Uso de sumários humanos nos idiomas português e inglês no modelo Cassiopeia

Jésyka Milleny Az. Gonçalves, Marcus Vinicius C. Guelpeli

Departamento de Sistemas e Computação
Universidade Universidade Federal dos Vales do Jequitinhonha e Mucuri – (UFVJM) – Diamantina, MG – Brazil

jesykamilleny@gmail.com, marcus.guelpeli@ufvjm.edu.br

Abstract.*This study proposes the modification in the pre-processing model Cassiopeia, which was originally used in automatic summarization. In our study, the use of human summaries, comparing and checking which provides, in Cassiopeia, a result with the best grouping. The analysis is performed with the same areas of the original model, medical and legal, in Portuguese and English. The performance is measured in terms of the internal and external metrics and the end has the efficiency results from the use or non-human in cluster summaries Cassiopeia model.*

Keywords: *Summarization, Grouping, Clustering, Cassiopeia, Metric Internal, External Metrics, Statistical Tests.*

Resumo. *Este estudo propõe a modificação na etapa de pré-processamento do modelo Cassiopeia, onde originalmente foi usado a sumarização automática. Neste trabalho houve a utilização de sumários humanos, comparando-os e verificando qual proporciona, ao Cassiopeia, um resultado com o melhor agrupamento. A análise é realizada com os mesmos domínios do modelo original, médico e jurídico, nos idiomas português e inglês. O desempenho será medido em termos das métricas internas e externas e ao ao final tem-se os resultados da eficiência ou não do uso de sumários humanos no agrupamento do modelo Cassiopeia.*

Palavras-Chave*: Sumarização, Agrupamento, Clusterização, Cassiopeia, Métricas Internas, Métricas Externas, Testes Estatísticos.*

1. Introdução

Segundo Guepeli *et al.* (2010) atualmente tem-se valorizado muito a área da Mineração de Textos – MT, pois o avanço das tecnologias vem ocasionando a alta velocidade de acesso e propagação de dados na internet, o que dificulta cada vez mais às pessoas o tratamento e assimilação destas informações, uma vez que a capacidade humana de leitura e registro é limitada. Zoubi e Rawi (2008) afirmam que selecionar as informações que melhor correspondem aos seus interesses facilitam o processamento e a recuperação dessas, sendo de suma importância o aperfeiçoamento das técnicas de agrupamento de dados para fins de pesquisas trazendo aos usuários resultados mais concisos e precisos.

Ao estudar o trabalho de Guelpeli (2012), onde foi proposto o modelo Cassiopeia,que é um agrupador de texto hierárquico, juntamente com um novo método para definição do corte de Luhn(LUHN, 1958) observa-se que não foram feitas avaliações utilizando um sumário humano – SH. Foi desenvolvido ainda um trabalho por *Delgado et al.*, (2013), que usou apenas o domínio jornalístico no idioma português e inglês. Surge então a questão: Com a utilização de

SH em seu pré-processamento, existiria um ganho considerável em relação aos SA nos agrupamentos gerados pelo modelo Cassiopeia nos domínios jurídico e médico e nos idiomas português e inglês? Com base nesta indagação, este trabalho realizou testes com a utilização de SH e SA, a fim de obter resultados dos quais possam ser analisados e comparados, para que se possa garantir a utilização do melhor sumário para o modelo Cassiopeia.

