UNIVERSIDADE FEDERAL DOS VALES DO JEQUITINHONHA E MUCURI
DEPARTAMENTO DE COMPUTAÇÃO
CURSO DE GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

MODELO CASSIOPEIA: ANÁLISE DO DESEMPENHO NOS IDIOMAS ESPANHOL E ITALIANO NO DOMÍNIO JORNALÍSTICO

Jéssica Lopes Gonçalves da Silva

Diamantina - MG
2014

UNIVERSIDADE FEDERAL DOS VALES DO JEQUITINHONHA E MUCURI
DEPARTAMENTO DE COMPUTAÇÃO
CURSO DE GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

MODELO CASSIOPEIA: ANÁLISE DO DESEMPENHO NOS IDIOMAS ESPANHOL E ITALIANO NO DOMÍNIO JORNALÍSTICO

Autor:
Jéssica Lopes Gonçalves da Silva

Orientador:
Marcus Vinícius Carvalho Guelpeli

Trabalho de Conclusão de Curso apresentada ao Curso de Sistemas de Informação da Universidade Federal dos Vales do Jequitinhonha e Mucuri – UFVJM, como parte dos requisitos exigidos para a obtenção do título de Bacharel em Sistemas de Informação.

Diamantina - MG
2014

Monografia de projeto final de graduação sob o título "Modelo Cassiopeia: Análise do Desempenho nos Idiomas Espanhol e Italiano no Domínio Jornalístico", defendida por Jéssica Lopes Gonçalves Da Silva e aprovada em 23 de julho de 2014, em Diamantina, Minas Gerais.

Banca Examinadora:

Prof. Dr. Marcus Vinicius Guelpeli
Orientador

Profª. MSc. Cláudia Beatriz Berti

Prof. MSc. Athila Rocha Trindade

Prof. MSc. Eduardo Pelli

A minha família por todo apoio que me foi dedicado
em mais esta decisão de minha vida, e em especial à minha mãe Elizabete
que acreditou em mim, mesmo quando eu não acreditava. Te amo mãe!

AGRADECIMENTOS

Esta é uma vitória muito importante para mim. Tenho que agradecer a muitas pessoas que sempre me apoiaram e me inspiraram.

Em primeiro lugar agradeço a Deus, pela força que sempre me deu em todos os momentos de minha vida. Pelas vezes em que acreditava está sozinha e o Senhor me mostrou a sua presença, provando para mim mesma o quanto sou forte e capaz.

Agradeço a minha mãe Elizabete, pelo amor incondicional, pelo carinho, apoio que sempre me concedeu. Pelas conversas relaxantes e inspiradoras. Mãe, você me inspirou e me ajudou a seguir este caminho, graças a você estou chegando ao fim de mais esta batalha. Essa vitória não é minha é nossa.

Agradeço a meu querido avô Eurípedes, você pode não estar presente fisicamente, mas eu sei que você está feliz e orgulhoso de mim. Você foi uma das pessoas que sempre acreditou em mim. Dedico esta vitória a você, meu querido, meu velho, meu amigo.

A minha querida vovozinha Juraci, pelos conselhos que sempre me deu, pela confiança que teve em mim. Obrigada por me ensinar a gostar das coisas simples da vida.

A minha tia Eliane, pelas conversas que tivemos, pelos livros que compartilhamos. Pelos simples momentos que passamos juntas, pelas palhaçadas e bobagens que tivemos. Estes momentos foram muito importantes para mim, foi uma distração em meio à turbulência que me encontrava.

A minha Dinda e a minha tia Flávia, pelos momentos que vivemos juntas, pelos ensinamentos que me passaram. Pelas conversas engraçadas que tivemos. Pelas comidas deliciosas que vocês fizeram e eu simplesmente comi. Agradeço aos meus tios e aos meus primos por tudo que fizeram por mim.

A minha amiga Renata pelas conversas, pelo apoio e por sua amizade. Ao Heider por todos os trabalhos que fizemos juntos até altas horas, mas que tiveram resultados espetaculares foi muito bom te conhecer. Aos outros amigos que sempre me apoiaram. Muito obrigada pela amizade.

A meu orientador Marcus Guelpeli, muito obrigado pela sua ajuda, paciência dedicação e disponibilidade. Você por vezes teve que me explicar e até mesmo desenhar para que eu compreendesse o que estava sendo dito. Obrigada por essa oportunidade.

A UFVJM e aos professores por compartilharem seus conhecimentos e valores. As suas aulas deixarão muita saudade. Agradeço, em especial, aos professores Áthila, Eduardo e Cláudia, que gentilmente aceitaram fazer parte da avaliação deste trabalho. A professora Geruza, por ter me orientado no estágio, e os técnicos Alan e Oscar.

Agradeço o pessoal da PROACE, por terem me acolhido e recebido de braços abertos. Muito obrigada pela oportunidade de aprendizado.

Agradeço ao computador por sempre ter estado presente, apesar de que de vez em quando você falhava. Mas mesmo assim não vivo sem você.

Enfim, muito obrigada a todos.

Aceitar o que não pode mudar revela sabedoria. Confie na vida. Ela sempre sabe o que é melhor para você. Sempre que surgir um pensamento ruim, não lhe dê importância e pense em alguma coisa boa. Tudo o que você dá importância passa a fazer parte de seu interior. Quanto mais otimista você for, mais coisas boas atrairá para sua vida Os desafios só surgem em nosso caminho quanto já temos como enfrentá-los e vencê-los. A vida não joga para perder. Acalme seu coração e confie.

Pelo espírito de Lucius. Psicografia de Zíbia Gasparetto

RESUMO

A clusterização de dados é uma técnica de Mineração de Dados para fazer agrupamentos automáticos de dados segundo seu grau de semelhança. O critério de semelhança faz parte da definição do problema. A proposta deste trabalho é apresentar um estudo sobre o comportamento do modelo Cassiopeia ao agrupar textos jornalísticos nos idiomas espanhol e italiano. Para isso utilizou-se dois *corpus*, um em italiano e um em espanhol. Neste trabalho foram utilizados textos sumarizados com quatro taxas de compressão, 50%, 70%, 80% e 90%. Os resultados das simulações foram validados com os testes estatísticos ANOVA de Friedman e Coeficiente de concordância de Kendall. Ao final deste trabalho têm-se os resultados da eficácia ou não do modelo Cassiopeia ao agrupar textos jornalísticos nos idiomas espanhol e italiano.

Palavras chave: agrupamento, clusterização, sumarização automática, métricas internas, métricas externas.

ABSTRACT

The data clustering is a technique of Data Mining to make automatic data groupings according to their degree of similarity. The criterion of similarity is part of the problem definition. The purpose of this paper is to present a study on the behavior of the cluster model Cassiopeia journalistic texts in Spanish and Italian languages. For this we used two corpu ¬ s, one in Italian and one in Spanish. In this study four summarized text compression ratios, 50%, 70%, 80% and 90% were used. The simulation results were validated with the Friedman ANOVA and Kendall coefficient of concordance. At the end of this work are the results of the effectiveness or otherwise of the Cassiopeia model by grouping journalistic texts in Spanish and Italian languages.

Keywords: Grouping, clustering, summarization, internal metrics, external metrics.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

ANOVA – *ANalysis Of VAriance*

MTPLNAM – Mineração de Textos e Processamento de Linguagem Natural e Aprendizado de Máquina

OTS – Open Text Summarizer

PLN – Processamento de Linguagem Natural

SA – Sumarização Automática

SUMÁRIO

1. INTRODUÇÃO

O desenvolvimento da internet trouxe para a sociedade atual uma grande quantidade de informação e a maior parte dela está na forma textual. A facilidade de acesso e publicação de documentos na rede permitiu a construção de um acervo informacional muito grande. (SHONS, 2007). Entretanto as pessoas não têm tempo para absorver todo esse volume de informação. De um lado, existe uma grande quantidade de informação disponível e de outro, uma dificuldade de recuperar o que seja relevante.

Com o avanço da tecnologia e da alta velocidade de acesso e propagação de dados via internet, tem-se valorizado muito a área de mineração de textos. Isto se deve principalmente ao alto volume de dados sendo enviada a cada momento na rede , sendo assim de suma importância o aperfeiçoamento de técnicas de agrupamento de textos para fins de refinamento de pesquisas com o intuito de trazer ao usuário resultados mais concisos e precisos para suas pesquisas.

O crescimento das informações textuais, oriundos em grande parte pelas facilidades encontradas hoje em gerar e armazenar informações em meios eletrônicos, e a dificuldade posterior de recuperar estas informações, proporcionou o surgimento do que é denominado de sobrecarga de informações (WIVES, 1999).

Uma saída para resolver o problema desta sobrecarga de informação pode ser conseguida com o uso de sumários e agrupamento de textos. Porém, a tarefa é exaustiva se for efetuada de forma manual. É uma tarefa que exige grande esforço intelectual, habilidade, experiência e conhecimento do assunto abordado.

Baseada neste contexto Guelpeli (2012) apresenta o Modelo Cassiopeia como um agrupador de texto hierárquico, que utiliza a sumarização de textos em seu pré-processamento para possibilitar um desempenho relativamente maior que outros modelos descritos na literatura, visando melhorar a precisão na recuperação dos documentos, a coesão e acoplamentos dos grupos de documentos formados, gerar agrupamentos a partir de domínios distintos, ou seja, ser independente do idioma.

O principal propósito deste trabalho é avaliar o comportamento do modelo Cassiopeia ao agrupar textos nos idiomas espanhol e italiano, que foram sumarizados por quatro sumarizadores: *BLMSumm, Copernic Summarizer, Intellexer Summarizer Pro e OTS*. Estes sumarizadores foram escolhidos, porque sumarizam em ambos os idiomas. Nos teste realizados por Guelpeli (2012) no modelo Cassiopeia foram

realizados com textos nos idiomas português e inglês. As simulações foram efetuadas com quatro tamanhos diferentes, isto é, com 10%, 20%, 30% e 50% de tamanho em relação ao texto fonte com textos pertencente ao domínio jornalístico. Para tornar esta avaliação possível foi utilizado um *corpus*[1] no idioma italiano (OLIVEIRA; GUELPELI, 2014) e um *corpus* no idioma espanhol (FERNANDES; GUELPELI, 2014).

Como metodologia foi realizada revisão bibliográfica, sumarização dos textos nos idiomas espanhol e italiano, testes de avaliação de desempenho do Cassiopeia no agrupamento de textos e validação das métricas *Coesão, Acoplamento, Coeficiente Silhouette, Recall, Precision e F-Measure.*

Com este trabalho espera-se poder contribuir, principalmente com a área de recuperação das informações, em virtude dos textos agrupados serem similares entre si. Espera-se, ainda, considerando a proposta deste trabalho, contribuir apresentando uma avaliação de desempenho do comportamento do modelo Cassiopeia ao agrupar textos nos idiomas italiano e espanhol.

1.1. Motivação

Nos dias atuais, com o avanço tecnológico, a área de mineração de textos é muito promissora, porém pouco estudada com relação a outras áreas de tecnologias. Estima-se que nos próximos anos a área de Mineração de Textos será de suma importância, principalmente para a internet devido ao grande volume de informações textuais na rede mundial de computadores.

Examinada a bibliografia levantada, dentro da área de recuperação de informação e na subárea de agrupamento de texto, uma questão que não foi amplamente estudada por Guelpeli (2012), foi a de como será o desempenho do Modelo Cassiopeia ao agrupar textos diferentes dos idiomas proposto em seu trabalho.

Acredita-se que com este trabalho, possa ser realizada uma análise sobre o comportamento do modelo Cassiopeia ao agrupar textos nos idiomas espanhol e italiano no domínio jornalístico, reafirmando assim sua independência quanto ao idioma

[1] *Corpus* é um coleção de textos selecionados e organizados seguindo critérios para servir pesquisa científica.

1.2. Problema:

Não existem resultados ao simular o desempenho do modelo Cassiopeia ao agrupar textos do domínio jornalístico nos idiomas espanhol e italiano.

1.3. Hipótese:

O modelo Cassiopeia obterá desempenho satisfatório ao agrupar textos do domínio jornalístico nos idiomas espanhol e italiano.

1.4. Contribuição:

As contribuições deste trabalho foram as seguintes:

- Análise do desempenho do Modelo Cassiopeia ao agrupar textos sumarizados por dois sumarizadores profissionais – *Copernic Summarizer e Intellexer Summarizer Pro*, um sumarizador da literatura – *BLMSumm* e um *software* livre – *OTS (Open Text Summarizer);*
- Análise e comparação dos resultados do agrupamento de texto utilizando sumários gerados com 50%, 30%, 20% e 10% de tamanho em relação texto fonte;
- Análise e comparação dos resultados do agrupamento variando os idiomas espanhol e italiano;
- Afirmar nos dois idiomas estudados o que apresenta melhor resultado de desempenho nos agrupamentos de texto do modelo Cassiopeia.

1.5. Metodologia da Pesquisa

A metodologia adotada nesta pesquisa compreende: leitura bibliográfica, métodos quantitativos com teste de hipótese em bases públicas no idioma espanhol e italiano, análise do desempenho do modelo Cassiopeia através de métricas internas e externas, visando dar suporte a toda análise realizada no trabalho, baseada na área de mineração de texto. Foram consultados GUELPELI (2012), LOPES (2011), NOGUEIRA (2009), BARION (2008), WIVES (2004).

1.6. Estrutura da Proposta

Capítulo 2 – Fundamentação Teórica

O Capítulo 2 apresenta o estudo da arte do Agrupamento de Textos e Sumarização e os principais conceitos dessa área. Os quais são base de estudo para tal trabalho. Também será apresentado o Modelo Cassiopeia, e suas fases de pré-processamento, processamento e pós-processamento.

Capítulo 3 – Metodologia

Neste Capítulo, será apresentada a metodologia. Serão descritos os métodos de elaboração dos testes, o *corpus* utilizado nos experimentos e suas estatísticas, os sumarizadores que foram utilizados.

Capitulo 4 - Resultados

O Capítulo 4 apresentará os resultados obtidos no experimento, será realizada a comprovação da hipótese, assim como a aplicação dos testes estatísticos e a exibição dos resultados alcançados.

Capítulo 5 - Conclusões

No Capítulo 5, serão discutidas as considerações alcançadas com os resultados obtidos, as limitações e os trabalhos futuros.

