

COMPARANDO SUMÁRIOS DE REFERÊNCIA HUMANOS COM EXTRATOS IDEAIS NO PROCESSO DE AVALIAÇÃO DE SUMÁRIOS EXTRATIVOS

Carlos Henrique Delgado
*UBM – Universidade de Barra Mansa
Barra Mansa – Rio de Janeiro – Brasil
henrique_chd@yahoo.com.br*

Caroline Evangelista Vianna
*UBM – Universidade de Barra Mansa
Barra Mansa – Rio de Janeiro – Brasil
caroline_evianna@yahoo.com.br*

Marcus Vinicius C. Guelpli
*UBM – Universidade de Barra Mansa
Barra Mansa – Rio de Janeiro – Brasil
guelpli@ubm.br*

RESUMO

O acesso aos meios de distribuição da informação está cada vez mais facilitado. Por isto, empresas e instituições das mais diversas áreas do conhecimento têm disponibilizado muitos de seus produtos em meios digitais. Devido ao grande volume de informações e ao pouco tempo para as leituras dos textos de diversas naturezas surgiu a necessidade de simplificar e resumir essas informações a fim de otimizar o processo de absorção dos conteúdos pelos leitores. O principal desafio da sumarização automática de textos é a obtenção de sumários automáticos com qualidade semelhante aos sumários gerados por seres humanos. Contudo, para medir o quão se aproximam os sumários automáticos dos sumários humanos faz-se necessária uma avaliação comparativa entre ambos. Essa avaliação geralmente é feita por seres humanos ou também pode ser realizada automaticamente por ferramentas de avaliação. Este trabalho apresenta um estudo detalhado destas diferentes formas de avaliação de sumários automáticos e tem por hipótese, mostrar que a avaliação baseada em sumários de referência humanos são melhores do que a avaliação baseada em extratos ideais gerados automaticamente.

PALAVRAS-CHAVE

Extratos Ideais, Sumários de Referência Humanos, Sumários Automáticos, Rouge e Avaliação.

1. INTRODUÇÃO

Sumários são textos reduzidos que transmitem as ideias principais e mais relevantes de um texto original, de forma clara e objetiva sem perda da informatividade. A utilização de sumários é muito útil atualmente, considerando a grande quantidade de informações disponíveis e a pouca disponibilidade de tempo para a leitura.

A sumarização automática visa à produção de sumários gerados automaticamente baseando-se em seu texto original. O principal objetivo da sumarização automática é extrair conteúdo a partir de um texto original gerando um texto reduzido com as informações mais relevantes (Mani, 2001).

De acordo com (Mani, 2001), um sumário pode ser denominado *extract* (extrato) ou *abstract* (resumo). Um extrato é um texto onde as sentenças utilizadas são copiadas e organizadas seguindo a sequência do texto original. No entanto, o resumo é composto por sentenças reescritas ou rearranjadas, não se limitando apenas à cópia das sentenças do texto original. Essa distinção diferencia os sumários pela forma como o conteúdo a ser

sumarizado é organizado no texto, sendo assim o foco deste trabalho são os Sumários Automáticos (SA) extrativos.

O grande problema desta área é gerar SA que não percam sua informatividade. Sendo assim a avaliação de SA é muito subjetiva, pois depende fundamentalmente da avaliação humana e também por não existir um consenso entre os avaliadores humanos frente aos resultados obtidos com os SA. Para aumentar a complexidade desta avaliação, existem ainda dois fatores determinantes que são: a quantidade de SA gerados e a quantidade de tempo gasto para avaliação dos SA.

Este trabalho apresenta um estudo da avaliação de SA tendo como base Extratos Ideais e a avaliação baseando-se em sumários de referência humanos. Para realizar os experimentos neste trabalho, utilizou-se o corpus TeMário 2004(Pardo e Rino, 2003), que contém 100 textos jornalísticos com seus respectivos extratos ideais gerados pela ferramenta Gerador de Extrato Ideal(GEI) (Pardo e Rino, 2004) e seus sumários de referência humanos.

Para a avaliação dos SA, foi utilizado o pacote *Recall-Oriented Understudy for Gisting Evaluation*(ROUGE), (Lin e Hovy, 2003) adotado em conferências internacionais dedicadas ao tema, onde o mais importante chama-se TAC (*Text Analysis Conference*) que é realizado anualmente. O ROUGE utiliza sumários de referência humanos para comparação com SA. Outra forma adotada para a avaliação de sumários foi a comparação dos SA com seus respectivos extratos ideais gerados pelo GEI.