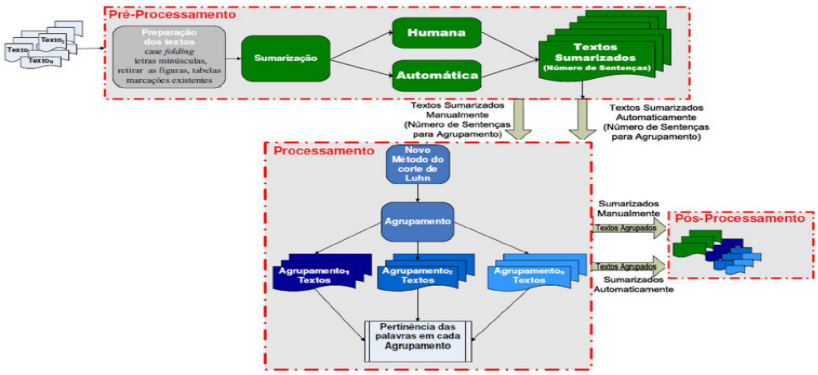

Figura 01. Adaptação Proposta ao Modelo Cassiopéia

2. Metodologia

A metodologia deste trabalho usa a Corpora descrita na seção 2.1. Foram escolhidos os sumarizadores citados na seção 2.2. Para o cálculo do percentual compressão de cada sumário, foi utilizada a aplicação da regra de três a qual será explicada na seção 2.3. O modelo de sumarização e agrupamento desenvolvido será descrito na seção 3.4. Para a análise dos resultados foram escolhidas as métricas externas e internas que são utilizadas na mensuração do processo de agrupamento e comentadas detalhadamente na seção 2.5.

2.1. Corpus

O corpus utilizado é o original no estudo de Guelpeli (2012), sendo todo este disponibilizado pelo próprio Guelpeli para realização deste trabalho. Serão utilizados cem amostras de textos fontes no domínio Médico no idioma português, cem amostras de textos fontes no domínio Médico no idioma inglês, cem amostras de textos fontes no domínio Jurídico no idioma português, cem amostras de sumários humanos no domínio Médico no idioma português, cem amostras de sumários humanos no domínio Médico no idioma inglês, cem amostras de sumários humanos no domínio Jurídico no idioma português. Será realizado também a sumarização dos textos fontes, se ambos os domínios, para gerar cem extratos para cada sumarizador utilizado.

2.2. Sumarizadores

A sumarização será feita para os textos no idioma inglês com o uso de dois sumarizadores profissionais da língua inglesa (*Copernic Summarizer e Intellexer Summarizer Pro*) disponibilizado respectivamente em: http://www.copernic.com/en/ products/summarizer/ e http://summarizer.intellexer.com/order_summarizer_pro.php.Nos textos no idioma português foi utilizado um sumarizador da língua portuguesa (SuPor 2) e o sumarizador profissional da língua inglesa (Intellexer Summarizer *Pro*), mas que também tem a possibilidade de sumarizar no idioma português.

2.3. Cálculo de Compressão

A sumarização será realizada utilizando a compressão por percentual de palavras. Esta será calculada com base na comparação entre número de palavras do texto fonte e seu sumário humano, para os idiomas Português e Inglês nos domínios Médico e para o idioma Português no domínio Jurídico. Após isto, é realizada aplicação a Equação 1, que nos permitirá obter o percentual de compressão individual para cada sumário humano.

$$C = \frac{P_{sh} \times 100}{P_{tf}} \qquad (1)$$

Onde: C é o percentual de compressão, Psh é o número de palavras do sumário humano e Ptf é o número de palavras do texto fonte.

Será realizada então a sumarização automática de cada texto fonte com o mesmo percentual de compressão do sumário humano, garantindo assim a integridade dos testes.

2.4. Modelo de Sumarização e Agrupamento

A sumarização automática foi realizada no corpus em português e inglês de todos os textos fontes com os sumarizadores apresentados na seção 2.2, obtendo assim, sumários automáticos com equivalência no número de palavras em relação aos sumários humanos. Foi feito então os agrupamentos no Cassiopeia, onde são utilizadas as métricas de coesão, acoplamento, *coeficiente silhouette* e *precision, recall, f-measure* que serão descritas na seção 2.5 a seguir. Com estes valores será possível saber se existe um ganho com o uso de sumários humanos no modelo Cassiopeia.