2. FUNDAMENTAÇÃO TEÓRICA

Neste Capítulo serão abordados os principais conceitos para a fundamentação deste trabalho. Inicialmente um conceito mais amplo de agrupamento e seus métodos e a descrição de suas propriedades e um conceito mais específico de agrupamento de texto. O capítulo apresenta ainda o conceito de sumarização, sumarização automática e seus elementos de acordo com a literatura. Será apresentado o conceito de *corpus,* além de apresentar o Modelo Cassiopeia e seu funcionamento. Por fim, serão tratados os conceitos e fundamentações das métricas utilizadas neste trabalho para validação da hipótese.

2.1. Mineração de texto

A mineração de é um processo de descoberta de conhecimento, que utiliza técnicas de análise e extração de dados a partir de textos, frases ou apenas palavras. Envolve a aplicação de algoritmos computacionais que processam textos e identificam informações úteis e implícitas, que normalmente não poderiam ser recuperadas utilizando métodos tradicionais de consulta, pois a informação contida nestes textos não pode ser obtida de forma direta, uma vez que, em geral, estão armazenadas em formato não estruturado.

Os benefícios da mineração de textos pode se estender a qualquer domínio que utilize textos (LOH, 2001), sendo que suas principais contribuições estão relacionadas à busca de informações específicas em documentos, à análise qualitativa e quantitativa de grandes volumes de textos, e a melhor compreensão do conteúdo disponível em documentos textuais.

Na prática a mineração de texto define um processo que auxilia na descoberta de conhecimento inovador a partir de documentos textuais, que pode ser utilizado em diversas áreas de conhecimento, como por exemplo, agrupamentos de textos, extração de conhecimento e recuperação de informação.

2.2. Agrupamentos

Segundo Wives (2004) a palavra agrupamento é utilizada na literatura como *clusterização,* tradução do termo *clustering*, e os grupos formados por esse processo são conhecidos como *clusters*. Na literatura, a denominação do processo é ampla: aglomeração, clusterização ou, simplesmente, agrupamento.

O processo de agrupamento é um processo no qual os elementos de uma base de dados são posicionados de tal maneira que formem grupos onde cada elemento tenha a maior similaridade com qualquer outro elemento do mesmo grupo. (BERKHIN, 2002).

Os algoritmos de agrupamento têm as seguintes propriedades: agregação de pontos no espaço (densidade), grau de dispersão nos pontos presentes no agrupamento (variância), raio ou diâmetro (dimensão), disposição dos pontos no espaço (forma), e isolamento dos agrupamentos no espaço (separação). (WIVES, 2004)

A partir dessas propriedades surgem diferentes tipos de agrupamentos, que podem ser hiperesféricos (Figura1-a), alongados (Figura 1-b), curvilíneos (Figura 1-c) ou possuir estruturas mais diferenciadas (Figura 1-d). (BERKHIN, 2002) e (WIVES, 2004).

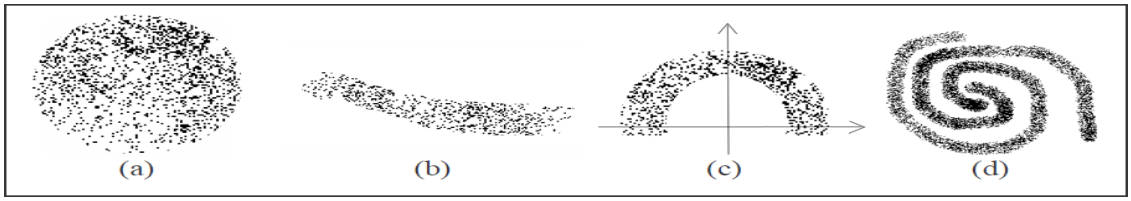

Figura 1. Diferentes tipos de agrupamentos (WIVES, 2004).

2.2.1. Agrupamento de Texto

O agrupamento de texto é um processo no qual se tem uma base de textos com conteúdo similares, cujo objetivo é ter maior conhecimento sobre os textos e suas relações. Assim este processo consiste em reunir uma base de textos de padrões desconhecidos em agrupamentos que possuam significado pertinente.

De acordo com Guelpeli (2012) o problema de agrupamento de texto pode ser definido como: dada uma base de texto T, devem-se agrupar os elementos de T de maneira que os textos mais similares sejam colocados no mesmo grupo, e os menos similares, em grupos distintos. Sendo assim, dado um conjunto com *n* elementos T = {T_1, T_2,..., T_n}, obtém-se um conjunto de *k* agrupamentos G = {G_1, G_2,..., G_k}, cujos elementos de um determinado agrupamento G_i são similares entre si, mas não são similares aos elementos contidos em um conjunto G_j qualquer, onde $i \neq j$. Dessa forma pode-se definir:

$$\bigcup_{i=1}^{k} G = T,\ G_i \neq \varnothing,\ para\ 1 \leq i \leq k\ \ G_i \cap G_j = \varnothing,\ para\ 1 \leq i, j \leq k, i \neq j \quad (1)$$

O agrupamento de texto é composto por três etapas: pré-processamento, processamento e pós-processamento.

Para Nogueira (2009) a etapa de pré-processamento é a parte mais crítica, pois determina a boa qualidade dos agrupadores textuais. Existem, nessa fase, técnicas usadas para reduzir os atributos, como por exemplo, a retirada das *stopword*[2], a radicalização de termos com *stemming*[3]. Esta fase é a que demanda mais tempo em virtude de ser responsável por tratar os textos que serão utilizados em todo o processo.

A etapa de pré- processamento consome cerca de 60% de todo processo de agrupamento, é uma etapa de vital importância para o bom funcionamento das etapas seguintes (PASSOS, 2005).

Na fase processamento, a redução de dimensionalidade tem que ser realizada para viabilizar o processamento dos textos. Entre as diferentes maneiras de efetuar a redução de dimensionalidade, está a proposta de Wives (2004), que seleciona os atributos que identificam os pesos das palavras, ou seja, a importância que cada palavra tem no texto, identificando assim sua "força" de representação na base de textos.

Na fase de pós-processamento obtêm-se os textos agrupados por similaridade e hierarquicamente. É nesta fase também que se valida à qualidade do processo de agrupamento com as métricas externas *Recall*, *Precision e F-Measure* e pela métrica interna Coesão, Acoplamento e Coeficiente *Silhoutte*.

2.3. Métricas para análise do agrupamento de texto

De acordo com Halkidi *et al*. (2001), a avaliação dos agrupamentos pode ser dividida em três grandes classes de métricas: internas ou não supervisionadas; externas ou supervisionadas e relativas.

Para Guelpeli (2012), a métrica relativa tem como objetivo encontrar o melhor conjunto de grupos que um algoritmo de agrupamentos pode definir, a partir de certas suposições e parâmetros. A avaliação de um agrupamento é realizada por comparações entre esse agrupamento, gerados pelo mesmo algoritmo, mas com diferentes parâmetros de entrada. Como a métrica relativa tem a função de avaliar e comparar os agrupamentos gerados pelo próprio algoritmo e não focar na comparação entre o método proposto e outros na literatura, esta métrica não foi a avaliação adotada para analisar este trabalho. Dessa forma, as métricas mais adequadas são as internas e as externas.

[2] *Stopword* são as palavras mais frequentes em um texto, por exemplo, artigos, pronomes e preposições.
[3] *Stemming* é uma técnica que consiste em reduzir palavras relacionadas a uma forma mínima comum, de forma que possam ser combinada sob uma única representação, chamada *stem*.

Nas métricas internas ou não supervisionadas, são utilizadas apenas informações contidas nos grupos gerados para realizar a avaliação dos resultados, ou seja, não são usadas as informações externas. As medidas mais utilizadas, de acordo com Tan *et al.* (2006) e Aranganayagil e Thangavel (2007), para este fim, são Coesão, Acoplamento e Coeficiente de *Silhoutte.*

Para as métricas externas ou supervisionadas, os resultados dos agrupamentos são avaliados por uma estrutura de classes pré-definidas, que refletem a opinião de um especialista humano. Para esse tipo de métrica na opinião de Tan *et. al.* (2006) são utilizadas medidas como: *Precision, Recall,* e como medida harmônica destas duas, *F-Measure*.

2.3.1. Métricas Internas

Coesão(C):

$$\boldsymbol{C} = \frac{\sum_{i>j} Sim(Pi, Pj)}{\frac{n(n-1)}{2}} \tag{2}$$

Onde *Sim (Pi, Pj)* é o cálculo da similaridade entre os textos i e j pertencentes ao agrupamento P, n é o numero de textos no agrupamento *P*, *e Pi e Pj*são membros do agrupamento P. (Guelpeli, 2012).

A *Coesão* representada na Equação 2, mede a similaridade entre os elementos do mesmo agrupamento. Quanto maior a similaridade entre eles, maior a coesão deste agrupamento (KUNZ; BLACK, 1995).

Acoplamento (A):

$$\boldsymbol{A} = \frac{\sum_{i>j} Sim(Ci, Cj)}{\frac{n_a(n_a-1)}{2}} \tag{3}$$

O Acoplamento mede a similaridade média de todos os pares de elementos, sendo que um elemento pertence a um agrupamento e o outro não pertence a esse mesmo agrupamento (KUNZ; BLACK, 1995).

Como é representado na Equação 3, *C* é o centroide de determinado agrupamento, presente em *P*, *Sim (Ci, Cj)* é o cálculo da similaridade do texto i pertencente ao agrupamento P e o texto j não pertence a *P*, *Ci* centroide do agrupamento

P e *Cj* é centroide do agrupamento *Pi* e n_a é o número de agrupamentos presentes em *P*.(Guelpeli, 2012).

Coeficiente Silhouette (S):

$$S = \frac{\left(b(i) - a(i)\right)}{\max\left(\mathrm{a(i)}, \mathrm{b(i)}\right)} \tag{4}$$

O Coeficiente *Silhoutte* representado na Equação 4, baseia-se na ideia de quando um objeto é similar aos demais membros do seu grupo, e de quanto este mesmo objeto é distante de outro grupo. Assim, essa medida combina as medidas de coesão e acoplamento (ARANGANAYAGIL; THANGAVEL, 2007) e (ZOUBI ; RAWI, 2008).

Onde *a (i)* é a distancia média entre o *i-ésimo* elemento do grupo e os outros do mesmo grupo. O *b(i)* é o valor mínimo de distância entre o *i-ésimo* elemento do grupo e qualquer outro grupo, que não contém o elemento, e *max* é a maior distância entre *a (i)* e *b(i),* (Guelpeli, 2012).

O *Coeficiente Silhoutte*de um grupo é a media aritmética dos coeficientes calculados para cada elemento pertencente ao grupo, sendo apresentada na Equação 5 a seguir, onde o valor de *S* situa-se na faixa de 0 a 1. (GUELPELI, 2012).

$$\bar{s} = \frac{1}{N}\sum_{i=1}^{N} S \tag{5}$$

2.3.2. Métricas Externas

Recall (R):

$$R = \frac{n(A)}{n\left(A \cup D\right)} \tag{6}$$

O *Recall* representado na Equação 6 mede a proporção de objetos corretamente alocados a um agrupamento, em relação ao total de objetos da classe associada a este agrupamento (RIJSBERGEN, 1979) e (MANNING *et. al*., 2008).

Onde n(A) é o número de elementos do subconjunto A de acertos e n(D) é o número de elementos do subconjunto D de falsos negativos[4] e n(A ∪ D) é o número total de elementos da classe correspondente (GUELPELI, 2012).

[4] Falsos negativos são elementos que deviam ter sido alocados a um grupo e que foram alocados a outros.

Precision(P):

$$P = \frac{n(A)}{n(A \cup B)} \tag{7}$$

A *Presicion* representada na Equação 7, mede a proporção de objetos corretamente alocados a um agrupamento, em relação ao total de objetos deste agrupamento (RIJSBERGEN, 1979) e (MANNING *et al*., 2008).

Onde n(A) é o número de elementos do subconjunto de A de acertos e n(B) é o número de elementos do subconjunto de B de falsos positivos e n(A ∪ B) é o número total de elementos do grupo. (GUELPELI, 2012).

F-Measure(F):

$$F = 2 * \frac{Precision(P) * Recall(R)}{Precision(P) + Recall(R)} \tag{8}$$

O *F-Measure* representado na Equação 8, é a medida harmônica entre o *Precision* e o *Recall* que, no *F-Measure*, assume valores que estão no intervalo [0,1]. O valor zero indica que nenhum objeto foi agrupado corretamente, o valor um, que todos os objetos estão contidos corretamente agrupados. Assim, um agrupamento ideal deve retornar um valor igual a um (RIJSBERGEN, 1979) e (MANNING *et al*., 2008).

Cada uma das medidas descritas é calculada para cada um dos grupos obtidos fornecendo assim a qualidade de cada grupo. A medida de avaliação, para todo o agrupamento, é obtida através do calculo da média entre cada um das medidas de todos os grupos, apresentada na equação 8. (GUELPELI, 2012).

2.4. Problema da Alta Dimensionalidade

O problema da alta dimensionalidade introduzido na literatura por Richard Bellman (1961) refere-se ao problema causado pelo aumento exponencial no volume de informações, decorrentes da adição de dimensões extras a um espaço matemático, ou seja, divisões e uma região do espaço em células regulares, que crescem exponencialmente com a dimensão do espaço.

O uso elevado de atributos (palavras) gera alta dimensionalidade e que para manter a capacidade de discriminação do atributo é necessário manter a baixa dimensionalidade dos dados. Segundo Guelpeli (2012) na recuperação da informação

este problema pode ser descrito na forma de atributos de um *corpus*, ou seja, a relação entre o número de documentos da coleção, a quantidade de palavras distintas que aparece no total da coleção, e a que aparece em cada documento.

Nogueira (2009) cita algumas técnicas para redução da alta dimensionalidade e dos dados esparsos, porém a técnica mais utilizada na literatura, de acordo Quoniam (2001), Cummins e O'Riordan (2005) é o corte Luhn (LUHN, 1958) que se baseia na Lei de Zipf, conhecida como Princípio do Menor Esforço.