O objetivo deste trabalho consiste em comparar as duas formas de avaliação: uma por sumários de referência humanos e outra por extratos ideais, sendo assim, a hipótese deste trabalho consiste em mostrar que avaliar por meio de sumários de referência humanos é melhor do que por extratos ideais. A sumarização foi baseada em número de sentenças para compor os SA. A métrica *Recall*, *Precision* e *F-Measure* foram usadas para mensurar os resultados. Visando a validação dos resultados obtidos, foram utilizados os testes estatísticos ANOVA de Friedman e o coeficiente de concordância de Kendall.

Os resultados obtidos são animadores, pois demonstram que a hipótese deste trabalho manteve-se em todos os experimentos realizados para o domínio escolhido.

Na sessão 2 apresenta-se uma abordagem dos conceitos principais da técnica da SA Extrativo. São apresentados na sessão 3, os diferentes tipos de sumários de referência utilizados para a avaliação. A sessão 4 descreve com mais detalhes os experimentos realizados neste trabalho, os algoritmos usados na SA e as formas utilizadas para a avaliação. Na sessão 5 serão apresentados os resultados obtidos por meio da pesquisa realizada. Por fim, a sessão 6 apresenta as conclusões e os trabalhos futuros.

2. SUMARIZAÇÃO AUTOMÁTICA EXTRATIVA

A Sumarização começou a ser estudada na década de 50 com o método das palavras-chaves proposto por Luhn (1958). A partir deste marco, foram desenvolvidos e estudados outros métodos de sumarização, todos com o mesmo propósito, o de obter sumários automáticos com qualidades próximas aos sumários gerados por seres humanos.

A sumarização extrativa segue a abordagem empírica definida por (Pardo e Rino, 2003), por onde se tem métodos estatísticos ou superficiais que identificam os seguimentos mais relevantes do texto-fonte, produzindo os extratos através da justaposição das sentenças extraídas, sem qualquer modificação.

Os textos originais são processados na etapa inicial de análise e uma representação interna com todo o conteúdo dos textos é produzida. Essa representação deve ser formal o suficiente para ser processada automaticamente. A etapa de redução realiza o processo de sumarização com número de sentenças desejadas para obter o sumário. Por fim, na etapa de síntese, as sentenças extraídas são organizadas para a formação do sumário automático.

3. AVALIAÇÃO DE SISTEMAS DE SUMARIZAÇÃO

Em 1996, Sparck Jones e Galliers apresentaram um modelo para a avaliação de sistemas de sumarização automática. Todos os tópicos que deveriam ser levados em consideração na avaliação destes sistemas são descritos neste documento. De acordo com Sparck Jones e Galliers (1996), a avaliação de sistemas de SA pode ser classificada como intrínseca, quando é avaliada a qualidade do próprio sistema, ou extrínseca,

quando é avaliada sua utilização em outras aplicações. Uma avaliação também pode ser dita on-line, quando faz uso de humanos para julgar os sumários e quando se usa apenas métricas automáticas para a avaliação, é chamada off-line. Por último, classifica-se uma avaliação como autônoma ou comparativa. A autônoma é realizada quando se julga o sistema separadamente ou individualmente, já na avaliação comparativa, se estabelecem métricas em comum entre diversos sumarizadores para uma comparação entre os mesmos.

Seguindo as classificações de (Sparck Jones e Galliers, 1996), as avaliações abordadas nesse trabalho são as intrínseca, off-line e comparativa. Para realizar a avaliação off-line, faz-se necessário o uso de boas métricas automáticas. Os sumários Extrativos são avaliados pela sua informatividade através da medida de precisão (*Precision*) e cobertura (*Recall*) e a média *F-Measure*. As fórmulas apresentadas encontram-se em (Guelpli, 2008).

- *Recall*: número de sentenças do sumário automático presentes no sumário de referência / número de sentenças do sumário automático;
- *Precision*: número de sentenças do sumário automático presentes no sumário de referência / número de sentenças do sumário de referência;
- *F-Measure*: $((P \times R) / (P + R)) \times 2$, o *F-Measure* combina as métricas de *Recall* e de *Precision*. O resultado *F-Measure* é um indicativo de que, quanto mais próximo de 1, melhor é o sumário e resultados mais próximos de 0, demonstram que os sumários são ruins.

