textual.De acordo com Pardo (2008):

> É importante que se diga que as etapas de SA se assemelham muito ao processo realizado na grande subárea de Geração Automática de Textos dentro de PLN (vide,

> por exemplo, Reiter e Dale, 2000). Muitas vezes, aliás, o processo de SA pode ser visto e tratado como um processo de geração textual.(Thiago Pardo,2008, P. 7)

3.GIST SUMMARIZER

Gist Summarizer é um SSA ou sumarizador automático, que utiliza a abordagem superficial. Também chamado de GistSumm, tem como principio a ideia principal do texto para realizar a sumarização. O sumarizador identifica a sentença mais significativa do texto fonte, a qual é chamada de sentença-gist.

O GistSumm foi o primeiro sumarizador automático disponível para a língua portuguesa, e apesar de ser disponível para o português brasileiro, pode ser utilizado para qualquer língua natural porque seus métodos são independentes de língua. Segundo Pardo (2002):

> Ainda, como extensão do sistema, é importante ressaltar a possibilidade de personalização do GistSumm para outras línguas naturais, o que pode ser feito pelo simples acoplamento de novos repositórios de dados da língua desejada, isto é, de um léxico e de uma stoplist específicos. Caso o GistSumm venha a ser usado em domínios específicos, a stoplist também pode ser substituída por uma outra específica do domínio.(Pardo,2002,P.19)

Inicialmente o GistSumm foi feito para realizar sumarização monodocumento.Com o avanço das pesquisas e as necessidades do mercado, começou a realizar também sumarização multidocumento, que como comentado anteriormente, é aquela em que utiliza vários textos-fonte. Atualmente o GistSumm também realiza sumários de teses e artigos científicos, que são textos mais elaborados e com estrutura mais rígida.

A sentença-gist é aquela que contém as palavras que aparecem mais vezes no texto(s)-fonte. É a partir dela que são escolhidas as demais sentenças que irão compor o extrato.De acordo com Pardo (2008, P.8) “o sistema identifica a sentença mais importante do texto-fonte, chamada sentença-gist. A seguir, procura por sentenças que complementam esta sentença,as quais, juntamente com a sentença-gist, constituem o extrato.”

O GistSumm segue três etapas (GUELPELI,2012), a saber: segmentação sentencial, ranqueamento das sentenças e seleção de sentenças.

Na etapa de segmentação sentencial, ocorre a escolha das sentenças pelos sinais de pontuação, tais como exclamação, interrogação e ponto final. As sentenças são marcadas por colchetes e cada uma recebe uma numeração especifica para poder ser localizada dentro do texto.

No ranqueamento das sentenças, o usuário do sumarizador escolhe o método para ranquear as sentenças, ou seja, atribuir uma pontuação. Esse processo de pontuação é feito em algumas sub-etapas (PARDO, 2002):

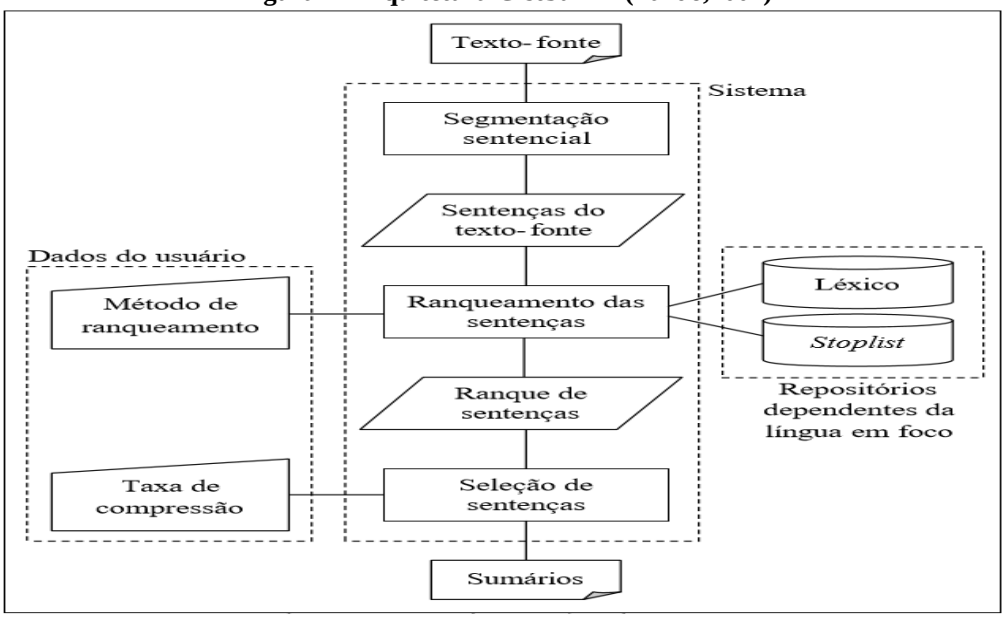
- Criação de vetores das sentenças: As sentenças são convertidas em vetores de palavras, cada palavra é colocada em um vetor na mesma ordem em que aparecem na sentença.
- *Case folding*, troca por canônicas e remoção de *stopwords*: Nessa etapa, o *case folding* transforma todas as palavras em letras minúsculas; o *stemming* troca as palavras por suas canônicas ou seja, palavras que seguem a mesma estrutura na gramática, e por último as *stopwords* são excluídas do texto e o cálculo de importância das sentenças pode ser realizado.
- Pontuação da sentença: A pontuação que é dada à sentença pode ser escolhida pelo método da palavra-chave, onde se têm que a ideia principal do texto pode ser encontrada pelas palavras-chaves. Por esse método, cada vetor recebe uma pontuação de acordo com a frequência das palavras que o compõem no texto.Aquela sentença que tiver maior pontuação no *rank* será a sentença-gist. É nessa sub-etapa que é utilizada a lista de *stopwords*, as *toplist*,ou seja, lista de palavras que devem ser excluídas na hora de sumarizar o texto-fonte.

Na última fase, a de seleção de sentenças, é calculada a média da pontuação das sentenças no texto-fonte, tal média é chamada de *baseline* e é utilizada como limite para cortar as possíveis sentenças que irão compor o sumário. A sentença que tiver a maior pontuação,será a sentença-gist e aquelas que tiverem pontuação acima do estabelecido pela *baseline* irão compor o sumário juntamente com a *gist sentence*.