2.4.1. Lei de Zipf

A Lei de Zipf é conhecida como Princípio do Menor Esforço. A Curva de Zipf representada na Figura 2 é uma distribuição estatística e específica utilizada em agrupamento, a qual se encontra em raros fenômenos estocásticos. Um deles é a distribuição da frequência da ocorrência de palavras em um texto, em que nas ordenadas $\mathbf{f}$, se tem um valor dessa frequência, e nas abscissas $\mathbf{r}$, o valor da posição de ordenação relativa dessa palavra, em termos da sua frequência em relação ao das outras palavras do texto. Para a curva de Zipf de uma dada amostra específica, tem se $\mathbf{f}.\mathbf{r} = \mathbf{k}$, em que $\mathbf{k}$ será uma constante específica para essa amostra. Quanto mais próximo do eixo Y mais frequente serão as palavras enquanto que, quanto mais próximo do eixo X, menos frequente serão as palavras, chegando até a frequência de uma ocorrência em todo o texto.

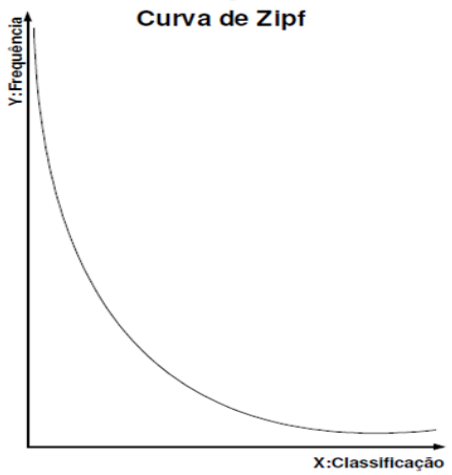

Figura 2. A Curva de Zipf (GUELPEL, 2012)

2.4.2. Corte de Luhn

Luhn (1958) propôs uma técnica para encontrar termos relevantes, assumindo que os mais significativos para discriminação de um conteúdo de um documento estão em um pico imaginário entre dois cortes como mostra a Figura 3.

Foi então proposto o primeiro corte na Curva de Zipf, que tem como finalidade retirar as *stopwords*. As *stopwords* são palavras com mais frequências, que para Pardo (2002) não trazem muita informatividade para o texto, são palavras como pronomes, interjeições e artigos.

Após o primeiro corte a quantidade de palavras é relativamente menor, porém se encontra outro grande problema, as palavras com menos frequência, que são as palavras específicas, encontradas apenas uma única vez nos documentos, as quais fazem com que em uma representação matricial, contribuam para um grande número de dados esparsos, também conhecidos como ruídos. Com isso foi proposto o segundo corte na Curva de Zipf, eliminando estas palavras que apareciam em apenas uma vez em todo o texto.

Com o primeiro e o segundo corte, surge o pico imaginário, um processo heurístico e fonte de estudos de pesquisas atuais (GUELPELI, 2012). Para Quoniam (2001), a Curva de Zipf, com o corte de Luhn (Figura 3), possuí três áreas distintas. Na área I, encontram-se as informações com maior frequência; na área II, as informações interessantes e, na área III os ruídos.

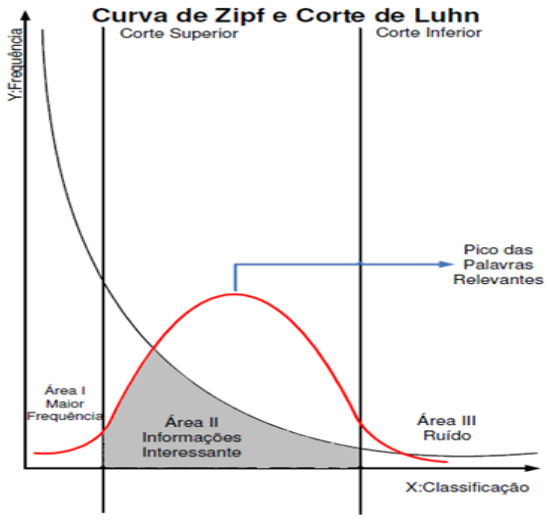

Figura 3. Curva de Zipf com os cortes de Luhn. (GUELPELI, 2012)

2.5. Sumarização

Sumarizar significa reduzir o tamanho de um determinado texto, ou seja, fazer um resumo. Com o avanço tecnológico e com a necessidade de um método mais rápido ao invés de alocar uma pessoa a fim de criar sumários de textos, no final da década de 50, a partir do trabalho de (LUHN, 1958), teve início a sumarização superficial baseada em métodos estatísticos.

Para Barion e Lago (2008) o processo de sumarização seleciona as informações do texto, tornando a descrição mais compacta, mas mantendo a mesma informação. É uma técnica bastante utilizada em mineração de textos com o intuito de identificar palavras ou frases mais importantes dos documentos.

O sumário é composto por palavras-chave e sentenças significativas. A partir da definição das sentenças significativas a estas são atribuídos pesos maiores e, com base nos pesos, é efetuada a seleção para a composição no sumário.

Quando se fala de sumarização, segundo Pardo (2008), é necessário referir-se ao que se entende por um sumário. No campo da sumarização humana, por exemplo, encontram-se vários tipos de sumários: resenhas de notícias jornalísticas, sinopses do movimento da bolsa de valores, sumários de textos novelísticos, extratos de livros científicos, resumos de previsões meteorológicas, dentre outros.

Existem dois tipos de denominação para os sumários, *extract* (extrato) ou *abstract* (resumo). Um extrato é um texto onde as sentenças utilizadas são copiadas e organizadas, de acordo com o texto fonte. O resumo é composto por sentenças reescritas ou rearranjadas, não se limitando a somente à cópia das sentenças do texto fonte.

Para Guelpeli (*et al*, 2010), o grande problema da área de sumarização automática é gerar um sumário automático que não faça com que o texto perca sua informatividade. Outro problema nessa área é a avaliação de um sumário automático que é muito subjetiva, pois depende fundamentalmente de uma avaliação humana. Para aumentar a complexidade da avaliação, existem ainda dois fatores determinantes que são: a quantidade de sumários automáticos gerados e a quantidade de tempo gasto para avaliação dos mesmos.

2.6. *Corpus*

De acordo com Oliveira e Guelpeli (2014) a palavra *corpus* vem do latim “corpo”, “conjunto de textos” e em linguística de *corpus* se refere a uma coleção de textos selecionados e organizados segundo critérios para servir à pesquisa científica.

Para Gandin (2009) um *corpus* representa um conjunto de textos em formato eletrônico que podem ser lidos e manipulados por *software* adequados à pesquisa em linguística.

De um modo geral, *corpus*, na área da Linguística, indica uma coleção de textos reunidos, de áreas variadas ou não, com um propósito específico de análise. Ele difere se, portanto, de uma coletânea (coleção de trechos de obras) ou de uma antologia (uma coleção de textos de autores consagrados), que reúnem obras ou parte de obras dispersas com um intuito didático ou simplesmente comercial. Segundo Bidermann (2001) um *corpus* linguístico informatizado é uma coletânea de textos selecionados segundo critérios linguísticos, codificados de modo padronizado e homogêneo. Essa coletânea pode ser tratada mediante processos informáticos.

2.7. Modelo Cassiopeia

O Modelo Cassiopeia, mostrado na Figura 4 foi proposto para ser um agrupador de texto hierárquico, com novo método para definição do corte de Luhn na curva de Zipf. Seu funcionamento é dividido em três macroetapas: pré-processamento, processamento e pós-processamento.

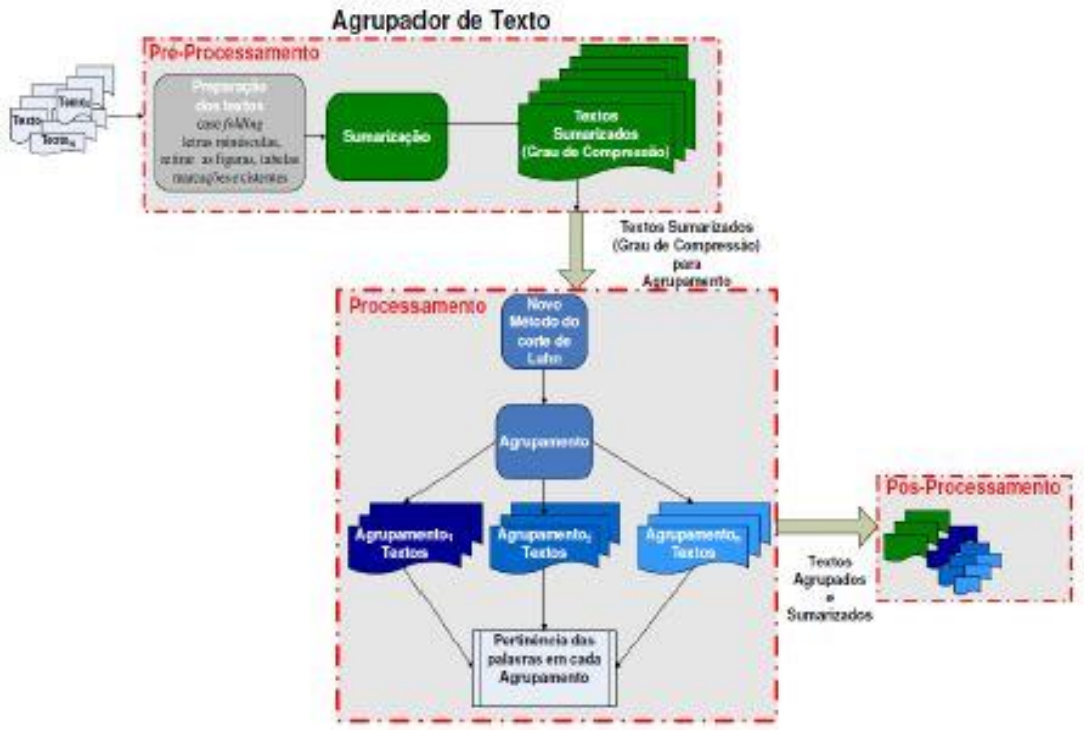

Figura 4. Modelo Cassiopeia (GUELPELI, 2012)

O processo se inicia com a entrada de textos, que são tratados para o processo computacional na etapa de pré-processamento, utilizando a técnica *case folding* (WITTEN *et. al.* 1994) que coloca todas as letras minúsculas, exclui todas as imagens, tabelas e marcações, entre outros cuidados implicando um formato compatível para serem processados.

Ainda nesta etapa é utilizado o processo de sumarização com a finalidade de diminuir o número de palavras, obtendo-se a ideia principal do texto, através da criação de um sumário contendo as palavras mais significativas do texto. Guelpeli (2012) afirma que a sumarização possibilita o uso de um espaço amostral que consegue reduzir o problema da alta dimensionalidade e dos dados esparsos, alem de viabilizar a permanência das *stopwords* tornando o Cassiopeia independente do idioma.

A redução de palavras não relevantes é considerável, precisamente pela definição do grau de compressão do sumário, ou seja, o percentual de sentenças a serem retiradas do texto, sendo este definido pelo usuário dependendo do nível de conhecimento que tenha sobre o assunto e o grau de interesse do mesmo.

A etapa de processamento utiliza o processo de agrupamento de textos hierárquicos e um algoritmo para agrupar os textos com similaridade. O agrupamento hierárquico é usado quando não se conhecem os elementos do domínio disponível,

procurando assim separar, automaticamente, os elementos em agrupamentos por algum critério de similaridade (RIZZI *et al*.,2000) e (LOH, 2001).

Com a utilização da sumarização de textos na etapa de pré-processamento Guelpeli (2012) propôs uma nova abordagem para o método de corte Luhn, onde é inserido um corte médio na distribuição da frequência das palavras, mostrado na Figura 5, o qual consiste em estabelecer a frequência média, do conjunto de palavras baseado na Curva de Zipf, através da frequência relativa, selecionando 25 palavras à esquerda e 25 à direita da média de palavras, os vetores de palavras dos agrupamentos, por questão de dimensionalidade, adotam uma truncagem que, segundo Wives (2004), é de 50 posições, não sendo necessário um valor maior. De acordo com Wives (2004), o uso de um vetor com mais posições não garante palavras com boas características, mas causa aumento do processamento computacional.

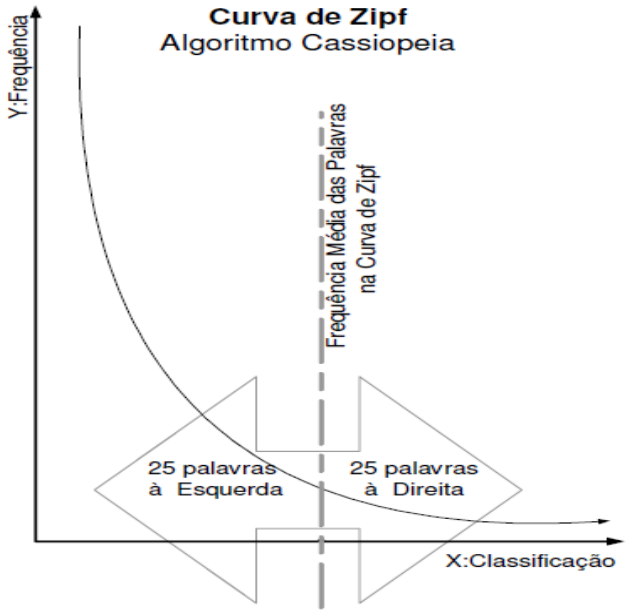

Figura 5. Seleção de atributos modelo Cassiopeia (GUELPELI, 2012)

Por fim na etapa de pós-processamento, cada um dos agrupamentos, terá por similaridade, um conjunto de textos. A organização dos textos de forma hierárquica obtida no pós-processamento é importante para a área da recuperação da informação, pois a estrutura gerada possibilita maior grau de informatividade nos textos agrupados pode atenuar a sobrecarga de informação.

2.8. Testes Estatísticos

Os testes estatísticos têm por objetivo comparar condições experimentais, podendo auxiliar e fornecer respaldo científico aquelas que tenham validade e aceitabilidade no meio cientifico (GUELPELI, 2012). É necessário aferir se existe uma diferença real nas amostras examinadas ou se os resultados são devido ao acaso.

De acordo com Callegari-Jacques (2007) os testes estatísticos podem ser agrupados em paramétricos e não paramétricos. Nos testes paramétricos, os valores da variável estudada devem ter distribuição normal ou aproximação normal. Já os valores não paramétricos, também chamados de distribuição livre, não têm exigências quanto ao conhecimento da distribuição da variável na população.