3.1 Sumários de referência

Neste trabalho serão abordados dois tipos de sumários de referência. O primeiro será o extrato ideal, gerado pela ferramenta GEI, e o outro será o sumário de referência humano.

3.1.1 Sumários gerados pela ferramenta GEI

O objetivo da ferramenta GEI é gerar Extratos Ideais automaticamente, baseando-se em sumários de referência humanos. Essa ferramenta foi desenvolvida seguindo a metodologia da medida do co-seno de Salton (1970;1989), que utiliza matriz de similaridade. O GEI gera seus Extratos ideais tendo como referência os textos originais e seus Sumários de Referência Humanos.

3.1.2 Sumários de referência gerados por humanos

Os sumários de referência humanos são aqueles gerados por uma pessoa a partir de um texto original. Para a obtenção destes sumários é necessário que uma pessoa entenda o assunto (especialista) a ser lido a fim de identificar e captar a ideia principal do texto. Depois da identificação do assunto principal é produzido um sumário com todo o entendimento do texto original pelo especialista. Esse sumário não é extrativo, e sim reescrito com transformações textuais.

3.2 Avaliação ROUGE

Atualmente, uma das métricas mais utilizadas é a chamada ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) (Lin e Hovy, 2003). Basicamente, a ROUGE é um pacote de avaliação de sumários, que surgiu para possibilitar a comparação direta entre um SA e o respectivo sumário humano usando como métrica n-gramas. Desta forma, pode-se descartar a avaliação com extratos ideais (GEI). Com essa ferramenta é possível analisar a proximidade de qualidade dos SA em relação aos sumários de referência humano. De uma forma geral, a ROUGE calcula o grau de informatividade dos extratos.

Segundo as definições de (Lin e Hovy, 2003), a Rouge faz o cálculo baseando-se nos conjuntos de palavras em sequência (n-gramas) em comum entre os sumários de referência humanos e os SA. Quanto maior o número de palavras em comum entre os sumários, maior será a nota dada ao SA.

3.3 Avaliação pela métrica do Perfil

O sumarizador Perfil (Guelpli,2008) possui uma funcionalidade capaz de gerar resultados baseados nas métricas *Recall*, *Precision* e *F-Measure*.

Para fazer a avaliação dos SA pelo Perfil, coloca-se texto gerado por qualquer SA, o texto original e o extrato ideal (gerado pelo GEI) a partir do texto original.

Com as entradas de textos definidas, a funcionalidade do Perfil fará a comparação entre o SA e o extrato ideal, gerando então os resultados.

4. EXPERIMENTOS

Para a realização dos experimentos foi utilizado o corpus de textos jornalísticos - TeMário 2004 (Pardo e Rino, 2003). Este corpus é dividido em cinco sessões contendo vinte textos cada uma com seus respectivos sumários de referência humanos e extratos ideais gerados pela ferramenta GEI (Gerador de Extratos Ideais). As sessões do TeMário 2004 são divididas em: Especial, Internacional, Mundo, Política e Opinião.

Neste trabalho foram utilizados sumarizadores da literatura, Perfil (Guelpeli, 2008), GistSumm(Pardo, 2003), o método Text Mining (Espina e Rino, 2002) e um sumarizador profissional - Subject Search Summarizer de propriedade da Krlyoff Technologies. Foi ainda implementado um sumarizador com uma função aleatória denominado Baseline.

Os experimentos foram realizados basicamente em duas grandes etapas. A primeira etapa foi a da avaliação utilizando extratos ideais e a segunda foi com sumários de referência humanos. A funcionalidade do Perfil foi usada como avaliação em um primeiro momento e depois foi utilizada a avaliação pelo pacote ROUGE com intuito de uniformizar as saídas para as métricas, todos com n-gramas.

Com a utilização dos cinco métodos de sumarização aplicados a cada um dos 100 textos do corpus TeMário 2004 (Pardo e Rino, 2003), foi gerado um total 500 sumários, ou seja, 100 sumários para cada método.

Para a validação dos resultados foi utilizado os testes estatísticos ANOVA de Friedman e o coeficiente de concordância de Kendall, métodos estes usados em conferências internacionais de avaliação de SA como TAC (*Text Analysis Conference*), a mais relevante dentro da área de SA.