Ainda na fase de seleção de sentenças o usuário seleciona a taxa de compressão. Essa informação dada pelo usuário é muito importante, pois é a partir dela que o GistSumm irá criar o sumário. A quantidade de sentenças que são usadas para compor o sumário é proporcional a taxa de compressão, ou seja, quando o percentual de taxa de compressão é atingido, o sistema para de colocar as sentenças no sumário.

A figura 2, mostra de forma exemplificada como é a estrutura interna do GistSumm:

Figura 2- Arquitetura GistSumm (Pardo,2002)



Fonte: Disponível em: PARDO,2002.

3.1Ambiente virtual do Sumarizador Automático Gist Summarizer

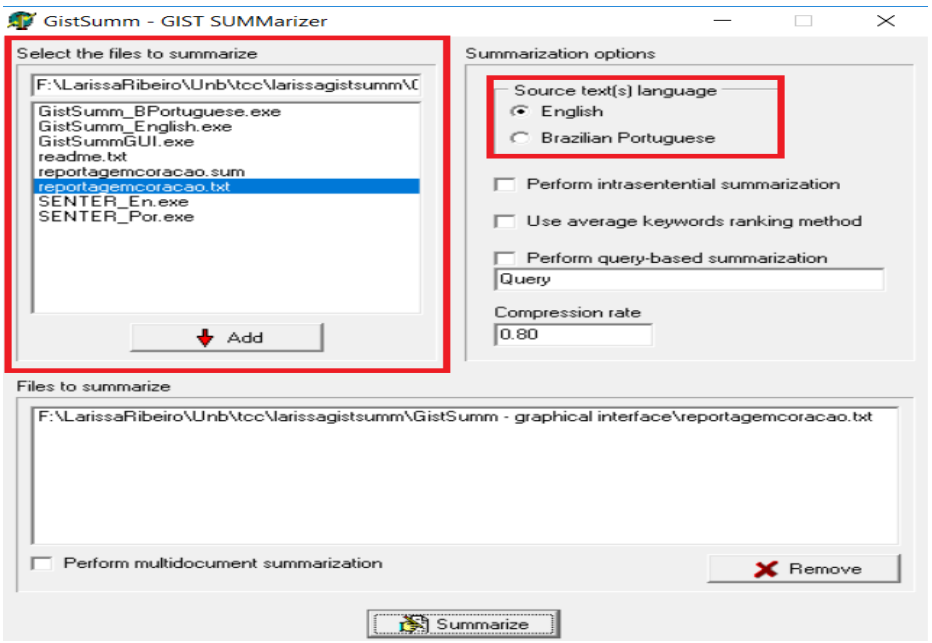
O GistSumm é um software disponível para download no link <metashare.elda.org > e no site do Núcleo Intralinguístico de Linguagem Computacional (NILC), gratuitamente. É um software simples e de fácil acesso.

O usuário irá ter acesso a uma pasta que contém a interface do GistSumm.Na interface existe um campo com os documentos inseridos na pasta. O usuário irá escolher um deles, ou vários textos. Além do mais, os documentos que serão sumarizados, precisam estar na mesma pasta onde se encontra a interface do GistSumm. Logo ao lado, há um campo para que o usuário possa escolher o idioma em que o texto fonte foi escrito.

A interface do GistSumm é na língua inglesa, e não conta com a opção para mudar o conteúdo da interface para outro idioma.Apesar de ser o primeiro sistema disponível para a língua portuguesa, não contém opções na interface para que seja escolhido o idioma português.Só existem dois campos, inglês e português que são as opções para escolher o idioma do texto de entrada.

Essa questão se remete à usabilidade do sistema, uma questão importante quando falamos de transmissão de informação. Uma vez que o usuário não é um falante do inglês ou tem pouco ou nenhum conhecimento da língua inglesa, pode ter problemas para usar adequadamente o sistema, se sentindo desconfortável e acabar por não optar utilizar o sumarizador.

Figura 3 - Interface GistSumm.



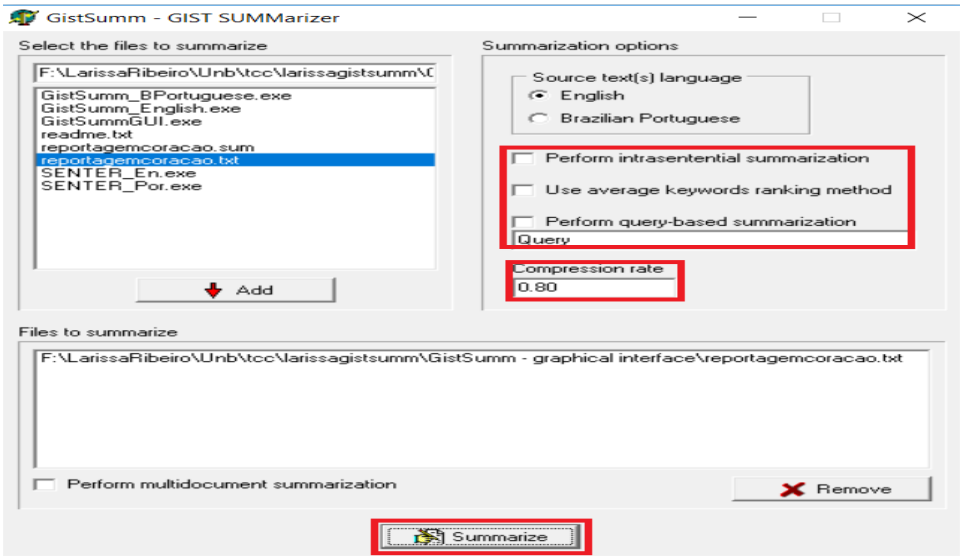
Fonte: Disponível em : < http://www.nilc.icmc.usp.br/nilc/index.php/tools-and-resources>

Abaixo, o usuário irá marcar as opções que deseja para a sumarização do texto. Essas opções são:sumarização intrasentencial, método de palavras chaves e a sumarização baseada na *performance query*. Há ainda, outro campo que deve ser preenchido pelo usuário, é o campo de taxa de compressão, "*COMPRESSION RATE*". O usuário precisa colocar a porcentagem da compressão que deseja para seu sumário

Nesse sistema, quando fala-se em taxa de compressão, temos outra questão dentro de usabilidade. O usuário que não tem conhecimento em sistemas de sumarização, geralmente não compreende o que taxa de compressão quer dizer e qual a sua necessidade. Não existe um campo explicando qual a função dessa opção e como irá influenciar no sumário. É importante essa explicação para que o usuário se sinta mais confortável quando estiver utilizando o sumarizador.