Neste trabalho a escolha dos testes foi feita de acordo com o Anexo A que traz um diagrama proposto por Callegari-Jacques (2007). O diagrama indica que os testes mais adequados para validar esta pesquisa são o ANOVA de Friedman e o coeficiente de concordância de Kendall. Isto porque os dados aqui examinados são ordinais, as amostras seguem uma distribuição anormal, sendo, portanto, necessário usar testes não paramétricos.

2.8.1. ANOVA de Friedman

O ANOVA de Friedman é um teste não paramétrico que é utilizado para comparar resultados obtidos por três ou mais amostras relacionadas, numa distribuição bivariada (CALLEGARI-JACQUES, 2007). O teste não utiliza os dados da amostra diretamente, mas, sim a ordem ocupada por eles. O ANOVA ordena os resultados da amostra para cada um dos casos e depois calcula a média das ordens pra cada uma das amostras.

2.8.2. Coeficiente de Concordância de Kendall

O teste de coeficiente de concordância de Kendall tem como objetivo normalizar o ANOVA de Friedman. É um teste não paramétrico que gera uma avaliação de concordância, com ranques estabelecidos nos experimentos, e assim, mede a diferença entre a probabilidade das classificações estarem na mesma ordem ou em ordens diferentes. O resultado ocupa um valor no intervalo entre 0 e 1: quanto mais próximo de

1 estiver, maior é a concordância, e quanto menos próximo de 1 estiver, menor é a concordância (CALLEGARI-JACQUES, 2007).

2.9. Trabalhos Correlatos

Existem diversos trabalhos na área de agrupamento de textos e sumarização automática. Porém não foram encontrados trabalhos que analisassem o desempenho de agrupamentos precedido de sumarização, utilizando textos nos idiomas espanhol e italiano.

Wives (2004) utiliza agrupamentos de textos com lógica difusa, e para representação de conhecimento textual utiliza representação de conceitos. Conceitos são estruturas capazes de representar objetos e ideias presentes nos texto. O trabalho de Wives (2004) apresenta uma forte manipulação na fase de pré-processamento, e uma dependência muito forte do domínio no qual está sendo trabalhado, e também nesta fase o autor utiliza uma lista de *stopwords.* No processamento é utilizado uma estrutura hierárquica aglomerativa e o algoritmo Cliques com operadores difusos para o cálculo da similaridade.

No trabalho de Maria *et. al.*(2008) é apresentado uma ferramenta que organiza os agrupamentos e define conceitos, denominada *ClusteringTookit* - CLUTO. Os autores escolhem um *corpus* e é determinado um domínio especifico. Na fase de pré-processamento é empregado à ferramenta FORMA (etiquetagem morfológica e lematização) e duas medidas para atribuir valores aos termos e selecioná-los em limiar de similaridade.

3. METODOLOGIA

Neste Capítulo é descrito a metodologia adotada, a clusterização dos textos no idioma espanhol e italiano, no domínio jornalístico. Serão apresentados os *corpora* utilizados e as estatísticas que a compõem. No capítulo ainda apresentado uma descrição dos sumarizadores automáticos utilizados.

3.1. *Corpus*

O *corpus* utilizado neste trabalho é denominado *Corpus* Italiano (OLIVEIRA; GUELPELI, 2014) e *Corpus* Espanhol (FERNANDES; GUELPELI, 2014), são constituídos de linguagem escrita e possuem textos no domínio jornalístico. Os *corpora* possuem 200 arquivos de texto, sendo 100 do idioma espanhol e 100 do idioma italiano. Na Figura 6 é apresentado o diagrama do *corpus* do domínio jornalístico.

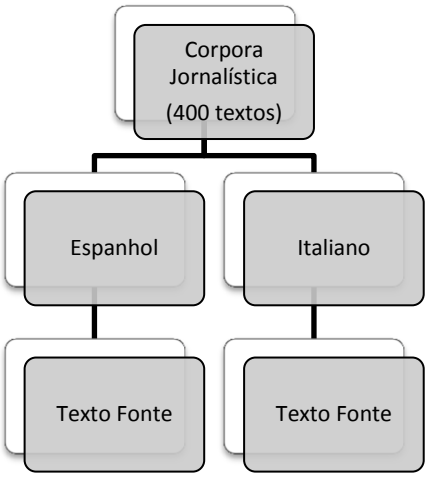

Figura 6. Diagrama dos Corpora usados neste trabalho

3.1.1. *Corpus* Espanhol

O *corpus* espanhol é constituído de 100 textos, que foram extraídos de dois jornais de grande visibilidade global – *El País* da Espanha e pode ser encontrado no *link* http://elpais.com/, e *El Clarín*, da Argentina, disponível em http://www.clarin.com/ (Fernandes e Guelpeli, 2014). A Tabela 1 apresenta as estatísticas dos 100 textos, dividido em 10 categorias, totalizando 69.093 palavras, com media geral de 6.909,3 palavras por categoria.

Tabela 1. Estatísticas dos textos fonte do domínio jornalístico (FERNANDES e GUELPELI, 2014).

Categoria	Palavras	Média de palavras por texto
Ciencias	5902	590,2
Deportes	5459	545,9
Economía	4177	417,7
Entreternimiento	6136	613,6
G-20	7390	739
Inmobiliario	9816	981,6
Negocios verdes	8273	827,3
Politica	6246	624,6
Salud	6087	608,7
Tecnología	9607	960,7
TOTAL	69093	6909,3
Média Geral	6909,3	690,9

3.1.2. *Corpus* Italiano

O *Corpus* Italiano foi composto por 100 notícias coletadas de três jornais de visibilidade global – *EuroNews* e pode ser encontrado no *link* http://www.euronews.com/, *La Reppublica* disponível em http://www.repubblica.it/, *Il CorrieredellaSera*, disponível em http://www.corriere.it/ . O domínio jornalístico é formado por 10 categorias.

Na Tabela 2 são apresentadas as estatísticas dos textos das 10 categorias que formam o domínio jornalístico. Como pode ser observada a categoria que possui os menores textos é a de Economia. No total dos 100 textos, são 41.516 palavras, com média geral de 4.051,6 por categoria.

Tabela 2. Estatísticas dos textos fonte do domínio jornalístico (OLIVEIRA E GUELPELI, 2014).

Arquivos	Palavras	Média de palavras por texto
Economia	931	93
G-20	5515	552
Green Business	6144	614
MercatoImmobiliare	2964	296
Politica	3381	338
Salute	5874	587
Scienza	6076	608
Sport	1434	143
Tecnologia	5115	512
Trattenimento	3082	308
Total	**40516**	**4051**
Média Geral	**4051,6**	**405,1**

3.2. Sumarizadores Automáticos

A sumarização automática foi realizada para os textos do idioma espanhol com o uso de dois sumarizadores profissionais da língua espanhola, o *Copernic Summarizer* e o *Intellexer Summarizer Pro*. Nos textos no idioma italiano foi utilizado o sumarizador profissional *Intellexer Summarizer Pro* que sumariza em qualquer idioma, o *Copernic Summarizer* não sumariza textos especificamente em italiano, mas sumariza textos de origem latina, como francês e espanhol. Além desses, os textos em espanhol e italiano também foram sumarizados com sumarizador da literatura *BLMSumm* e um sumarizador de código aberto *OTS*.

3.2.1. *Copernic Summarizer*

O *Copernic Summarizer* resume arquivos de texto, *emails*, páginas *web* e documentos em PDF. É desenvolvido pela *Copernic* Inc. e está disponível em: http://www.copernic.com/en/products/summarizer/. Permite a geração de resumos e/ou palavras chave de um texto fonte. Neste trabalho foi utilizada a versão *Trial*.

3.2.2. *Intellexer Summarizer Pro*

O *Intellexer Summarizer Pro* é desenvolvido pela *Effective Soft* e permite resumir arquivos de texto, e-mails e páginas web. Sumariza utilizando abordagem estatística e profunda. Para a abordagem estatística resume em qualquer idioma[5], para a profunda, porém, sumariza somente em inglês. Neste trabalho foi utilizada a abordagem estatística. A versão *Trial* foi utilizada neste trabalho e pode ser encontrada no *link*: http://summarizer.intellexer.com/.

3.2.3. *BLMSumm*

O *BLMSumm* é um sumarizador automático proposto por Oliveira e Guelpeli (2011) que tem como característica a independência do idioma e do domínio, justamente por usar o método denominado Cassiopeia proposto por Guelpeli (2012) na escolha de suas palavras.

Para gerar os sumários o *software* permite fazer combinações entre diferentes métodos de classificação de sentenças, tais como *Page Rank, Ts-Isf, Keywords*, Saliência, Grau e diversos algoritmos de meta-heurísticas para geração de sumários como Subida de Encosta, Guloso, Têmpera Simulada, Algoritmo Genético, Busca Tabu e Exame de Partículas.

Os sumários gerados pelo *BLMSumm* neste trabalho foram feitos a partir da combinação do método de classificação de sentenças *PageRank* com o algoritmo Têmpera Simulada[6]. O *PageRank* calcula a frequência das palavras e/ou sentenças encontradas num texto. O algoritmo Têmpera Simulada utiliza um método de sumarização feito a partir de iterações, assim, os sumários são gerados, comparados e classificados a cada iteração do algoritmo (OLIVEIRA; GUELPELI, 2012) e (OLIVEIRA; GUELPELI 2011). A versão utilizada neste trabalho foi disponibiliza por Oliveira (2011).

3.2.4. *Open Text Summarizer – OTS*

O *OTS* é um *software* de código aberto que possibilita ao usuário optar entre a geração de sumários e/ou palavras chaves e permite sumarização em 26 idiomas. Pode ser efetuado o *download* no computador, possuindo versão para *Linux e Windows* ou

[5] Informação obtida através de contato por e-mail com fabricantes do *IntellexerSumamrizer Pro.*
[6] Esta escolha foi feita seguindo sugestão proposta por Marcelo Arantes de Oliveira, um dos autores do BLMSumm. Segundo ele, esta combinação, provavelmente, poderia gerar melhores resultados.

pode ser utilizado através de sua interface na internet. Neste trabalho utilizou-se a versão *online* do sumarizador disponível em <http://www.splitbrain.org/services/ots>.

3.3. *Corpora* gerada com o processo de Sumarização

De acordo com Oliveira (2014) o processo de sumarização automática é dividido em duas etapas: (1) sumarização propriamente dita, (2) limpeza dos sumários através da remoção dos cabeçalhos.

A sumarização automática foi feita com quatro taxas de compressão: 50%, 70%, 80% e 90 %, ou seja, os sumários foram gerados, respectivamente, com 50%, 30%, 20% e 10% do tamanho em relação ao texto original. No total foram realizadas 3.200 sumarizações. Na Figura 7 é apresentado o diagrama dos *corpora* obtido com o processo de sumarização.

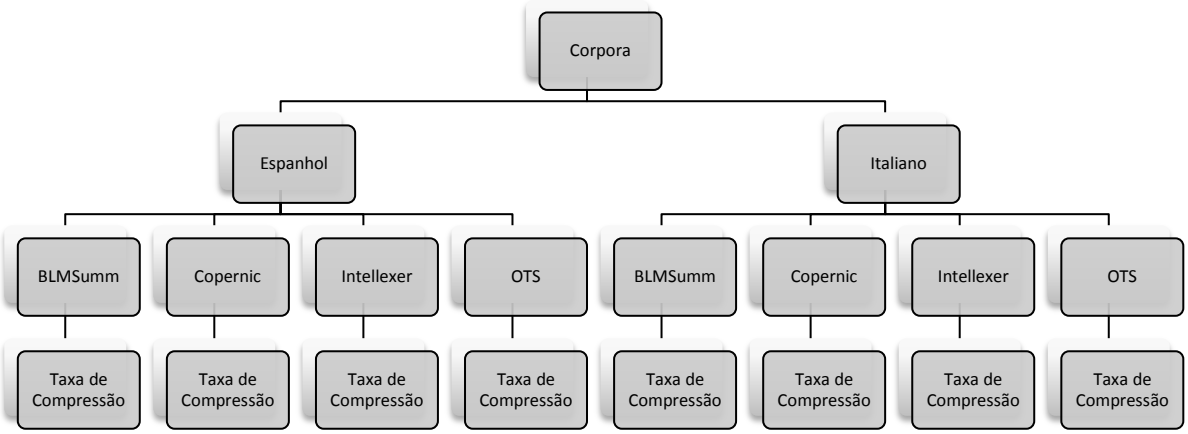

Figura 7. Diagrama dos Corpora gerada com o processo de sumarização

A segunda etapa consistiu na limpeza dos cabeçalhos dos sumários, pois os sumários gerados pelo *Copernic* e pelo *Intellexer* possuem cabeçalho com algumas informações do texto, como por exemplo, as estatísticas gerais e palavras-chave.

No processo de clusterização foram realizados agrupamentos dos textos sumarizados que foram gerados pelos sumarizadores da sessão 3.2. Foram utilizados, para os testes, 1.600 textos sumarizados do idioma espanhol e 1.600 textos sumarizados do idioma italiano. Os mesmos foram duas vezes submetidos ao Cassiopeia, uma para geração das métricas internas (Coesão, Acoplamento e Coeficiente *Silhouette*) e outra para métricas externas (*Recall*, *Precision e F-Measure*), permitindo assim, a mensuração dos agrupamentos gerados.

4. RESULTADOS

Neste Capítulo são apresentados os resultados das simulações realizadas com o *Corpus* italiano e espanhol aplicados no modelo Cassiopeia. Os resultados foram divididos por métricas – internas e externas e subdivididos por idioma – espanhol e italiano e com os graus de compressão usados ao longo deste trabalho 50%, 70%, 80% e 90%. Foram gerados gráficos de Coesão, Acoplamento e Coeficiente *Silhouette*, além de gráficos de *Recall*, *Precision* e *F-Measure*, porém, para organizar a apresentação dos resultados, neste Capítulo, foram apresentados somente os gráficos de gerais de Coeficiente Silhouette (média harmônica) e *F-Measure* (média harmônica) e suas respectivas validações através dos testes estatísticos ANOVA de Friedman e Coeficiente de Concordância de Kendall. Devido a grande quantidade de gráficos, os gráficos de Coeficiente *Silhouette*, Coesão e Acoplamento serão apresentados no Apêndice A e os gráficos de *F-Measure*, *Recall e Precision* no Apêndice B. É apresentada uma comparação entre os resultados obtidos pelas quatro taxas de compressão aplicadas. É mostrada também uma comparação entre os resultados alcançados pelos dois idiomas estudados com objetivo de constatar qual deles apresentou melhores resultados. Por fim, é apresentada a comprovação da hipótese.