4.2 Experimentos com Extratos Ideais

Para a realização dos experimentos utilizando os extratos ideais foi necessária a geração dos SA para os cem textos do corpus TeMário 2004, utilizando o sumarizador Perfil. Baseando-se no número de sentenças geradas em cada sumário do Perfil, foram gerados os SA nos outros sumarizadores para uma avaliação justa.

Com os sumários gerados por cada sumarizador, foi utilizado posteriormente um recurso do Perfil capaz de gerar resultados baseados nas métricas *Recall*, *Precision* e *F-Measure*. Para a geração dos resultados com estas métricas, fez-se necessário alguns parâmetros de entrada descritos na figura 1.

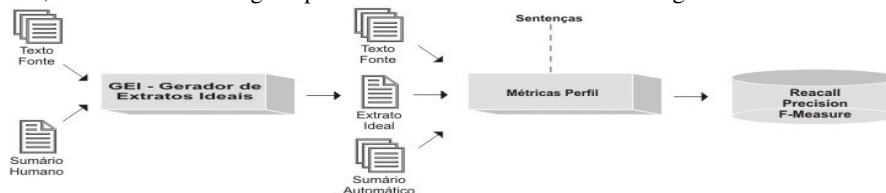


Figura 1. Avaliação pelos Extratos Ideais.

Após a entrada do texto original, extrato ideal e o sumário automático de cada sumarizador, a funcionalidade fará a comparação entre o SA e o extrato ideal, gerando então os valores para as métricas.

Tendo os resultados gerados, foram organizados todos os resultados para a realização dos cálculos da média, do coeficiente de variação, da variância, do desvio padrão, da soma acumulada dos resultados, e dos testes estatísticos ANOVA de Friedman e do coeficiente de concordância de Kendall.

4.3 Experimentos com Sumários de Referência Humanos

Para a realização dos experimentos com o pacote ROUGE, utilizou-se os sumários gerados automaticamente pelos sumarizadores e os sumários de referência humanos contidos no TeMário 2004 (Pardo e Rino, 2003).

Na figura 2 mostra-se uma estrutura simplificada da geração do processo de avaliação do pacote ROUGE.

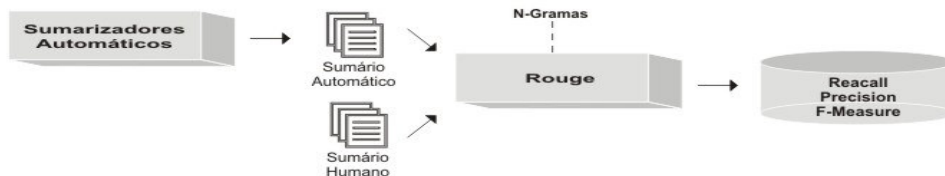


Figura 2. Esquema simplificado da avaliação de sumários utilizando-se a ferramenta ROUGE.

Como mostrado na figura 2, a ferramenta ROUGE necessita de um sumário de referência humano e o SA. Após a entrada dos textos, o pacote ROUGE gera os resultados baseando-se no conjunto de palavras em sequência (n-gramas) para a comparação entre os sumários. Os resultados finais mostram os valores dos SA com as métricas *Recall*, *Precision* e *F-Measure* baseado no sumário de referência humano.

A figura 3 apresenta a estrutura para o experimento utilizando os extratos ideais do GEI desempenhando o papel de sumário de referência humano da figura 2.

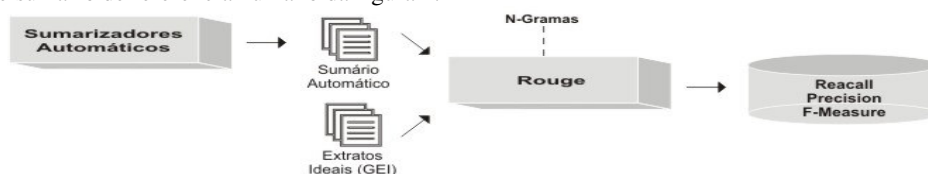


Figura 3. Estrutura da avaliação baseada nos extratos ideais (GEI) como sumários humanos.

5. RESULTADOS OBTIDOS

Com os resultados dos experimentos obteve-se um ordenamento dos sumarizadores. A tabela 1 mostra os resultados dos diferentes experimentos usando a medida *F-Measure* que é uma medida harmônica entre as medidas *Recall* e *Precision*.

Tabela 1. Ordenamento dos sumarizadores e métodos utilizados.