A Figura 02, a seguir, esboça o Modelo de Sumarização e Agrupamento gerados para o domínio médico no idioma português. Nos testes foram realizados o agrupamento do texto-fonte (sem sumarização) e dos textos sumarizados que foram gerados pelos sumarizadores automáticos apresentados na seção 2.2. Foram utilizados, para os testes, cem textos para os textos fontes em cada domínio, cem sumários humanos de cada texto fonte e cem textos de cada sumarizador. Os mesmos foram submetidos duas vezes ao Cassiopeia, uma para geração das métricas Externas (*Recall, Precision e F-Measure*) e outra para métricas Internas (Coesão, Acoplamento e *Coeficiente Silhouette*), permitindo assim, a mensuração dos agrupamentos gerados.

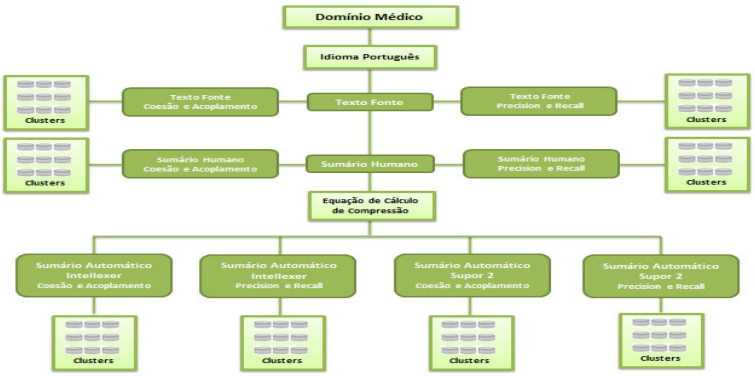

Figura 02. Modelos de Sumarização e Agrupamento gerados para o Domínio Médico no idioma Português.

Os demais modelos de sumarização e agrupamento, para o domínio médico no idioma inglês e para o domínio jurídico no idioma português, foram estruturados de forma análoga.

2.5. Mensuração

A mensuração dos resultados será realizada com a aplicação das métricas internas (coesão, acoplamento e *coeficiente silhouette*) e métricas externas (*recall, precision e Fmeasure*) que são detalhadas em Guelpeli (2012).

3. Resultados Obtidos

Devido ao grande volume das medidas utilizadas para aferir os agrupamentos dos experimentos e consequentemente dos resultados obtidos, serão apresentados os gráficos das medidas harmônicas, na métrica externa que é *F-Measure* e na métrica interna que é o *Coeficiente Silhouette* para cada domínio específico nos ítens 3.1 Domínio Jurídico e 3.2 Domínio Médico, respectivamente.

3.1 Domínio Jurídico

A Figura 3 mostra os resultados dos testes no domínio Jurídico no idioma português, com medida externa F-Measure. Observa-se que o sumário automático Intellexer obteve valores de F-Measure superiores ao sumário humano.

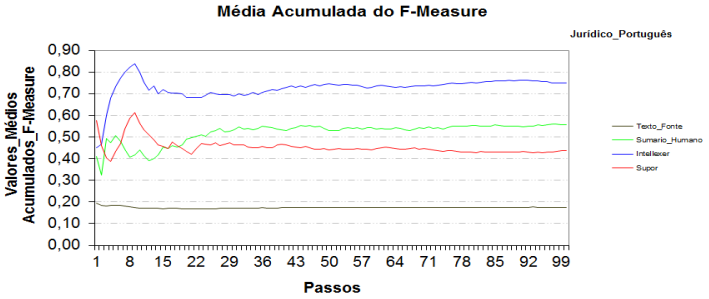

Figura 3. Resultados obtidos pelo modelo Cassiopéia usando a medida externa F-Measure no idioma Português do domínio Jurídico.

A Figura 4 mostra os resultados dos testes no domínio Jurídico no idioma português, com medida interna Coeficiente Silhouette. Observa-se que os sumários automáticos Intellexer e Supor obtiveram valores de Coeficiente Silhouette superiores ao sumário humano.