Depois de escolhidos os requisitos para o sumário, o usuário deverá clicar no botão "*SUMMARIZE*".Os campos citados estão marcados na figura abaixo.

Figura 4 - Interface GistSumm.



Fonte: Disponível em : < http://www.nilc.icmc.usp.br/nilc/index.php/tools-and-resources>

É importante ressaltar que o GistSumm só aceita textos em formato TXT, o que constitui um obstáculo para o uso do sistema de sumarização, já que uma grande parte dos textos acadêmicos, que são um dos focos do GistSumm, são gravados em formato PDF, o formato mais comum disponibilizado na web.

Os sumários ou os textos de saída do GistSumm são em geral muito satisfatórios, conseguem trazer a mensagem principal do texto de forma bem sintetizada. Porém, os sumários não são separados em tópicos, quando esses existem no texto-fonte, e não vêm com o título do texto que foi sumarizado. Isso pode dificultar um pouco a leitura do sumário ou a identificação de tópicos principais pelo leitor.

Sendo assim, pode-se perceber que existem vários elementos que podem dificultar o acesso ao sistema de sumarização automática GistSumm. Esses aspectos muitas vezes passam despercebidos pelos desenvolvedores, mas são de suma importância para que a SA continue a ser difundida e que o seu uso se torne cada vez mais comum no ambiente web.

4. SWESUM

De acordo com Dalianis (2000),o sumarizador automático SweSum, foi o primeiro SSA criado para a língua sueca.Entretanto, ele também pode ser utilizado para outras línguas assim como o espanhol, o francês e o inglês.

Criado por Martin Hassel e Hercules Dalianis, o sumarizador SweSum, segue o método de sentenças com maior pontuação. As sentenças que recebem pontuação elevada farão parte do sumário. Essa pontuação é dada de acordo o gênero jornalístico-que é o foco

desse sistema-, partindo do principio que textos desse gênero tem sua informação principal nas primeiras sentenças, então à essas sentenças é dada uma pontuação alta, enquanto as sentenças no final do texto recebem pontuação baixa. Apesar de ter sido feito com foco em textos jornalísticos, hoje o SweSum realiza também sumarização de textos acadêmicos.

O SweSum utiliza textos jornalísticos na linguagem HTML. Ele ignora o formato da página HTML, mas processa as *tags* que se referem ao texto presente na página.As *tags* HTML que indicam sentenças em negrito recebem uma pontuação elevada, isso se deve ao fato de que em alguns textos de jornais suecos os parágrafos começam com palavras em negrito (DALIANIS,2000).

O sumarizador SweSum utiliza alguns critérios para poder calcular as pontuações de cada sentença. As sentenças que contêm data e as sentenças que contêm as palavras chaves, ou *Term Frequency* recebem uma pontuação alta.Para poder saber quais são as palavras chaves que o sistema utiliza, existe um dicionário pronto com todas as palavras que carregam significado. Esses critérios são combinados e é dada a pontuação final para cada sentença.

Nesse sistema, o usuário pode entrar com palavras chave e dessa forma o sumário irá sair personalizado. Além disso, o usuário poderá escolher a taxa de compressão que deseja para saída do seu sumário.

O SweSum é um sumarizador muito difundido, e geralmente é o mais escolhido para fazer resumos de texto(s)-fonte na língua inglesa. A sua adaptação para outros idiomas, se deve ao fato do sistema mapear as formas flexionadas de outras línguas a partir de um léxico utilizado pelo próprio SweSum. Esse mapeamento é realizado comparando as palavras com o léxico, trazendo-as para sua raiz respectiva.

4.1 Ambiente Virtual do Sumarizador Automático SweSum

O sumarizador SweSum, conta com uma plataforma online, não é necessário que o usuário faça o download do sistema para que possa utilizá-lo. O sumarizador é disponibilizado na web de forma gratuita, e não necessita de *login* ou cadastro prévio para ser usado.

A página inicial do SweSum é disponibilizada para língua inglesa no link- <http://swesum.nada.kth.se/index-eng.html>. Logo no inicio da página é possível ver algumas bandeiras que representam os países com suas respectivas línguas principais. Ao contrário do GistSumm, o SweSum conta com Interface para várias línguas, auxiliando assim o usuário a navegar na página. Ao clicar em uma dessas bandeiras, o usuário irá ter acesso à página na língua selecionada.

Pode-se entender melhor essa explicação na imagem abaixo

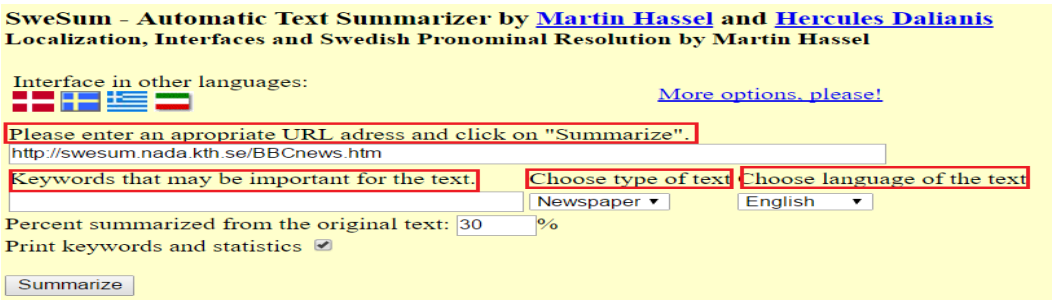
Figura 5 - Interface SweSum

Fonte: Disponível em: <http://swesum.nada.kth.se/index-eng-adv.html>

Já nessa mesma página, encontra-se um campo*"PLEASE ENTER AN APROPIATE URL ADRESS AND CLICK ON 'SUMMARIZE'"*, para entrada do texto a ser sumarizado. Esse texto precisa estar em HTML, ou seja, ali será colocado a URL da página que se deseja sumarizar. Logo abaixo do campo de entrada do texto, há outro campo com o enunciado "*KEYWORDS THAT MAY BE IMPORTANT FOR THE TEXT*"onde o usuário poderá escolher as palavras-chave para personalizar seu sumário.

Logo ao lado em "*CHOOSE TYPE OF TEXT"*o usuário poderá escolher o gênero do texto a ser sumarizado : jornalístico ou acadêmico.No campo *"CHOOSE LANGUAGE OF THE TEXT*" o usuário poderá escolher a língua em que o texto-fonte está escrito. As opções são:dinamarquês,inglês, francês,alemão,grego,norueguês,persa,espanhol, sueco e italiano.