F-MEASURE					
Extratos Ideais Métricas Perfil (Figura 1)		Avaliação por Sumários de Referência Humanos (Figura 2)		Extratos Ideais como Sumários de Referência Humanos (Figura 3)	
		GEI	0,5601		
Perfil	0,4217	Perfil	0,4602	Perfil	0,6180
Subject	0,3469	Subject	0,4584	Subject	0,5851
Baseline	0,3154	Baseline	0,4522	Baseline	0,5576
GistSumm	0,3129	GistSumm	0,4343	GistSumm	0,5393
TextMining	0,2580	TextMining	0,4066	TextMining	0,4950

A figura 4 mostra os resultados obtidos aplicando-se a avaliação baseada nos extratos ideais representado na figura 1 e descrito na tabela 1.

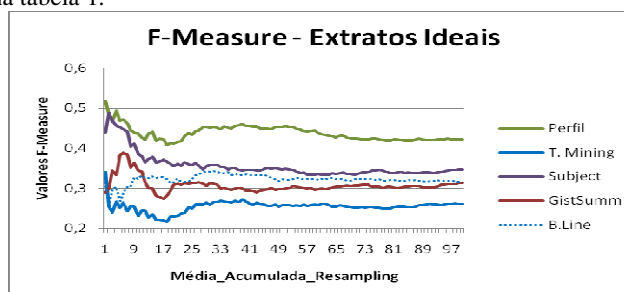


Figura 4. Resultados baseados na avaliação por extratos ideais.

De acordo com a figura 4, o algoritmo Perfil obteve o melhor resultado. A pior avaliação foi do método Text Mining. O Baseline representado pela linha pontilhada está abaixo do Perfil e do Subject e acima do algoritmo GistSumm e do método Text Mining.

Analisando a figura 4, observa-se que o algoritmo Perfil tem uma diferença significativa dos demais. Sendo assim, foram utilizados os testes estatísticos de ANOVA de Friedman e o coeficiente de concordância de Kendall para verificar se existe tal diferença em todas as amostras. O primeiro teste, ANOVA de Friedman – um teste não paramétrico, onde o objetivo envolve um ranking de cada fileira aleatória (ou bloco) junto, por sua vez, considera os valores dos *Ranks* por colunas de cada amostra e desta forma avalia se todas as amostras são significativamente diferentes produzindo um ordenamento de seus valores crescentes. Isto pode ser observado na tabela 2. O segundo teste, coeficiente de concordância de Kendall – também não paramétrico, usado para normalizar teste estatístico ANOVA de Friedman, tem a finalidade de gerar uma avaliação de concordância ou não com *Ranks* estabelecidos. Quanto mais próximo de 0, menor é a concordância e quanto mais próximo de 1, maior é a concordância. Desta forma pode-se observar que os valores de ordem médios e o coeficiente de concordância de Kendall estão próximos de 1, respectivamente 0,971858586 e 0,97214.

Tabela 2. Teste estatístico com Extratos Ideais.

Comparando amostras múltiplas selecionadas - Extratos Ideais Métricas Perfil			
N	100	Graus de liberdade	4
qui-quadrado	388,856	p-nível	0
Coef. de concordância de Kendall	0,97214	Ordem médio	0,971858586
	Ordem médio	Soma de ordens	Média
Perfil	4,98	498	0,437497
T. Mining	1,02	102	0,254668
Subject	4,02	402	0,356174
GistSumm	2,09	209	0,30759
B.Line	2,89	289	0,322513

A figura 5 apresenta os resultados produzidos pela avaliação baseando-se em sumários de referência humanos através do pacote ROUGE, representado na figura 2 e descrito na tabela 1.

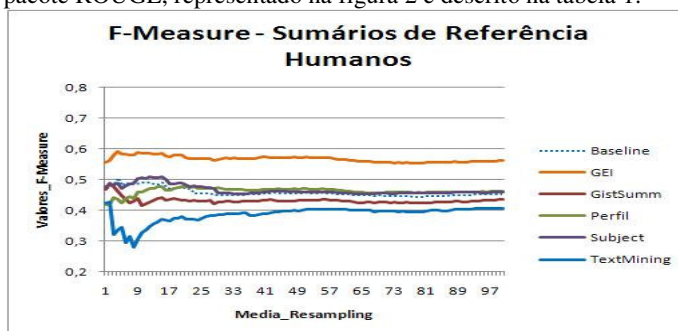


Figura 5. Resultados baseados na avaliação por sumários de referência humano no pacote ROUGE.