4.1. Resultados De Coeficiente *Silhouette*

4.1.1. Idioma Espanhol

4.1.1.1. Comparação dos resultados de Coeficiente *Silhouette*

Na Tabela 3 e na Figura 8 é apresentada a comparação realizada entre as taxas de compressão das médias gerais de Coeficiente *Silhouette* obtidas pelos quatro sumarizadores para os textos do idioma espanhol. O *OTS* obteve os melhores resultados com taxa de compressão de 50% e 90%. O *BLMSumm* obteve os resultados relativamente baixos. Com taxa de compressão de 80%, os melhores resultados foram obtidos pelo *Intellexer*. Quando foi aplicada a taxa de compressão de 70% os valores variaram muito pouco entre si.

Tabela 3. Comparação entre os resultados de Coeficiente *Silhouette* do idioma espanhol

	BLMSumm	**Copernic**	**Intellexer**	**OTS**
50%	0,291855388	0,269163352	0,25297643	0,977366
70%	0,962958284	0,968376644	0,970675016	0,967183
80%	0,890680408	0,941120924	0,95858926	0,87969
90%	0,839017896	0,907650324	0,811425883	0,91293

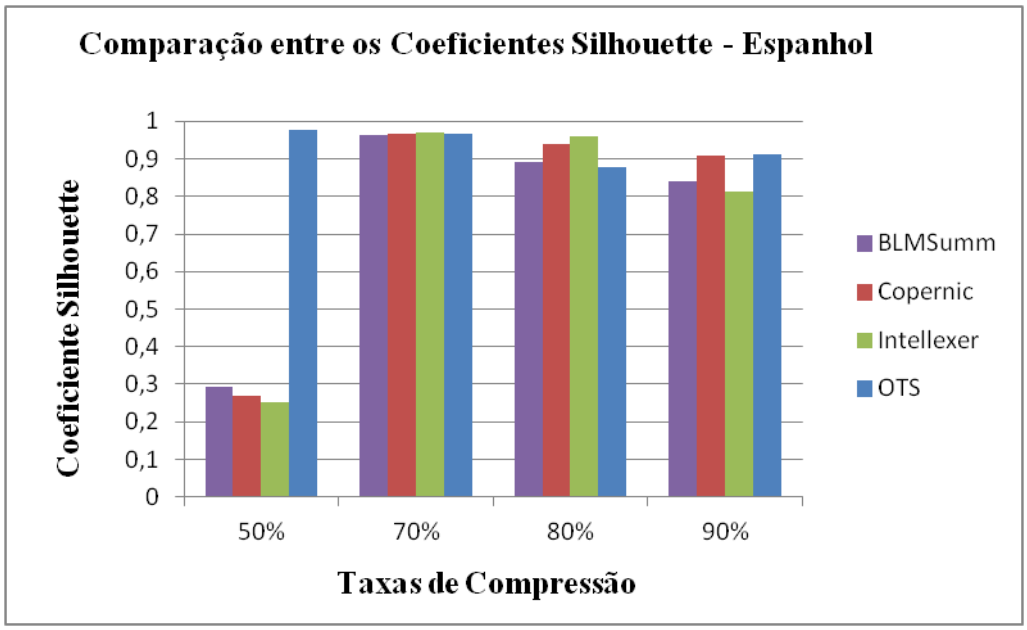

Figura 8. Comparação entre os Coeficientes *Silhouette* – Espanhol

4.1.2. Idioma Italiano

4.1.2.1. Comparação dos Resultados de Coeficiente *Silhouette*

Na Tabela 4 e na Figura 9 são apresentados uma comparação realizada entre as taxas de compressão das médias gerais de Coeficiente *Silhouette* obtidos pelos quatro sumarizadores para os textos do idioma italiano. O *Copernic* obteve os melhores resultados nas taxas de compressão de 50% e 80%. O *OTS* apresentou os melhores resultados com a taxa de compressão de 70% e o *BLMSumm* com taxa de 90%. O *Intellexer* apresentou os valores mais baixos.

Tabela 4. Comparação entre os resultados de Coeficiente *Silhouette* do idioma italiano

	BLMSumm	Copernic	Intellexer	OTS
50%	0,926855877	0,937363574	0,930938	0,936009
70%	0,884043448	0,903600711	0,894798	0,911962
80%	0,858169716	0,876980643	0,852351	0,873996
90%	0,905396922	0,784207061	0,810532	0,748561

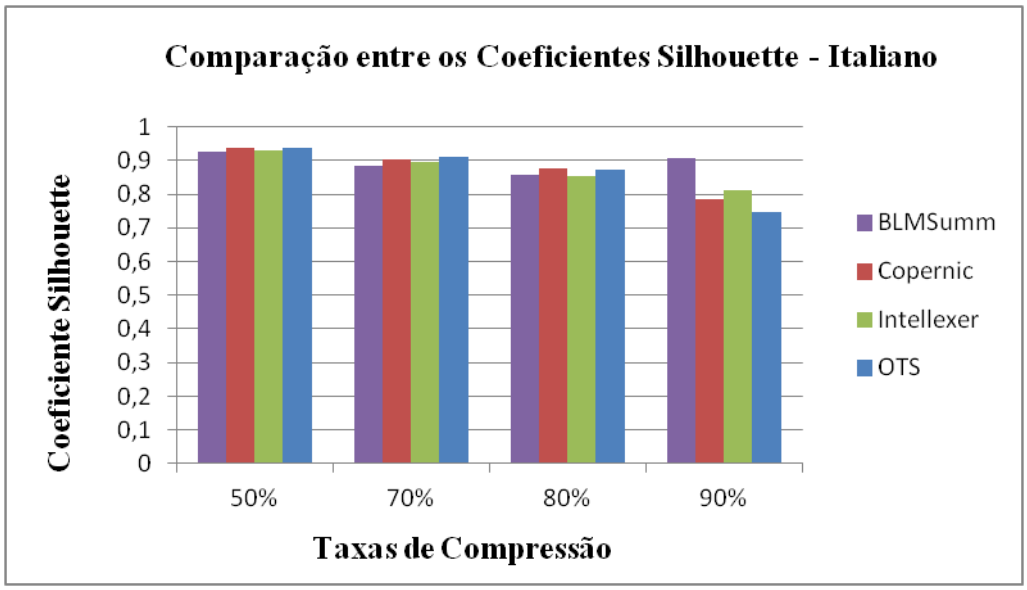

Figura 9. Comparação entre os resultados de Coeficiente *Silhouette*– Italiano

4.2. Resultados de *F-Measure*

4.2.1. Idioma Espanhol

4.2.1.1. Comparação dos resultados de *F-Measure*

Na Tabela 5 e na Figura 10 é apresentada uma comparação entre as quatro taxas de compressão das médias gerais de *F-Measure* alcançadas pelos quatro sumarizadores automáticos para os textos do idioma espanhol. O *BLMSumm* obteve os melhores resultados, exceto quando a taxa de compressão foi de 80%. O melhor resultado com taxa de compressão de 80% foi obtido pelo sumarizador *OTS*. O *Copernic* alcançou os valores mais baixos.

Tabela 5. Comparação entre os resultados de *F-Measure* - Espanhol

	BLMSumm	Copernic	Intellexer	OTS
50%	0,254048087	0,046741475	0,091940402	0,083125
70%	0,183201548	0,046741475	0,091940402	0,055297
80%	0,196613634	0,076971856	0,064782017	0,374934
90%	0,304125069	0,046825092	0,259990939	0,095503

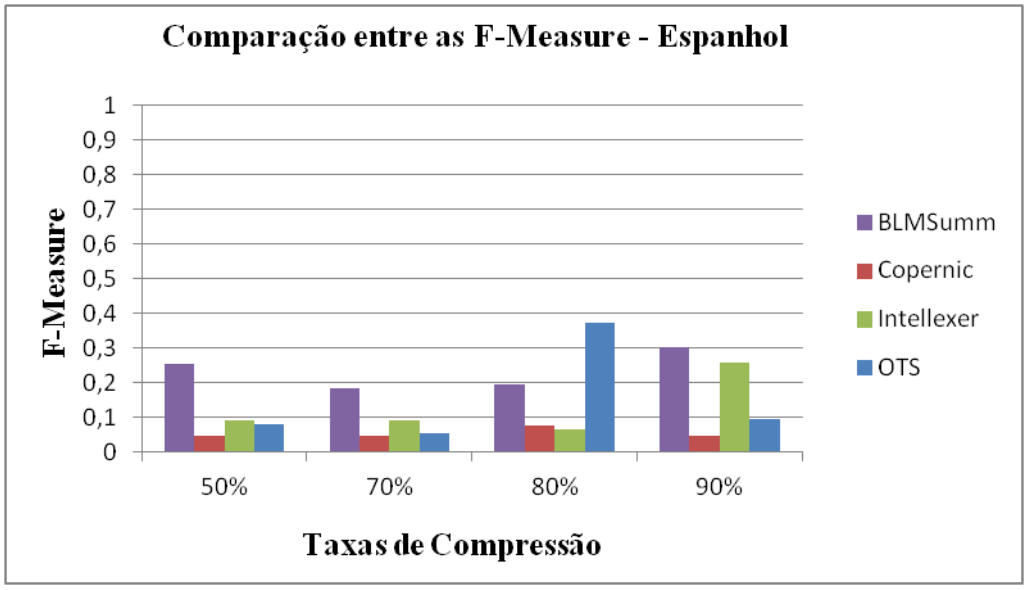

Figura 10. Comparação entre os resultados de *F-Measure*- Espanhol

4.2.2. Idioma Italiano

4.2.2.1. Comparação dos Resultados de *F-Measure*

Na Tabela 6 e na Figura 11 é apresentada uma comparação feita entre as taxas de compressão das médias gerais de *F-Measure* obtidas pelos quatro sumarizadores para os textos do idioma italiano. O *OTS* alcançou os melhores resultados em todas as taxas de compressão. O *BLMSumm*, o *Copernic* e o *Intellexer* apresentaram uma média de valores muito próximas.

Tabela 6. Comparação entre os resultados de *F-Measure* do idioma italiano

	BLMSumm	Copernic	Intellexer	OTS
50%	0,191090773	0,217797126	0,217033	0,228913
70%	0,208790726	0,221039499	0,17837	0,268184
80%	0,203593685	0,212355285	0,208107	0,280404
90%	0,204228583	0,225031589	0,21606	0,500175

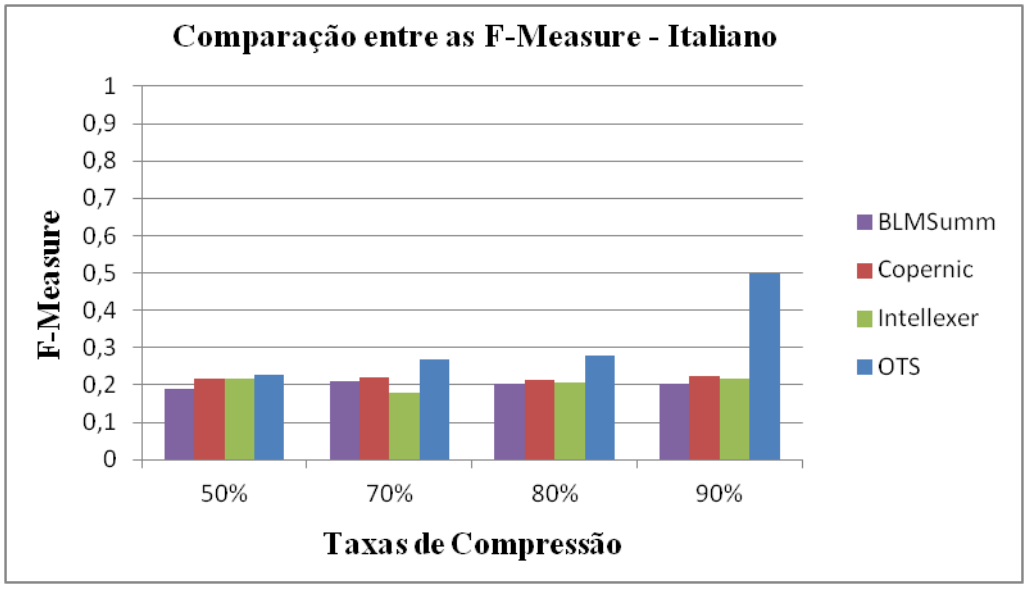

Figura 11. Comparação entre os resultados de *F-Measure* do idioma italiano

4.3. Comparação entre os idiomas espanhol e italiano

4.3.1. Coeficiente *Silhouette*

4.3.1.1. Taxa de Compressão 50%

Para taxa de compressão de 50% o idioma italiano obteve os melhores resultados de Coeficiente *Silhouette* para todos sumarizadores automáticos utilizados, exceto para o *OTS*, no qual o idioma espanhol apresentou melhores resultados, como pode ser observado na Tabela 7, bem como, na Figura 12.

Tabela 7. Comparação de Coeficiente *Silhouette* entre os idiomas espanhol e italiano com taxa de compressão de 50%

	BLMSumm	Copernic	Intellexer	OTS
Espanhol	0,291855388	0,269163352	0,25297643	0,977366
Italiano	0,926855877	0,937363574	0,930938153	0,936009

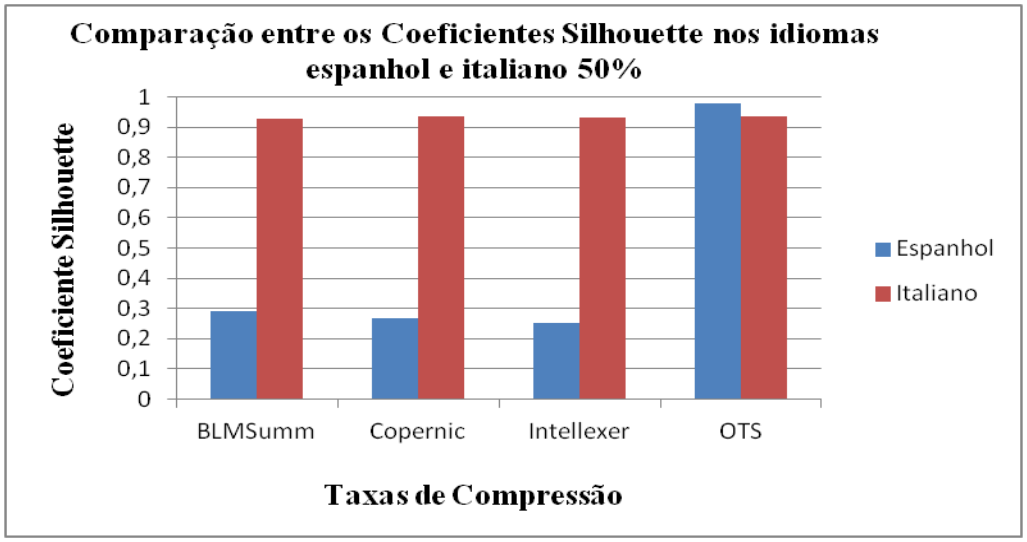

Figura 12. Comparação de Coeficiente *Silhouette* entre os idiomas espanhol e italiano com taxa de compressão de 50%

4.3.1.2. Taxa de Compressão 70%

Na Tabela 8 e na Figura 13 observa-se que quando a taxa de compressão aplicada foi de 70% o idioma espanhol obteve os melhores resultados de Coeficiente *Silhouette* para todos os sumarizadores.