Figura 6 - Campos de entrada da URL, palavras-chave, tipo textual e idioma.



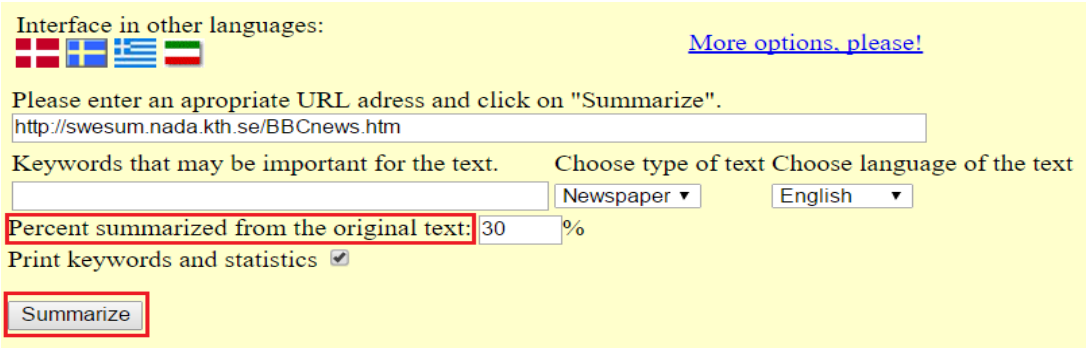
Fonte: Disponível em: <http://swesum.nada.kth.se/index-eng-adv.html>

O penúltimo passo é a escolha da taxa de compressão do texto fonte no campo *"PERCENT SUMMARIZED FROM THE ORIGINAL TEXT* ", essa etapa é de suma importância, pois o tamanho do sumário irá depender dessa escolha.

Uma das qualidades do SweSum é explicação dos campos, ao invés de ter apenas palavras indicando o que precisa ser colocado no campo, o sumarizador disponibiliza breves explicações, como o enunciado da taxa de compressão citado acima, que diferentemente do GistSumm traz uma breve explicação, e não apenas enunciado “TAXA DE COMPRESSÃO”.

Por último o usuário deverá clicar em "*SUMMARIZE*" para dar inicio a sumarização.Assim que a sumarização é feita, o site já encaminha o usuário para outra página com o sumário pronto.

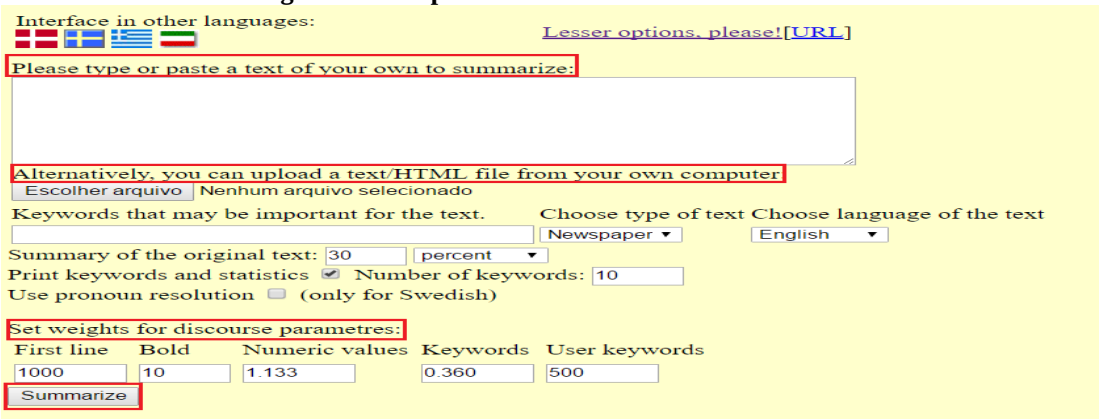
Figura 7 - Campo de entrada compressão e botão de summarize.



Fonte: Disponível em: <http://swesum.nada.kth.se/index-eng-adv.html>

Ao clicar em “MORE OPTIONS, PLEASE!”, o usuário é levado à outra página. Nessa página, terá um campo "PLEASE TYPE OR PASTE A TEXT OF YOUR OWN TO SUMMARIZE" onde o usuário poderá digitar um texto ou colar um texto para ser sumarizado.Logo abaixo o usuário poderá entrar com um texto em .TXT ou página HTML salvo em seu próprio computador no campo "ALTERNATIVELY, YOU CAN UPLOAD A TEXT/HTML FILE FROM YOUR OWN COMPUTER", basta clicar no botão para escolher o arquivo.Logo abaixo no campo "SET WEIGHTS FOR DISCOURSE PARAMETRES" o usuário poderá entrar com valores para parâmetros de discurso.Depois basta clicar em "SUMMARIZE" para obter o sumário do texto-fonte.

Figura 8 - Campos de entrada de texto do SweSum.



Fonte: Disponível em: <http://swesum.nada.kth.se/index-eng-adv.html>

O SweSum conta com um bom nível de usabilidade.Sua interface é simples e os campos estão dispostos de maneira clara e qualquer usuário, mesmo que seja leigo no uso de sumarizadores automáticos, pode utilizá-lo sem grandes problemas.Apesar do número de idiomas da interface ainda ser um pouco limitado, já é um grande avanço e já alcança um grande número de usuários de diferentes línguas e regiões do mundo todo.

Os textos de saída do SweSum são muito satisfatórios.O sumário vêm com o título do texto fonte e seus respectivos tópicos. Assim, mesmo que o texto seja um resumo, ele continua com a mesma estrutura do texto-fonte.

Nesse sistema de sumarização ocorre o mesmo problema que acontece no GistSumm: os textos de entrada não podem ser em extensão PDF.Ainda, para sumarizar é necessário que os textos estejam em extensão TXT.

Sendo assim, o sistema de sumarização automática SweSum conta com recursos muito bons para sumarização e é de fácil acesso pelo usuário.Entretanto, ainda tem algumas falhas, que, se corrigidas, podem favorecer consideravelmente o uso da SA.

5. APLICAÇÕES DA SA

5.1 - Textos jornalísticos

Uma das principais aplicações da SA são os textos do gênero jornalístico.São textos de grande repercussão no ambiente web, já que trazem noticias do cotidiano, relevante a todos os cidadãos. São textos com grande quantidade de acesso, independente da fonte.