A figura 5 mostra os extratos ideais criados pelo GEI como um SA. Isto serve para identificar o quanto ele se aproximou do sumário de referência humano. Sendo assim, o GEI obteve o melhor desempenho seguido pelo algoritmo Perfil. Novamente pode-se constatar que o método Text Mining e o algoritmo GistSumm obtiveram o resultado abaixo do Baseline. A tabela 3 mostra os testes estatísticos usando ANOVA de Friedman e o coeficiente de concordância de Kendall. Pode-se observar que os valores de ordem médios e o coeficiente de concordância de Kendall estão próximos de 1, respectivamente 0,912438672 e 0,913314286 provando assim o ordenamento correto estabelecido pela ferramenta ROUGE dentro da comparação entre sumário de referência humano e os SA do experimento.

Tabela 3. Teste estatístico com Sumários de Referência Humanos no pacote ROUGE.

Comparando amostras múltiplas selecionadas - Avaliação por Sumários de Referência Humanos			
N	100	Graus de liberdade	5
qui-quadrado	45,6571429	p-nível	0
Coef. De concordância de Kendall	0,913314286	Ordem médio	0,912438672
	Ordem médio	Soma de ordens	Média
B.Line	3,22	322	0,456926105
GistSumm	2,07	207	0,430939741
Perfil	4,47	447	0,461595424
Subject	4,22	422	0,465141291
T.Mining	1,02	102	0,385953463

A Figura 6 mostra os resultados gerados utilizando os extratos ideais como os sumários de referência humanos representado na figura 3 e descrito na tabela 1. De acordo com a figura 6, o algoritmo Perfil obteve novamente o melhor resultado. O Baseline continua sendo maior que o método Text Mining e algoritmo GistSumm.

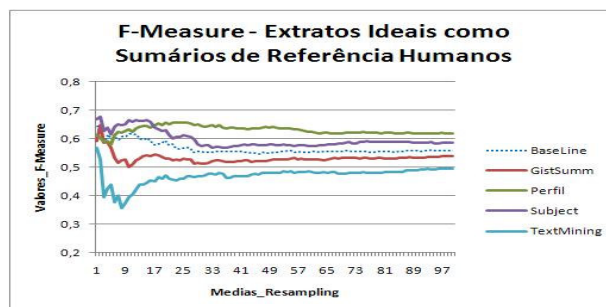


Figura 6. Resultados baseados em extratos ideais como sumários de referência humanos.

Conforme mostrado na tabela 4 observa-se que os valores de ordem médios e o coeficiente de concordância de Kendall estão próximos de 1, respectivamente 0,950606061 e 0,9511, verificados por meio dos testes estatísticos ANOVA de Friedman e o coeficiente de concordância de Kendall. Importante ressaltar que os testes estatísticos ANOVA de Friedman e o coeficiente de concordância de Kendall são teste oficiais do principal evento de avaliação de sumários que é realizado anualmente o TAC (*Text Analysis Conference*).

Tabela 4. Teste estatístico com Extratos Ideais como Sumários de Referência Humanos no pacote ROUGE.

Comparando amostras múltiplas selecionadas - Extratos Ideais como Sumários de Referência Humanos			
N	100	Graus de liberdade	4
qui-quadrado	380,44	p-nível	0
Coef. De concordância de Kendall	0,9511	Ordem médio	0,950606061
	Ordem médio	Soma de ordens	Média
B.Line	3,03	303	0,564558495
GistSumm	2,02	202	0,530585116
Perfil	4,79	479	0,628786933
Subject	4,16	416	0,595939862
T.Mining	1	100	0,469297352

Para exemplificar a figura 7, foi utilizado o melhor sumariador (Perfil), com seus resultados obtidos na tabela 1 (Extratos Ideais como Sumário de Referência Humano). Pode-se observar que o SA do Perfil obteve 61% de proximidade aos extratos ideais, sendo então, 39% distante dos extratos ideais. A distância do sumário humano para os extratos ideais foram calculados a partir dos resultados obtidos na tabela 1 (GEI como sumariador), onde obteve-se uma proximidade dos extratos ideais de 56% em relação ao sumário de referência humano, sendo 44% distantes do humano. Somando-se as distâncias totais entre os sumários humanos, extratos ideais e SA, têm-se um total de 83% de distância dos SA para os sumários de referência humanos.