Tabela 8. Comparação de Coeficiente Silhouette entre os idiomas espanhol e italiano com taxa de compressão de 70%

	BLMSumm	Copernic	Intellexer	OTS
Espanhol	0,962958284	0,968376644	0,970675016	0,967183
Italiano	0,884043448	0,903600711	0,894798385	0,911962

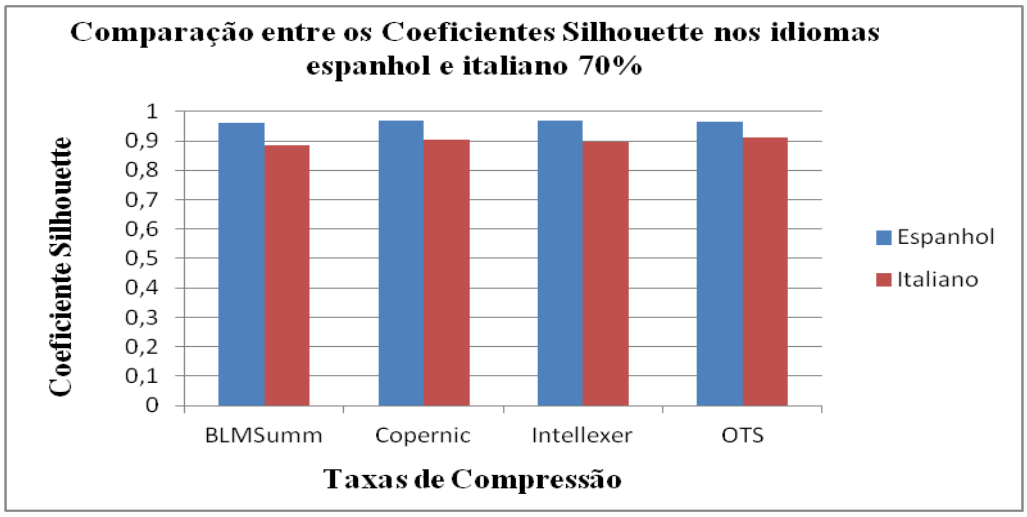

Figura 13. Comparação de Coeficiente *Silhouette* entre os idiomas espanhol e italiano com taxa de compressão de 70%

4.3.1.3. Taxa de Compressão 80%

Na Tabela 9 e na Figura 14 apresentam que o idioma espanhol com taxa de compressão de 80% obteve resultados mais satisfatórios para os quatro sumarizadores utilizados.

Tabela 9. Comparação de Coeficiente *Silhouette* entre os idiomas espanhol e italiano com taxa de compressão de 80%

	BLMSumm	Copernic	Intellexer	OTS
Espanhol	0,890680408	0,941120924	0,95858926	0,87969
Italiano	0,858169716	0,876980643	0,852351125	0,873996

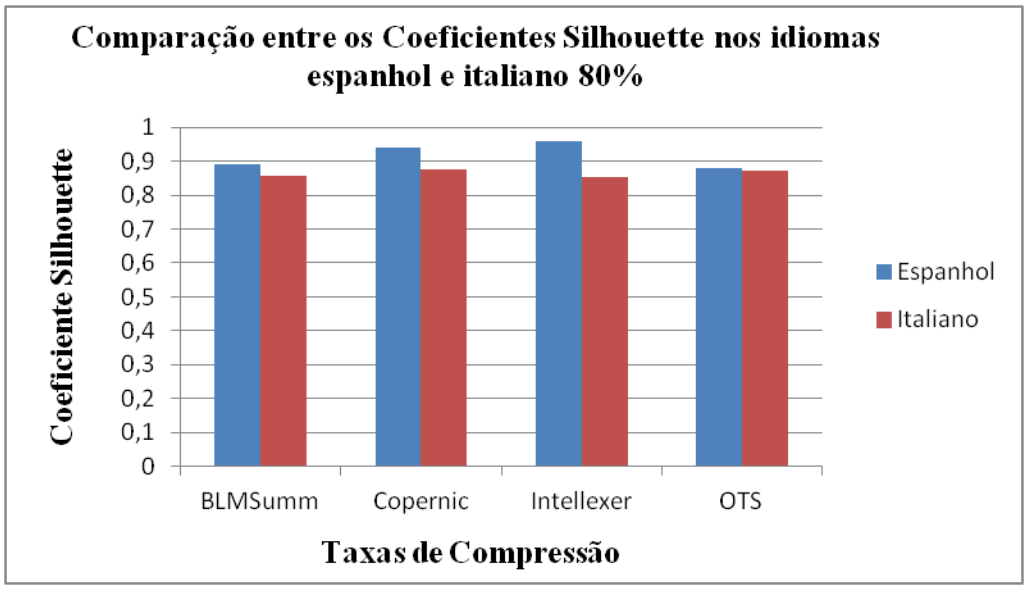

Figura 14. Comparação de Coeficiente *Silhouette* entre os idiomas espanhol e italiano com taxa de compressão de 80%

4.3.1.4. Taxa de Compressão 90%

Na Tabela 10 e na Figura 15 percebe-se que para a taxa de compressão de 90% o idioma espanhol alcançou resultados mais satisfatórios com os sumarizadores *Copernic* e *OTS*. O idioma italiano obteve melhor média harmônica com o *BLMSumm*. E com o *Intellexer* não houve diferença significativa entre os dois idiomas.

Tabela 10. Comparação de Coeficiente *Silhouette* entre os idiomas espanhol e italiano com taxa de compressão de 90%

	BLMSumm	Copernic	Intellexer	OTS
Espanhol	0,839017896	0,907650324	0,811425883	0,91293
Italiano	0,905396922	0,784207061	0,810531776	0,748561

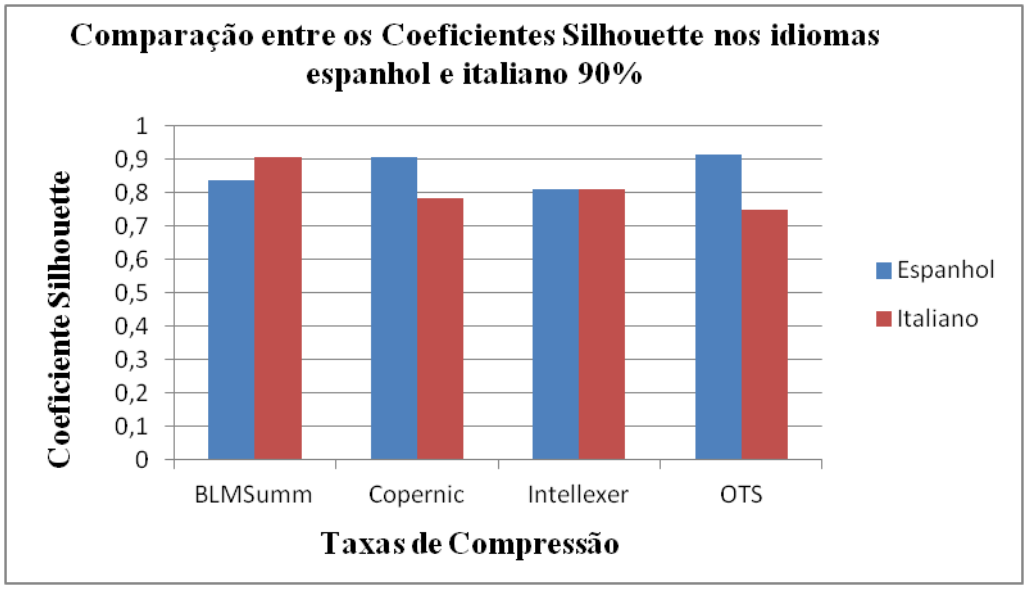

Figura 15. Comparação de Coeficiente *Silhouette* entre os idiomas espanhol e italiano com taxa de compressão de 90%

4.3.2. *F-Measure*

4.3.2.1. Taxa de Compressão 50%

Na Tabela 11 e na Figura 16 observa-se que quando aplicada a taxa de compressão de 50% o idioma italiano obteve os melhores resultados de média *F-Measure* para todos os sumarizadores, exceto para o *BLMSumm*, no qual o idioma espanhol apresentou melhores resultados.

Tabela 11. Comparação de *F-Measure* entre os idiomas espanhol e italiano com taxa de compressão de 50%

	BLMSumm	Copernic	Intellexer	OTS
Espanhol	0,254048087	0,046741475	0,09194	0,083125
Italiano	0,191090773	0,217797126	0,217033	0,228913

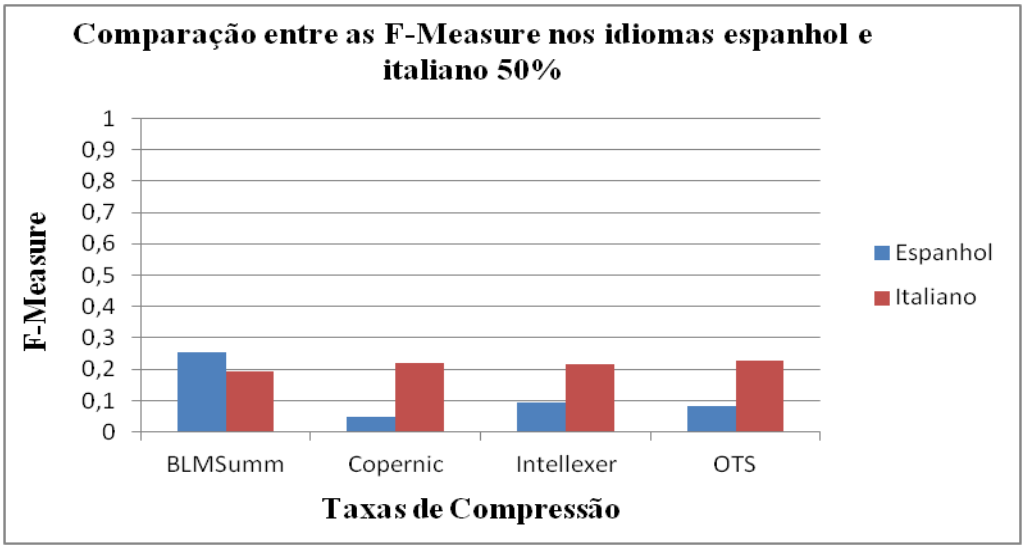

Figura 16. Comparação de *F-Measure* entre os idiomas espanhol e italiano com taxa de compressão de 50%

4.3.2.2. Taxa de Compressão 70%

Para taxa de compressão de 70% o idioma italiano obteve os melhores resultados de *F-Measure* para todos os sumarizadores automáticos utilizados, como se pode observar na Tabela 12, assim como na Figura 17.

Tabela 12. Comparação de *F-Measure* entre os idiomas espanhol e italiano com taxa de compressão de 70%

	BLMSumm	Copernic	Intellexer	OTS
Espanhol	0,183201548	0,046741475	0,09194	0,055297
Italiano	0,208790726	0,221039499	0,17837	0,268184

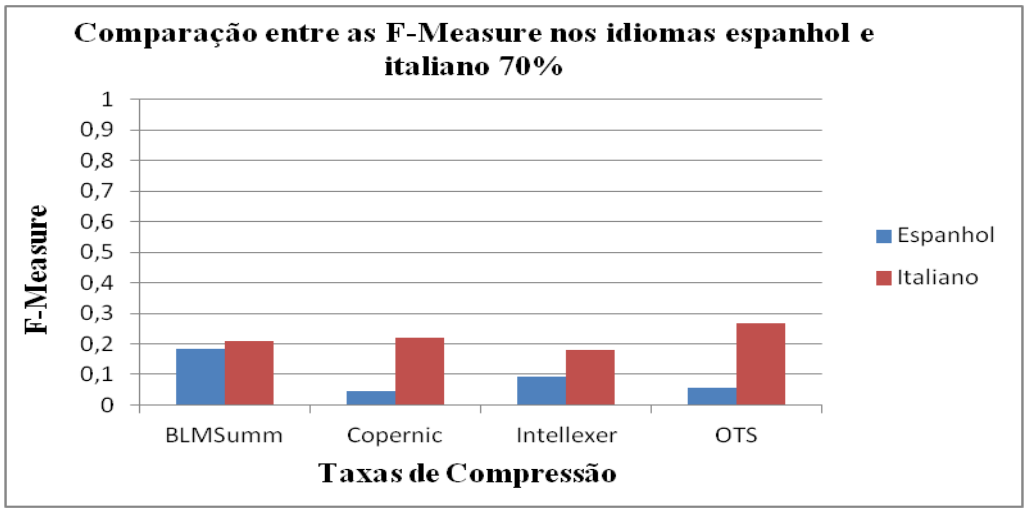

Figura 17. Comparação de *F-Measure* entre os idiomas espanhol e italiano com taxa de compressão de 70%

4.3.2.3. Taxa de Compressão 80%

Na Tabela 13 e na Figura 18 percebe-se que para taxa de compressão de 80% o idioma italiano obteve os melhores resultados com o *BLMSumm*, o *Copernic* e o *Intellexer*. O idioma espanhol alcançou melhor média harmônica com o OTS.