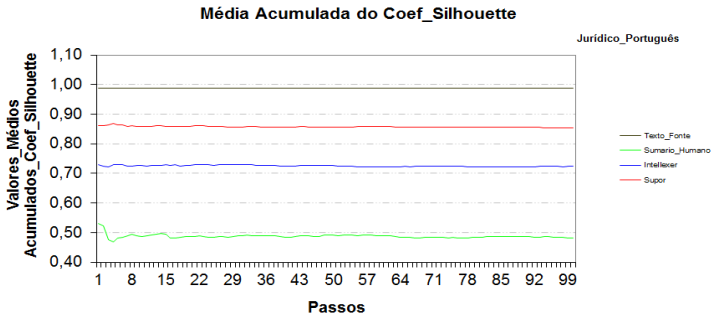

Figura 4. Resultados obtidos pelo modelo Cassiopéia usando a medida interna Coeficiente Silhouette no idioma Português do domínio Jurídico.

3.2 Domínio Médico

A Figura 5 mostra os resultados dos testes no domínio Médico no idioma Português, com medida externa F-Measure. Observa-se que os sumários automáticos Intellexer e Supor obtiveram valores de F-Measure superiores ao sumário humano.

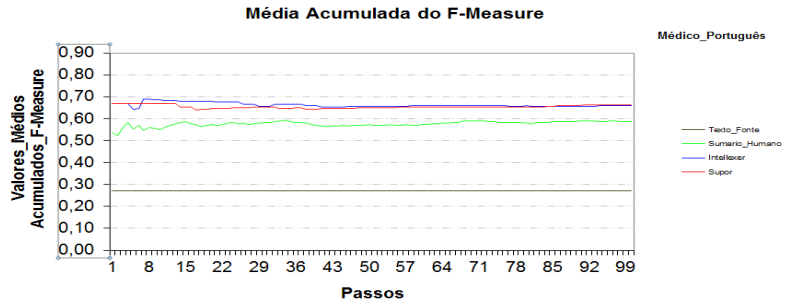

Figura 5. Resultados obtidos pelo modelo Cassiopéia usando a medida externa F-Measure no idioma Português do domínio Médico.

A Figura 6 mostra os resultados dos testes no domínio Médico no idioma português, com medida interna Coeficiente Silhouette. Observa-se que os sumários automáticos Intellexer e Supor obtiveram valores de Coeficiente Silhouette superiores ao sumário humano.

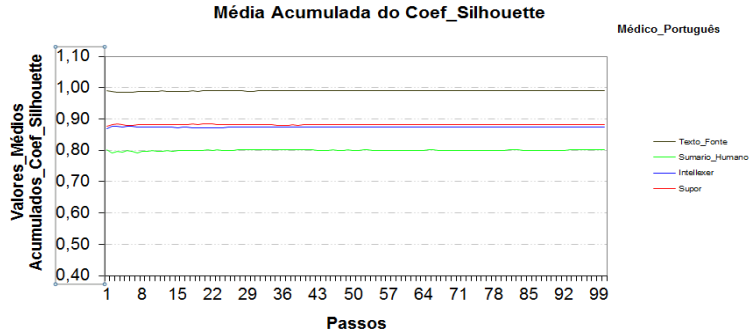

Figura 6. Resultados obtidos pelo modelo Cassiopéia usando a medida interna Coeficiente Silhouette no idioma Português do domínio Médico.

A Figura 7 mostra os resultados dos testes no domínio Médico no idioma Inglês, com medida externa F-Measure. Observa-se que o sumário automático Copernic obteve valores de F-Measure superiores ao sumário humano.

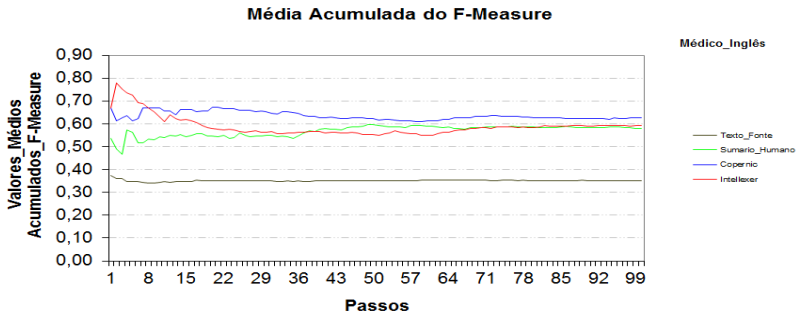

Figura 7. Resultados obtidos pelo modelo Cassiopéia usando a medida externa F-Measure no idioma Inglês do domínio Médico.