O usuário geralmente se encaminha para uma determinada página se o conteúdo em questão lhe chama atenção. Para escolher qual página web irá acessar, o leitor lê primeiramenteo resumo do texto original, para escolher aquele que o agrada. Realizar esse

resumos por meio de sumarização humana demandaria muito tempo.Com o uso do sumarizador automático, tanto o responsável pelo site pesquisado quanto o leitor, irão ganhar tempo, uma vez que o texto-fonte é sumarizado de forma rápida, e traz um sumário coerente e coeso, com a mensagem principal do texto original para ser exposto nos *sites* de buscas e no próprio *site* onde está o texto fonte.

5.2 - Tradução automática

A tradução automática está sendo estudada como uma possível aplicação da SA. É de grande importância para o acesso de não falantes da língua portuguesa à conteúdos disponibilizados na web.

Para exemplificar o uso de sumários automáticos na tradução, pode-se pensar em um usuário com pouco domínio da língua portuguesa. Ao invés de precisar traduzir um texto inteiro ou fragmentos desconexos para entender o conteúdo de um site, textos explicativos, ou ainda, teses e dissertações para estudo, o leitor pode utilizar o sumário do texto escolhido e traduzi-lo em um bom tradutor automático.

Tradutores automáticos não têm muita precisão quando usados para traduzir grandes textos, mas para traduzir textos menores ou fragmentos de textos, a tradução é de alta qualidade e traz a ideia de forma precisa e conexa.

Tradutores automáticos hoje, já contam com recursos avançados, não traduzem apenas palavra por palavra, mas a palavra em seu contexto.Exemplos de bons tradutores automáticos estão no site português *Linguateca*- Centro de recursos para o processamento computacional da língua portuguesa - <http://www.linguateca.pt/>. Exemplo de tradutores automáticos disponíveis no site da Linguateca são : *officetranslator* , FALATUDO, *moses*, WinDi e *WordTranslator*. Alguns desses sistemas de tradução necessitam de pagamento para poder ser utilizado.

Para exemplificar a situação foi escolhido o tradutor automático *GoogleTranslate*, hoje já considerado um dos serviços líderes em tradução automática gratuito. É um serviço utilizado por mais de 200 milhões de pessoas e conta com aproximadamente 90 idiomas.

Foi escolhido um texto jornalístico, retirado do *site* <G1.globo.com> para ser sumarizado pelo GistSumm. Abaixo segue o texto original.

Figura 9 -Texto-fonte para sumarização automática

Coração artificial ajuda bebê a esperar por transplante em São Paulo

Aparelho impulsiona o sangue, dando suporte ao coração que não tem força suficiente para manter circulação sanguínea.

Um coração artificial está ajudando um bebê de 11 meses a esperar por um transplante, em São Paulo.

A batida que toca na reportagem é criação do homem. Ela parte de um coração artificial - o primeiro no Brasil colocado numa criança tão pequena como o Gustavo. Já faz mais de três meses que o garotinho tem o aparelho.

"Fez a cirurgia e colocou, e deu muito certo. Fiquei muito feliz, porque ele está vivo graças a isso", contou Luane Barrios, mãe do Gustavo.

"O aparelhinho tem duas partes: uma parte onde circula o sangue e outra parte onde o ar entra, e no meio tem uma membrana. Essa membrana impulsiona o sangue entra e sai do aparelhinho. Com isso, ele dá um suporte ao coração – o coração está batendo, está funcionando, mas ele não tem força suficiente para manter a circulação adequada, então o aparelho serve como uma força além, adjuvante, como um suporte ao coração", explicou Marcelo Jatene, diretor da Unidade Cirúrgica Cardio pediátrica do Incor.

O Gustavo tem o miocárdio, que é o músculo do coração, dilatado. E a única saída para ele é o transplante. Ele está na fila de espera e tem prioridade máxima.

"Esses chamados corações artificiais recentemente permitem que o paciente sobreviva esperando o transplante cardíaco, podendo ficar em ambiente hospitalar ou até ir para casa com o coração artificial. Vai ter uma vida bem próximo do normal ou normal, quando aparecer o doador ele é chamado e coloca o coração definitivo", afirmou Roberto Kalil Filho, presidente do conselho diretor do Incor.

O equipamento que o Gustavo está usando é importado, alemão, bem parecido com um que foi desenvolvido no Brasil pelo próprio Incor, o Instituto do Coração. O coração artificial mostrado na reportagem é do tamanho do de um adulto e já foi testado. Agora a equipe começa uma nova etapa: a avaliação de um modelo que tem o tamanho do coração artificial do Gustavo. Serve para uma criança. Só que, como é uma pesquisa ainda, não tem produção comercial.

A expectativa é que o custo do aparelho brasileiro seja menor do que o alemão.

"A ideia é ampliar a possibilidade de se utilizar essa terapia em pacientes que utilizam o sistema SUS, e também desenvolver tecnologia no país", disse Idágne Cestari, diretora de bioengenharia do Incor.

Até lá, segue a batalha do pequeno coração valente.

Jornal Nacional: O aniversário do Gustavo é...

Luane Barrios: No Natal. Meu presente de Natal. Com certeza, acho que vai ganhar o presentinho dele. Quem sabe um coraçãozinho, né?

O presente da Júlia já veio, e ela nem precisou da ajuda do aparelho. Há um mês, ela ganhou um coração novinho e sentiu a diferença.

"A primeira reação dela foi falar do batimento do coração, que ela sentia bater mais forte", contou Jéssica Cleim, mãe da Júlia.

Jornal Nacional: Como faz o barulho?

Júlia Pereira, 4 anos: Tum, tum.

Jornal Nacional: Você não tinha ouvido esse barulho antes?

Júlia Pereira: Não.

Jornal Nacional: E foi legal?

Júlia Pereira: Foi.

Fonte:Disponível em: <http://g1.globo.com/jornal-nacional/noticia/2015/12/coracao-artificial-ajuda-bebe esperar-por-transplante-em-sao-paulo.html>

O mesmo texto, foi utilizado como texto-fonte pelo GistSumm. Abaixo está o texto depois de sumarizado.

Figura 10 - Texto sumarizado pelo GistSumm

Aparelho impulsiona o sangue, dando suporte ao coração que não tem força suficiente para manter circulação sanguínea.

Um coração artificial está ajudando um bebê de 11 meses a esperar por um transplante, em São Paulo.

Com isso, ele dá um suporte ao coração – o coração está batendo, está funcionando, mas ele não tem força suficiente para manter a circulação adequada, então o aparelho serve como uma força além, adjuvante, como um suporte ao coração", explicou Marcelo Jatene, diretor da Unidade Cirúrgica Cardiopediátrica do Incor.