Tabela 13. Comparação de *F-Measure* entre os idiomas espanhol e italiano com taxa de compressão de 80%

	BLMSumm	Copernic	Intellexer	OTS
Espanhol	0,196613634	0,076971856	0,064782	0,374934
Italiano	0,203593685	0,212355285	0,208107	0,280404

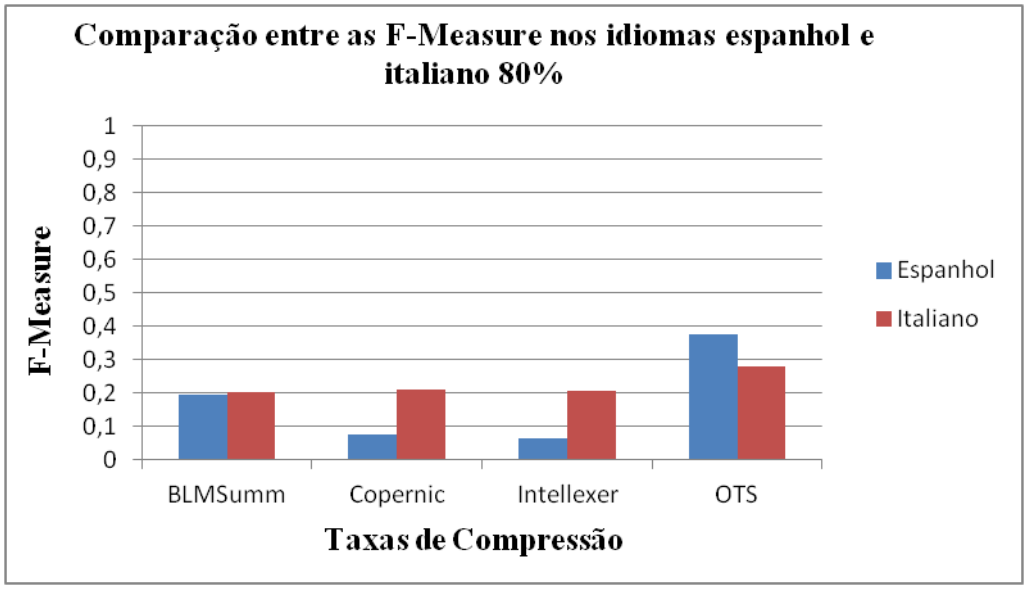

Figura 18. Comparação de *F-Measure* entre os idiomas espanhol e italiano com taxa de compressão de 80%

4.3.2.4. Taxa de Compressão 90%

Na Tabela 14 e na Figura 19 percebe-se que para taxa de compressão de 90% o idioma espanhol obteve resultados mais satisfatórios com o *BLMSumm* e o *Intellexer*. O idioma italiano alcançou melhor *F-Measure* com o *Copernic* e o *OTS*.

Tabela 14. Comparação de *F-Measure* entre os idiomas espanhol e italiano com taxa de compressão de 90%

	BLMSumm	Copernic	Intellexer	OTS
Espanhol	0,253328184	0,081640826	0,210933	0,158677
Italiano	0,204228583	0,225031589	0,21606	0,500175

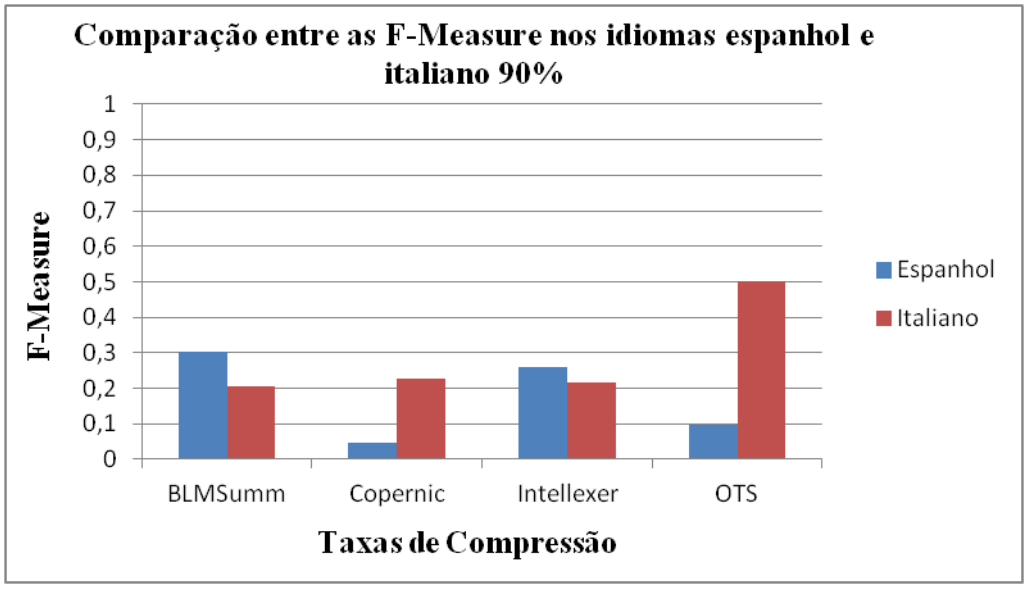

Figura 19. Comparação de *F-Measure* entre os idiomas espanhol e italiano com taxa de compressão de 90%

4.4. Comparações entre os dois domínios e as quatro taxas de compressão

4.4.1. Coeficiente*Silhouette*

A Tabela 15 e a Figura 20 trazem uma comparação entre os resultados de Coeficiente *Silhouette* para os idiomas espanhol e italiano e as quatro taxas de compressão.

Analisando o idioma italiano percebe-se que os textos agrupados que obtiveram resultados mais satisfatórios foram com compressão 50%. Neste idioma percebe-se que os sumarizadores apresentaram uma tendência específica: quanto menor as taxas de compressão melhores foram os resultados apresentados, com exceção do sumarizador *BLMSumm*, que apresentou essa tendência somente com as taxas de compressão de 50%, 70% e 80%.

No idioma espanhol, os textos agrupados que obtiveram os melhores resultados foram com as taxas de compressão de 70%, exceto o *OTS* que teve os melhores resultados com 50% compressão. Neste idioma o sumarizador*OTS* apresentou uma tendência especifica quanto menor a compressão melhor os resultados obtidos. O *Copernic*, o *BLMSumm* e o *Intellexer* apresentaram uma tendência contrária, ou seja, quanto maior a taxa de compressão melhor os resultados, com exceção das taxas de compressão de 80% e 90%.

Quando se compara os dois idiomas espanhol e italiano se conclui que o espanhol possui valores mais altos, salvo para o *BLMSumm*, *Copernic* e *Intellexer* com compressão de 50% e o *BLMSumm* com compressão de 90%.

Tabela 15. Comparação entre os resultados de Coeficiente *Silhouette* dos dois idiomas e das quatro taxas de compressão

	BLMSumm	Copernic	Intellexer	OTS
Espanhol 50%	0,291855388	0,269163352	0,25297643	0,977366
Italiano 50%	0,926855877	0,937363574	0,930938153	0,936009
Espanhol 70%	0,962958284	0,968376644	0,970675016	0,967183
Italiano 70%	0,884043448	0,903600711	0,894798385	0,911962
Espanhol 80%	0,890680408	0,941120924	0,95858926	0,87969
Italiano 80%	0,858169716	0,876980643	0,852351125	0,873996
Espanhol 90%	0,839017896	0,907650324	0,811425883	0,91293
Italiano 90%	0,905396922	0,784207061	0,810531776	0,748561

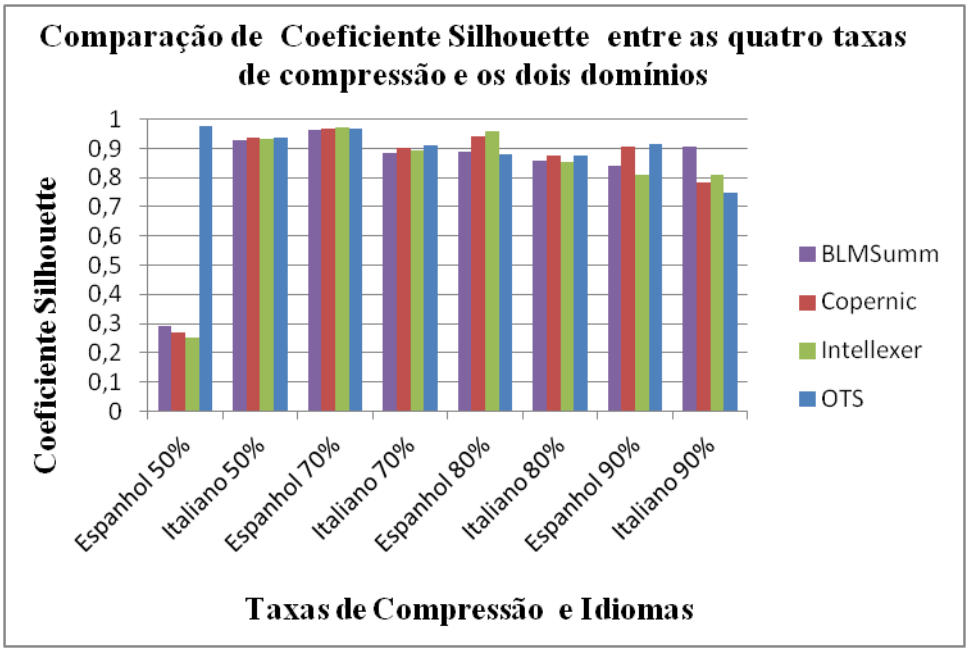

Figura 20. Comparação entre os resultados de Coeficiente *Silhouette* dos dois idiomas e das quatro taxas de compressão

4.4.2. *F-Measure*

A Tabela 16 e a Figura 21 trazem uma comparação entre os resultados *de F-Measure* para os idiomas espanhol e italiano e as quatro taxas de compressão.

No idioma italiano, observa-se que os agrupamentos de textos alcançaram melhores resultados com compressão de 90%, com exceção do *Intellexer* que obteve melhores resultados com taxa de compressão de 50%. Com este idioma, o *OTS* apresentou uma tendência específica: quanto maior a taxa de compressão, mais satisfatórios foram os resultados. Observou-se ainda, que houve pouca variação nos resultados entre as taxas de compressão 50% e 70%.

Analisando o idioma espanhol percebe-se que os agrupamentos de textos alcançaram os resultados mais satisfatórios com grau de compressão de 90%, com exceção do *Copernic* e do *OTS* que obtiveram melhores resultados com compressão de 80%.

Quando se compara os idiomas espanhol e italiano conclui-se que o idioma italiano possui os valores mais altos, com exceção do *OTS* com compressão de 80% e *BLMSumm* e *Intellexer* com 90%.

Tabela 16. Comparação entre os resultados de *F-Measure* dos dois idiomas e das quatro taxas de compressão

	BLMSumm	**Copernic**	**Intellexer**	**OTS**
Espanhol 50%	0,254048087	0,046741475	0,091940402	0,083125
Italiano 50%	0,191090773	0,217797126	0,217032899	0,228913
Espanhol 70%	0,183201548	0,046741475	0,091940402	0,055297
Italiano 70%	0,208790726	0,221039499	0,178370457	0,268184
Espanhol 80%	0,196613634	0,076971856	0,064782017	0,374934
Italiano 80%	0,203593685	0,212355285	0,20810724	0,280404
Espanhol 90%	0,304125069	0,046825092	0,259990939	0,095503
Italiano 90%	0,204228583	0,225031589	0,216060241	0,500175

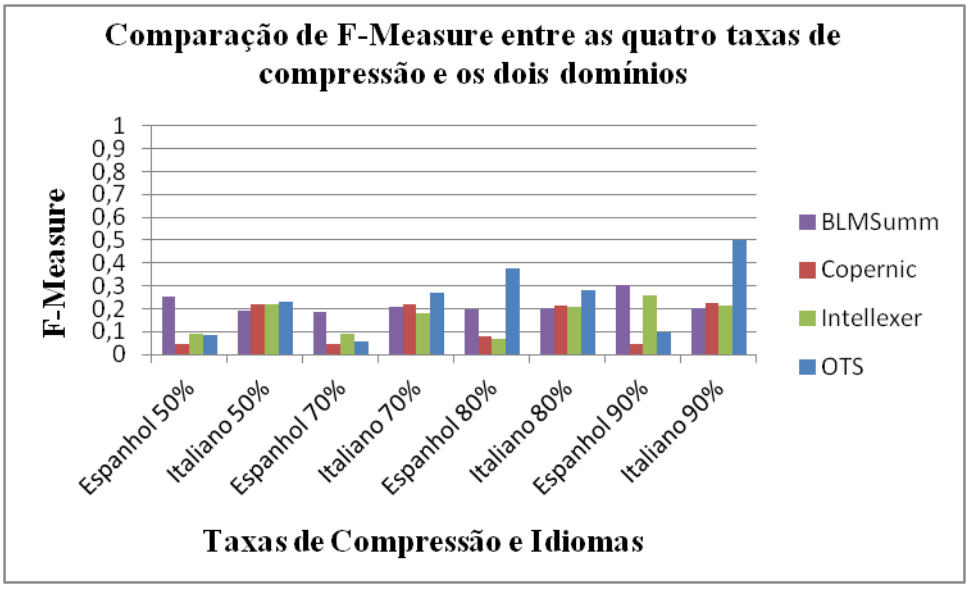

Figura 21. Comparação entre os resultados de *F-Measure* dos dois idiomas e das quatro taxas de compressão

4.5. Comprovação da Hipótese

A hipótese deste trabalho consiste na possibilidade de avaliar o desempenho do modelo Cassiopeia utilizando sumários nos idiomas espanhol e italiano. Formulando em um teste de hipóteses, tem-se que a hipótese nula consiste que o modelo Cassiopeia não obteve um bom desempenho ao agrupar textos nos idiomas espanhol e italiano. A equação 9 traz a representação da hipótese nula.

$$H_0 = K_{Agrupadores\_domínio\_idioma\_restrito} = K_{Cassiopeia\_domínio\_idioma\_restrito} \quad (9)$$

Onde:

H_0= hipótese nula;

$K_{Agrupadores\_domínio\_idioma\_restrito}$ = K amostras de textos agrupados que são sensíveis ao idioma em outros agrupadores.

$K_{Cassiopeia\_domínio\_idioma\_restrito}$ = K amostra de textos agrupados no modelo Cassiopeia.·.

Quando a hipótese nula for rejeitada, outra hipótese, a alternativa H_1 deve ser aceita, ou seja, o modelo Cassiopeia alcança um bom desempenho nos agrupamento dos textos, independente do idioma.