A Figura 8 mostra os resultados dos testes no domínio Médico no idioma Inglês, com medida interna Coeficiente Silhouette. Observa-se que o sumário automático Copernic obteve valores de Coeficiente Silhouette superiores ao sumário humano.

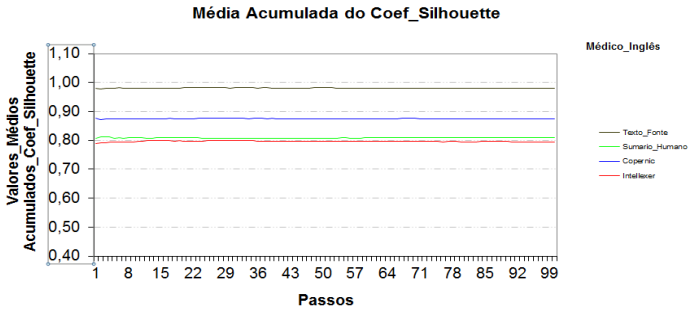

Figura 8. Resultados obtidos pelo modelo Cassiopéia usando a medida interna Coeficiente Silhouette no idioma Inglês do domínio Médico.

4. Considerações Finais

Adotando-se o modelo Cassiopéia, foi proposto uma adaptação na sua etapa de pré-processamento onde foi introduzido sumários humanos e sumários automáticos para análise e comparação dos agrupamentos gerados. Os resultados obtidos foram apresentados através das métricas *F-mesasure* (Figuras 3, 5 e 7) e *Coeficiente Silhouette* (Figuras 4, 6 e 8), onde foram obtidos resultados em que os sumários automáticos têm desempenho superior em 100% dos testes efetuados em relação aos sumários humanos. Observou-se ainda que os resultados no domínio jurídico são representados por linhas mais espaçadas entre si do que as linhas que representam o domínio médico. Isto é justificado devido ao tamanho do corpus utilizado, onde o que representa o domínio médico é mais extenso que o domínio jurídico, além dos domínios em si serem distintos.

O Sumário humano apresenta, para ser feito, a necessidade de alocação de grupos de pessoas, o que toma tempo para análise do texto fonte e ainda implica-se a questão da subjetividade, que varia de pessoa para pessoa. Logo, em questão de desempenho pode-se analisar a praticidade e a velocidade superior ao usar um sumário automático ao invés de um sumário humano, pois esse não necessita de tanto tempo e dinheiro para a sua execução. Após essa análise conclui-se que os sumários automáticos são mais eficientes que os sumários humanos, para o modelo Cassiopeia nas questões de praticidade, economia e qualidade dos agrupamentos gerados.

Referências Bibliográficas

Delgado, C.C.N.; Dias, H. D.; Guelpeli, M.V.C. Utilização de sumários humanos no modelo Cassiopeia Computer on the Beach 2013. Florianópolis - SC

Guelpeli, M.V.C.; Cassiopeia: Um modelo de agrupamento de textos baseado em sumarização. – Tese (doutorado) – Universidade Federal Fluminense. Programa de Pós-graduação em Computação, Niteroi, BR – RJ, Brasil, 2012.

Guelpeli, M.V.C.; e Garcia, A.C.B.; e Bernardini, F.C.; Analysis of Constructed Categories for Textual Classification using Fuzzy Similarity and Agglomerative Hierarchical Methods. “Emergent Web Intelligence” published by Springer Verlag in the series Studies in Computational Intelligence, 2010.

Zoubi, M.B.; e Rawi, M.; An Efficient Approach for Computing Silhouette Coefficients. Journal of Computer Science Volume 4 Page Nº. 252-255, 2008.