Agora a equipe começa uma nova etapa: a avaliação de um modelo que tem o tamanho do coração artificial do Gustavo.

Já com o texto sumarizado, o leitor poderá traduzi-lo no tradutor automático. Não será necessário traduzir um texto grande na integra,com o sumário,o individuo não falante da língua portuguesa poderá obter uma tradução precisa e coesa,que não aconteceria com um texto de grandes proporções como o da figura 9.

Abaixo o texto sumarizado, traduzido pelo tradutor automático *Google Translate.*

Figura 11 - Texto Traduzido pelo Tradutor Automático *Google Translate*

The device boosts the blood by supporting the heart that does not have enough strength to maintain blood circulation.
An artificial heart is helping an 11-month-old baby wait for a transplant in Sao Paulo.
With this, it gives support to the heart - the heart is beating, it is working, but it does not have enough strength to maintain proper circulation, so the device serves as a force beyond, adjuvant, as a support to the heart, "explained Marcelo Jatene, director of the Incorrect Cardio pediatric Surgical Unit.
Now the team begins a new stage: the evaluation of a model that is the size of Gustavo's artificial heart.

O idioma escolhido foi o inglês, essa escolha foi aleatória, apenas para exemplificar a situação de um não falante da língua portuguesa. Pode-se perceber a precisão da tradução do português para o inglês, a única falha foi a tradução de *Incor* para *Incorrect*, uma vez que Incor é uma sigla para Instituto do Coração e não foi reconhecida pelo tradutor. Pode-se analisar também o suporte que o sumarizador automático oferece para o tradutor automático.

5.3–Textos Acadêmicos

Textos acadêmicos também são um dos focos da SA. O sumarizador automático GistSumm por exemplo, já tem recursos para traduzir esse gênero textual.

Para acessar textos inseridos em bancos ou em sites de busca, os leitores precisam primeiro saber se o texto acadêmico é realmente pertinente para sua pesquisa. O titulo do texto auxilia a achar o tema que está sendo procurado, porém não se pode certificar que esse é mesmo o assunto que o leitor está precisando e seria muito desgastante para o leitor ler o texto

até saber se aquele conteúdo é ou não o que procura. É nesse contexto que os sumários podem ser adequados.

A sumarização automática é utilizada para elaborar tais sumários e trazer a mensagem principal do texto em questão. Não é mais necessário que o leitor precise se desgastar para encontrar o conteúdo que deseja. Além disso, não são necessários sumarizadores humanos para disponibilizar os resumos para cada texto.Os sumários podem ser feitos online ou por softwares disponibilizados para download.

Os artigos científicos, teses e dissertações já contêm em sua estrutura o resumo na língua mãe do autor e em outra língua, os *abstracts*. Porém esses resumos são de poucas linhas e não conseguem mostrar com precisão o conteúdo que está apresentado na obra, eles são apenas para exemplificar o tema abordado. Os sumários produzidos automaticamente conseguem apresentar os tópicos do texto e mostrar a mensagem principal abordada.

6. PROPOSTA DE FILTRO AUXILIAR PARA A SA

Analisando a importância da sumarização automática em várias áreas, foi proposta a criação de um software, chamado de *filtro auxiliar*. A proposta consiste na exclusão de *sessões* de artigos científicos que não são relevantes para o sumário,- por isso a nomenclatura "filtro" -, antes que o texto-fonte passe pelo sumarizador, ou seja, o texto-fonte será tratado e sairá pronto para ser sumarizado.Além dessa funcionalidade, o *filtro auxiliar* converteria o texto em *pdf* para *txt*, que é um dos problemas dos sumarizadores automáticos.

É importante ressaltar que *software* é um conjunto de instruções escritas para serem interpretadas por um computador para executar tarefas específicas. De acordo com o Dicionário Priberam da Língua Portuguesa, *software* é o “conjunto de programas,processos,regras e eventualmente, documentação, relativos ao funcionamento de um conjunto de tratamento de informação.”

As *sessões* a serem excluídas foram nomeadas de *stop sessions*, esse nome foi dado, baseado nas nomenclaturas já comentadas *stoplist* e *stopwords*, utilizadas na estrutura do GistSumm. As *stop sessions* são tópicos que compõem tradicionalmente a estrutura de textos acadêmicos, sendo alguns de caráter facultativo. Podemos citar alguns, como por exemplo: agradecimentos, lista de tabelas e bibliografia. Esses tópicos geralmente não são relevantes para o sumário porque não trazem nada do conteúdo especifico do texto, portanto não são pertinentes do ponto de vista semântico.

O *filtro auxiliar* tem como intuito preparar o texto e auxiliar o funcionamento do sumarizador automático, portanto, ele será um software auxiliar e será utilizado antes do

processo de sumarização. O usuário poderá escolher quais sessões deseja excluir, caso considere alguma importante para seu sumário, poderá incluí-la.

Foi escolhido como *corpus* para análise, a Biblioteca Digital Brasileira de Teses e Dissertações (BDTD) para exemplificar como o *filtro auxiliar* excluiria sessões consideradas irrelevantes. A BDTD tem como objetivo reunir em um portal de busca, as teses e dissertações defendidas em todo o País e por brasileiros no exterior.Foi lançada em 2002 e foi concebida pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) com o apoio da Financiadora de Estudos e Pesquisas (FINEP).

Como SSA que receberia o texto filtrado,foi escolhido o GistSumm, que melhor se aplica à língua portuguesa.

Na tabela 1,está a lista de *stop sessions*. São sessões que serão excluídas do texto-fonte, e um breve detalhamento de sua composição.Essas sessões foram escolhidas com base nas teses e dissertações disponibilizadas na BDTD e segundo as estruturas obrigatórias ou facultativas de textos acadêmicos baseados nas normas da Associação Brasileira de Normas Técnicas (ABNT).