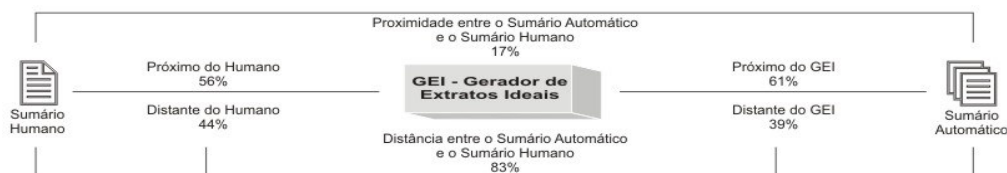


Figura 7. Distância entre Sumários Automáticos e Sumários Humanos passando pelo GEI.

Já na avaliação direta entre SA e sumários de referência humanos, como mostrado na figura 8, tem-se uma diminuição da distância entre os mesmos. Neste exemplo manteve-se apenas 54% de distância entre os SA gerados em relação aos humanos, sendo 46% a proximidade entre os sumários.

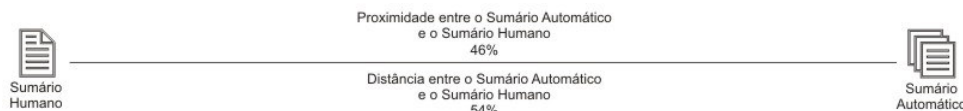


Figura 8. Distância entre Sumários Automáticos e Sumários Humanos.

6. CONCLUSÃO

De acordo com (Rino e Pardo, 2003), os extratos ideais podem não ser realmente ideais por nem sempre obterem o conteúdo relevante do texto original a ser sumarizado, como o faria um especialista humano. A metodologia do co-seno pode produzir extratos com sentenças insignificantes.

Os extratos ideais foram criados para possibilitar a avaliação de SA, uma vez que estes sumários são extrativos não sendo possível a comparação com os sumários de referência humanos, que contêm reescritas e adequações textuais. Mesmo em avaliações baseadas em extratos ideais, que por sua vez se baseiam em sumários humanos, observa-se notoriamente que a avaliação baseada em sumários de referência humanos é a melhor forma de avaliar o SA. Pode-se afirmar que na avaliação baseada nos extratos ideais do GEI, os SA ficam mais distantes dos sumários gerados por humanos. Importante destacar que os resultados obtidos se limitam a aplicação dos testes em domínio jornalístico, não sendo possível ainda, obter estas conclusões em outros domínios não estudados, sendo este um trabalho interessante para o futuro.

REFERÊNCIAS

- ESPINA, A.P and RINO, L.H.M, 2002, Utilização de Métodos Extrativos na Sumarização Automática de Textos. Relatórios Técnicos (NILC-TR-02-06). NILC – ICMC – USP. São Carlos, Brasil.
- GUELPELI, M.V.C and BERNARDINI, F.C and GARCIA, A.C.B, 2008, Todas as Palavras da Sentença como Métrica para um Sumarizador Automático. *Tecnologia da Informação e da Linguagem Humana – TIL*. Vila Velha, Brasil.
- LIN, C-Y. and HOVY, E.H., 2003, Automatic Evaluation of Summaries Using N-gram Cooccurrence Statistics. *In the Proceedings of the Language Technology Conference*. Edmonton, Canada.
- LUHN, H. P., 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, Vol. 2, pp. 159-165.
- MANI, I., 2001, Automatic Summarization. *John Benjamins Publishing Co.*, Amsterdam. Estados Unidos.
- PARDO, T.A.S. and RINO, L.H.M. and NUNES, M.G.V., 2003, GistSumm: A Summarization Tool Based on a New Extractive Method. *In the Proceedings of the 6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken*. Faro, Portugal.
- PARDO, T.A.S. and RINO, L.H.M., 2003, TeMário: Um Corpus para Sumarização Automática de Textos. Relatórios Técnicos (NILC-TR-03-09). NILC – ICMC – USP. São Carlos, Brasil.
- PARDO, T.A.S. and RINO, L.H.M., 2004, Descrição do GEI – Gerador de Extratos Ideais para o Português do Brasil. Relatórios Técnicos (NILC-TR-04-07). NILC – ICMC – USP. São Carlos, Brasil.
- SPARCK JONES, K. and GALLIERS, J.R., 1996. Evaluating Natural Language Processing Systems. *Lecture Notes in Artificial Intelligence*, Vol. 1083.