Tabela 1 - Lista de *Stop Sessions*

Capa	É composta dos tópicos: • Universidade • Departamento • Nível de graduação • Título • Nome do autor • Nome do orientador • Cidade • Data
Contra capa	É composta dos tópicos: • Universidade • Departamento • Nível de Graduação • Título • Nome do autor • Propósito do trabalho • Nome do orientador

Ficha Catalográfica	É composta dos tópicos: • Nome do autor • Título • Ano de publicação • Palavras-chave
Banca Examinadora	Essa sessão é composta dos nomes dos examinadores da defesa e espaço para suas assinaturas.
Agradecimentos	Essa sessão é destinada ao autor do texto acadêmico, para que possa fazer os agradecimentos que acha necessário.
Resumo	Nessa sessão, o autor faz um breve resumo em sua língua materna sobre sua obra, e escolhe as palavras-chaves do texto.
Resumo em outros idiomas	Nessa sessão o autor traduz o resumo para outro idioma de sua preferência.
Índice geral	O índice geral é composto dos tópicos do texto e suas respectivas páginas.
Índice de figuras/tabelas	É composto da numeração das figuras/tabelas e suas respectivas páginas.
Referências Bibliográficas/Bibliografia	Sessão destinada às obras utilizadas no trabalho, ano de publicação e seus respectivos autores.

Como os documentos disponibilizados no BDTD são de extensão *pdf*, o filtro receberia os arquivos na sua extensão de origem os converteria e disponibilizaria o texto com extensão *txt*, que é a extensão que os sumarizadores automáticos como o GistSumm recebem.

O filtro então contaria com duas funcionalidades: a exclusão de sessões não relevantes e a conversão de arquivo *pdf* para *txt*. O texto filtrado será, no caso, o texto-fonte utilizado pelo sumarizador automático, que irá disponibilizar um sumário mais objetivo e personalizado para o leitor, que na situação em questão, é o usuário do software.

METODOLOGIA

O desenvolvimento desse trabalho de conclusão de curso segue as seguintes etapas:

1. **Estudo da SA através de dois sumarizadores automáticos**

Os sumarizadores escolhidos para o estudo foram os seguintes: o GistSumm (Gistsummarizer) de criação de Thiago Pardo e o SweSum de criação de Martin Hessel e Hercules Dalianis. A escolha dos sumarizadores deveu-se ao desejo de mostrar o funcionamento de um sumarizador automático para a língua portuguesa e outro para outra língua estrangeira, para que pudesse ser feita uma análise e breve comparação entre ambos. O estudo e a análise desses sistemas observaram se estes são realmente eficientes para poder substituir a sumarização humana e qual o nível de sua relevância para a era atual, a era da informação.

O sistema de sumarização automática GistSummarizer, foi o primeiro sistema disponibilizado para a língua portuguesa.É um sistema bem elaborado, que utiliza a abordagem superficial e consegue obter bons resultados. É um sistema exemplo para futuros sumarizadores automáticos para a língua portuguesa, pois utiliza métodos precisos e fornece um sumário de qualidade. As inovações que esse sistema de sumarização trouxe para a língua portuguesa são de grande relevância para sumarizar textos acadêmicos e jornalísticos. Ainda é um sistema relativamente novo e que precisa de alguns aprimoramentos.

Já o sumarizador automático SweSum foi feito inicialmente para a sumarização de textos escritos na língua sueca. Depois de aprimoramentos, foi possível utilizá-lo para sumarização automática de outras línguas, como por exemplo: francês, espanhol e inglês. É um sumarizador completo e conta com uma interface acessível, já que o idioma pode ser alterado pelo usuário para que possa ser compreendido o conteúdo da página. Apesar de ainda ser disponível para um número limitado de idiomas, o SweSum é um sistema de sumarização reconhecido, principalmente para sumarização de textos da língua inglesa.Conta com um bom nível de usabilidade, sendo de fácil acesso para os usuários que não têm um bom conhecimento acerca de sistemas de sumarização automática.

Essa pesquisa foi feita com o objetivo de contribuir para toda a comunidade de usuários web, aqueles que necessitam de informações de forma rápida e buscam por sistemas de sumarização eficientes. O estudo traça as vantagens e desvantagens oferecidas pelos sistemas de sumarização apresentados, de forma a mostrar que ainda existem pontos que

podem ser aprimorados para que se garanta um bom nível de usabilidade, assim como a qualidade dos sumários produzidos por estes sumarizadores, o que pode ser alterado para que eles se tornem mais precisos e excelentes. As propostas que surgiram dessa análise auxiliarão aqueles que necessitam utilizar os sistemas de sumarização, independente da finalidade, se é para pesquisas pessoais ou para utilizar em atividades profissionais..

2. Adequações sobre o uso dos sumarizadores automáticos

Após análise dos dois sistemas de sumarização e do campo de Sumarização Automática, foram apontadas algumas falhas relativas à usabilidade e aos textos de saída, ou seja, os sumários produzidos. Essas observações foram feitas para que o usuário do sistema sinta-se mais confortável quando estiver utilizando o sumarizador. São questões relativas aos idiomas da interface dos sumarizadores, explicações dos campos e adequações do texto de saída.

Sendo assim, as alterações propostas foram a tradução da interface do GistSumm também para a língua portuguesa, já que foi o primeiro SSA disponível para esse idioma, a explicação de campos para leigos em SA,como a taxa de compressão e o sumário com o título do texto-fonte e seus respectivos tópicos, para que o usuário do sistema tenha um sumário de boa qualidade.

3. Proposta de criação de um filtro auxiliar

Após análise da estrutura interna de sumarizadores automáticos e de suas interfaces gráficas, foi proposta a criação de um software auxiliar para dar suporte à sumarização automática. Esse software terá como objetivos excluir sessões do texto-fonte que não são interessantes para o sumário e converter os arquivos de entrada para *.txt*, auxiliando assim o funcionamento desses sistemas.

O corpus utilizado para analisar esses textos e comparar as sessões que são obrigatórias ou ocorrem com frequência em textos acadêmicos foi a BDTD, disponível em: < http://bdtd.ibict.br/vufind/> .

4. Proposta de estudos futuros sobre SA e outros campos de processamento automático das línguas naturais

Com a analise do funcionamento dos SSA e a proposta de criação de um filtro auxiliar para apoiar a SA, é possível realizar pesquisas mais aprofundadas a cerca dos métodos

utilizados pela SA, novos dispositivos que possam servir de suporte para SA para otimizar seu processo e sobre as aplicações da sumarização automática, assim como a tradução automática, que utiliza sistemas de tradução automática, que também são subáreas da PLN.

Propostas para a otimização desses sistemas, ajudariam muito o campo da PLN, que está em evidência, devido aos computadores já terem se tornado parte da vida dos seres humanos.A otimização de processos como a sumarização automática traria muitos benefícios para a sociedade atual, que depende em grande parte de computadores.

CONSIDERAÇÕES FINAIS

De acordo com os estudos realizados nesse trabalho acerca de sumarização automática, podemos concluir que o uso de sumarizadores automáticos é muito importante para auxiliar os usuários web na atual conjuntura, onde a informação é disponibilizada virtualmente em larga escala. Como pode ser visto,o uso da sumarização automática alcança pessoas que estão em busca de informações jornalísticas na web, aquelas que necessitam de materiais para uma determinada pesquisa, assim como teses e dissertações e estrangeiros com pouco domínio da língua portuguesa que precisam buscar por determinado material disponibilizado apenas no idioma português. Obviamente, a SA ainda tem muito a ser aprimorada, mas com certeza será um trabalho gradual e extremamente útil na sociedade da informação.

Os estudos a cerca do processo realizado por um sumarizador automático para confeccionar os sumários, nos mostra que é um trabalho minucioso, e que precisa de muita cautela e pesquisas para que os resultados obtidos nos sumários possam ser satisfatórios. Tomando por base de estudos os sumarizadores automáticos GistSumm e SweSum, pode-se perceber um pouco da sua estrutura e como eles funcionam. Foi possível mostrar que ainda podem ser aprimorados, como por exemplo, a interface do GistSumm que apesar de ter sido o primeiro SSA disponível para a língua portuguesa, conta apenas com interface na língua inglesa.Como possíveis modificações, foi sugerido uma interface também na língua portuguesa.Também, foi sugerido um *filtro auxiliar*, que conta com métodos similares ao da abordagem superficial da SA, para que seja otimizado o processo de sumarização automática.

É importante ressaltar o uso dos tradutores automáticos. A tradução automática não é o foco do presente trabalho, porém está interligada com a SA. A tradução automática também é uma das subáreas da PLN e de muita importância para o processamento automático das línguas naturais.Outros trabalhos podem vir a explorar com mais profundidade o uso dos tradutores automáticos.

Finalmente, em uma era onde a informação se tornou o bem mais valioso, são necessários sistemas que possam ajudar a difundi-la para todos que necessitam.A sumarização automática como um processo de linguagem natural, auxilia nessa difusão de forma prática e rápida.É necessário que os profissionais da área da informação comecem a utilizar esses sistemas como auxilio para disponibilização da informação.

REFERÊNCIAS BIBLIOGRÁFICAS

ALEIXO, Priscila; PARDO, Thiago Alexandre Salgueiro. *CSTNews: um Córpus de Textos Jornalísticos Anotados segundo a Teoria Discursiva Multidocumento CST (Cross-document Structure Theory).* Série de Relatórios do Núcleo Interinstitucional de Linguistica Computacional NILC - ICMC - USP, NILC-TR-08-05, 2008, P 1- 8.

CUNHA, Murilo Bastos Da; CAVALCANTI, Cordélia Robalinho de Oliveira; *Dicionário de Biblioteconomia e Arquivologia.* Brasília: Briquet de Lemos/Livros, 2008,P.109,323,348.

DALIANIS, Hercules. *SweSum - a text summarizer for Swedish.*Interactions and Presentation Laboratory (IPLab), Numerical Analysis and Computing Science (NADA), Royal Institute of Technology (KTH), 2000, P.2.

DIAS, S. Márcio; GARAY,Alessandro Y. Bokan; CHUMAN Carla; BARROS, Cláudia D.; MAZIERO, Erick G; NOBREGA, Fernando A.A; SOUZA, Jackson W.C; CABEZUDO, Marco A. Sobrevilla; DELEGE, Marina; JORGE, María Lucía R. Castro; SILVA, Naira L.; CARDOSO, Paula C.F; FILHO, Pedro P. Balage; CONDORI, Roque E. López; MARCASSO Vanessa; FELIPPO, Ariani Di; NUNES, Maria das Graças V.; PARDO, Thiago A.S.*Enriquecendo o Córpus CSTNews - a Criação de Novos Sumários Multidocumento.*Universidade de São Paulo, 2014, P.1,2.

GRICE, Paul H. *Logic and Conversation*. In P. Cole and J. L. Morgan (eds.), Syntax and Semantics. Volume 3: Speech Acts, pp. 41-58. Academic Press, New York, 1975, P, 45-47.

GUELPELI, Marcus Vinicius Carvalho. *Cassiopeia : Um Modelo de Agrupamento de Textos Baseado em Sumarização.* Tese (Doutorado em Computação) - Niterói: Universidade Federal Fluminense, 2012, P.40,72.

MARTINS, Camilla Brandel; PARDO, Thiago Alexandre Salgueiro; ESPINA, Alice Picon; RINO, Lucia Helena Machado. *Introdução à Sumarização Automática.* Relatório Técnico RT -DC 002. Departamento de Computação, Universidade Federal de São Carlos.São Carlos-SP, Fevereiro, 2001, P.9-30.

PARDO, Thiago Alexandre Salgueiro. *Sumarização Automática: Principais Conceitos e Sistemas para o Português Brasileiro.* Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional NILC - ICMC - USP, 2008, P 1-7.

PARDO, Thiago Alexandre Salgueiro. *GistSumm: Um sumarizador Automático Baseado na Idéia Principal de Textos.* Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional NILC - ICMC - USP, NILC-TR-02-13, 2002, P 6,19.

RINO, Lucia Helena Machado. *Modelagem de Discurso para o Tratamento da Concisão e Preservação da Idéia Central na Geração de Textos.* Tese(Doutorado)-São Carlos: Instituto de Física de São Carlos, 1996, P.54.

ANEXO A - GLOSSÁRIO

C

CORPUS- "Reunião de fontes, isto é, informações e documentos orais ou escritos, concernentes a uma área específica de conhecimento". (Dicionário de Biblioteconomia e Arquivologia, 2008).

R

RESUMO - "Representação concisa e acurada do conteúdo de um documento; síntese de um documento; notação de conteúdo".(Dicionário de Biblioteconomia e Arquivologia,2008).

S

SOFTWARE - "Conjunto de programas que permite ao usuário escrever, depurar e modificar programas aplicativos." (Dicionário de Biblioteconomia e Arquivologia, 2008).

SUMÁRIO - "Pode tanto se referir a índice quanto a resumo propriamente dito.Em português, o termo sumário foi importado da língua inglesa e, por isso, foi adotado como padrão nos trabalho da área (Sumarização Automática)"(PARDO, 2008).

U

USABILIDADE - "Grau com que usuários específicos podem alcançar metas específicas em determinado ambiente, com eficácia, comodidade e de modo aceitável". (Dicionário de Biblioteconomia e Arquivologia, 2008).