Representação da hipótese alternativa através da Equação 10:

$$H_1 = K_{Cassiopeia\_domínio\_idioma\_restrito} > K_{Agrupadores\_domínio\_idioma\_restrito} \quad (10)$$

A hipótese alternativa foi baseada nas amostras obtidas com os testes de agrupamento utilizando o modelo Cassiopeia nos textos sumarizados, usando as métricas Coesão, Acoplamento, Coeficiente *Silhouette, Recall, Precisio*n e *F-Measure*.

Para comprovação da hipótese foi utilizado o teste estatístico ANOVA de Friedman, que considera que as diversas amostras são estatisticamente idênticas, em sua distribuição (hipótese nula ou H_0). A hipótese alternativa (H_1) aponta como elas são significativamente diferentes na sua distribuição e o teste de concordância de Kendall normaliza o teste estatístico de Friedman, com a finalidade de gerar uma avaliação de concordância, ou não, com ranques estabelecidos (GUELPELI, 2012).

4.6. Análise dos testes estatísticos

Para análise dos testes estatísticos foram geradas as tabelas apresentadas no Apêndice C. Nelas estão contidos os valores gerados para o teste ANOVA de Friedman e para o Coeficiente de Concordância de Kendall, obtidos com o *software* citado e explicado no mesmo apêndice. São dezesseis tabelas, oito para os testes do idioma espanhol com as métricas internas e externas, e oito para o idioma italiano com as métricas internas e externas.

As tabelas são compostas de informações como: N (número de amostras utilizadas); Graus de liberdade, qui-quadrado, p-nivel, os valores de ordem médio são comparadoscom os valores do Coeficiente de Concordância de Kendall, os valores de soma das ordens, ordem médio e média são usados para o ANOVA de Friedman.

Nos teste estatísticos, em todas as tabelas contidas no apêndice C, observou-se a rejeição da hipótese nula (H_0) e a aceitação da hipótese alternativa (H_1). Em virtude de todas as tabelas apresentarem valores de Ordem Médio e Coeficiente de concordância de Kendall próximos de 1, como exemplo é apresentado a Tabela 17.

Tabela 17. Teste Estatístico da métrica *F-Measure* – Idioma Espanhol

Comparando amostras múltiplas relacionadas			
N	30	*Graus de liberdade*	3
qui-quadrado	83,56	*p-nível*	0
Coef. de concordância de Kendall	0,9284	*Ordem médio*	0,926
	Ordem médio	**Soma de ordens**	**Média**
Copernic	1,7667	53	0,077
Intellexer Summarizer Pro	1,2333	37	0,0648
BLMSumm	3	90	0,1966
OTS	4	120	0,3749

5. CONCLUSÕES

A internet se consolida a cada dia como o principal meio de distribuição e armazenamento de informações, por este motivo o estudo de informações textuais tem demonstrado ser um importante foco de pesquisa, contribuindo muito com a área de Mineração de Texto e incentivando, cada vez mais, o aprimoramento de técnicas que possibilitem uma manipulação e recuperação destas informações, sendo este um importante diferencial para as pessoas e para as empresas.

O modelo Cassiopeia consiste no uso da sumarização automática no pré-processamento para redução da quantidade de atributos, podendo manter as *stopwords*, o que torna o modelo independente do idioma e ainda apresenta uma nova proposta para o corte de Luhn que o torna também independente do domínio. Este modelo apresenta um grande avanço em termos de precisão, recuperação da informação, coesão e acoplamento.

O principal objetivo deste trabalho foi apresentar uma avaliação do comportamento do modelo Cassiopeia ao agrupar textos nos idiomas espanhol e italiano. Foram avaliados textos sumarizados com os sumarizadores *BLMSumm*, *Copernic*, *Intellexer e OTS.*

Nos experimentos foram gerados *clusters* contentos textos sumarizados com as seguintes taxas de compressão 50%, 70%, 80% e 90%. Os resultados foram validados com os testes estatísticos ANOVA de Friedman e Coeficiente de Concordância de Kendall.

O Cassiopeia obteve um desempenho satisfatório ao agrupar textos nos idiomas espanhol e italiano no domínio jornalístico. Ao se observar as quatro taxa de compressão aplicada nos textos percebeu-se que os que os textos sumarizados obtiveram de maneira geral os resultados mais satisfatórios nos agrupamentos foram os textos com as taxas de compressão de 70% e 80%.

Observou-se também, que dentre os sumarizadores utilizados no pré-processamento, o que apresentou melhor desempenho foi o sumarizador *OTS*, seguido pelo sumarizador *BLMSumm*.

Analisando os dois idiomas estudados espanhol e italiano, percebeu-se que na maioria dos resultados apresentados os textos do idioma italiano apresentaram desempenho superior aos textos do idioma espanhol, isso pode ser explicado devido aos textos no idioma italiano serem menores e, portanto apresentam as informações

relevantes do texto de maneira direta, ao contrário dos textos no idioma espanhol que por serem maiores trazem outras informações, além das relevantes, prejudicando a informatividade do texto.

5.1. Limitações

Neste trabalho foram utilizados nas simulações apenas textos do domínio jornalístico, os *corpora* utilizados possuíam dois idiomas com apenas 100 textos cada. Acredita-se que este seja um fator limitante, considerando que os textos poderiam ser em número maior, abranger mais domínios e ter uma maior diversidade de idiomas.

5.2. Trabalhos Futuros

As simulações com o *BLMSumm* foram feitas, conforme descrito na seção 3.2.3., a partir da combinação do método de classificação de sentenças *PageRank* com o algoritmo Têmpera Simulada. Como este sumarizador obteve resultados muito satisfatórios nesse trabalho, sugere-se, para trabalhos futuros, variar tanto os métodos de classificação de sentenças quanto o algoritmo e observar se os resultados se mantêm ou variam. Outra sugestão é realizar uma avaliação mais ampla que envolva outros domínios e outros idiomas, sejam eles de origem latina ou não.

REFERÊNCIAS

ALMEIDA, L. G. P. **Análise de agrupamento para base de dados.** Petrópolis, RJ. : Laboratório Nacional de Computação Científica, 2007. xxii, 139 p.: il.: 29 cm. Dissertação (M. Sc.) – Laboratório Nacional de Computação Científica, 2007.

ARANHA, C. N. **Uma Abordagem de Pré-Processamento Automático para Mineraçãode Textos em Português: Sob o Enfoque da Inteligência Computacional**. Tese deDoutorado. PUC-Rio de Janeiro, Brasil, 2007.

ARANGANAYAGIL, S.; THANGAVEL, K. **Clustering Categorial Data Using Silhouette Coefficient as a Relocating Measure.** In International conference on computational Intelligence and multimedia Applications, ICCIMA, 2007, Sivakasi, India. Proceedings. Los Alamitos: IEEE 2007. p. 13-17.

BERKHIN, P. **Survey of Clustering Data Mining Techniques.** Accrue Software, San Jose, CA, 2002.

CALLEGARI-JACQUES, S. M. **Bioestatística: Princípios e Aplicações**. Porto Alegre: Artmed, p. 264, 2007.

Copernic Summarizer (Versão *Trial*) [Programa de Computador]. N. Harris Computer Systems. Disponível em: <http://www.copernic.com/en/products/summarizer/download.html>. Acesso em 12 ago. 2013.

CUMMINS, R., O'RIORDAN, C. E**volving General Term-Weighting Schemes for Information Retrieval: Tests on Larger Collections.** Journal Artificial Intelligence Review http://dl.acm.org/citation.cfm?id=1107370archive Volume 24 Issue 3-4, November 2005 Kluwer Academic Publishers Norwell, MA, USA, 2005.

DELGADO, C. C. N.; DIAS, H. D. **Utilização de Sumários Humanos no Modelo Cassiopeia** – Trabalho de Conclusão de Curso–Centro Universitário de Barra Mansa. Graduação em Engenharia da Computação, Barra Mansa, Brasil, 2012.

FERNANDES, H. M.; GUELPELI, M. V. C. Creación de corpus enlenguaespañola para suutilizaciónen testes acerca de Sumarización Automática. In: **6th InternationalConferenceon Corpus Linguistics (CILC2014**). Las Palmas de Gran Canaria, Espanha, 2014.

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.; BERNARDINI, F. C. ; GARCIA, A. C. B**. Todas as Palavras da Sentença como Métrica para um Sumarizador Automático.** In: Tecnologia da Informação e da Linguagem Humana-TIL, Web Media, 2008. p. 287-291,Vila Velha,Brasil, 2008.

GOLDSCHMIDT, R., PASSOS, E. **Data Mining: Um Guia Prático**. Livro Editora Campus - Rio de Janeiro: Elsevier, 2005.

HALKIDI, M.; BATISTAKIS Y.; VARZIRGIANNIS, M. **On clustering validation techniques**. Journal of Intelligent Information Systems, 17(2-3):107-145, 2001.

HOWLAND, P.; PARK, H. C**luster-Preserving Dimension Reduction Methods for Document Classification.** Book survey of text mining: clustering, classification, and retrieval Second . Editors BERRY, M. E CASTELLANO, M. Edition, Springer, Part I Clustering, pp 3- 24, 2007.

Intellexer Summarizer Pro. (Versão 3.1. *Trial*) [Programa de Computador]. Soft 112. Disponível em: <http://intellexer-summarizer-pro.soft112.com/>. Acesso em 12 ago. 2013.

KUNZ, T.; BLACK, J.P.: **Using Automatic Process Clustering for Design Recovery and Distributed Debugging.** IEEE Trans. Software Eng.515 527,1995.

LEVY, D. M. **To grow in wisdom: vannevar bush, information overload, and the life of leisure.** In JCDL (2005) p.281-286, 2005.

LOH, S. **Abordagem Baseada em Conceitos para Descoberta de Conhecimento em Textos** Universidade Federal do Rio Grande do Sul-Instituto de Informática-Curso de Pós-graduação em Ciência da Computação.Tese de Doutorado- UFRGS, 2001.

LOPES, M. C. S. Mineração de **Dados Textuais Utilizando Técnicas de Clustering para o Idioma Português** [Rio de Janeiro] 2004 XI, 180 p. 29,7 cm (COPPE/UFRJ, D. SC., Engenharia Civil, 2004) Tese - Universidade Federal do Rio de Janeiro. COPPE.

LOPES, G. A. W. **Um Modelo de Rede Complexa para Análise de Informações Textuais.** Dissertação apresentada ao Curso de Mestrado em Inteligência Artificial Aplicada à Automação Industrial do Centro Universitário da FEI, São Paulo, 2011.

LUHN, H. P. The Automatic Creation of Literature Abstracts. **IBM Journal of Research and Development**, vol. 2, 157-165, 1958.

MANNING, C. D.; RAGHAVAN, P.; SCHUTZE, H. **Introduction to Information Retrieval**, Cambridge University Press. 2008.

MARTINS, C. A. **Uma Abordagem para Pré-processamento de Dados Textuais em Algoritmos de Aprendizado**. Tese de Doutorado, Instituto de Matemática e Computação, Universidade de São Paulo, Brasil, 2003. http://www.teses.usp.br/teses/disponiveis/55/55134/tde-08032004-164855/.

NOGUEIRA, B. M. **Seleção não-supervisionada de atributos para Mineração de Textos**. 2009. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, São Paulo, Brasil, 2009.

OLIVEIRA, H. M. **Seleção de entes complexos usando lógica difusa.** Dissertação (Mestrado em Ciência da Computação) – Instituto de Informática, PUC-RS, Porto Alegre, 1996.

OLIVEIRA, M. A.; GUELPELI, M. V. C. *BLMSumm* – Métodos de Busca Local e Metaheurísticas na Sumarização de Textos. **Proceedings of the ENIA - VIII Encontro Nacional de Inteligência Artificial 2011**, p. 287 - 298, Natal, Brasil, 2011. Disponível em: <http://nlx.di.fc.ul.pt/~guelpeli/Arquivos/Artigo19.pdf>. Acesso em 18 jun. 2013.

OLIVEIRA, M. A.; GUELPELI, M. V. C. The Performance of BLMSumm: Distinct Languages with Antagonistic Domains and Varied Compressions. In: **The Second International Conference on Information Science and Technology (ICIST 2012**), p 609 – 614. Wuhan, China, 2012.

OLIVEIRA, R.R.; GUELPELI, M.V.C. Building a Corpus in Italian Written Language. In: **6th International Conference on Corpus Linguistics (CILC2014**). Las Palmas de Gran Canaria, Espanha, 2014. No prelo.

OLIVEIRA, R.R.; GUELPELI, M.V.C. Corpus in Italian of the Journalism and Medical Fields. In: **Second Asia Pacific Corpus Linguistics Conference (APCLC 2014),** Abstract. Hong Kong, China, 2014 b. No prelo.

PARDO, T.A.S.; RINO, L.H.M. **TeMário: Um Corpus para Sumarização Automática de Textos.** Série de Relatórios do NILC. NILC-TR -03-09, 2003.

PARDO, T.A.S. **GistSumm: Um Sumarizador Automático Baseado na Ideia Principal de Textos.** Série de Relatórios do NILC. NILC-TR-02-13, 2002.

PARDO, T. A. S. **Sumarização automática: principais conceitos e sistemas para o português brasileiro.** São Paulo: Núcleo Interinstitucional de Linguística Computacional – NILC, Universidade de São Paulo, 2008. 14 p. (Rel. Técnico NILC-NILC-TR-08-04/ 2008).

PEIXOTO, M.D.F; BATISTA, M.G.T.R.H; CAPELO, M.J.T.S.P. **Categorização de Textos .** <http://www.di.ubi.pt/~api/text_categorization.pdf> Acesso em: 01 JUL. 2012.

QUONIAM, L.; TARAPANOFF, K.; ARAUJO JUNIOR, R. H. de; ALVARES, L. **Inteligência obtida pela aplicação de *data mining* em base de teses francesas sobre o Brasil**.*Ci. Inf.* [online]. 2001, vol.30, n.2, pp. 20-28. ISSN 0100-1965. http://dx.doi.org/10.1590/S0100-19652001000200004

RIJSBERGEN, C. J. **Information Retrieval.** Book London: Butterworths, 1979.

ROTEM, N. OTS – Open Text Summarizer (0.5.0) [Programa de Computador].*GPL licence*. Versão online disponível em: <http://www.splitbrain.org/services/ots/>. Acesso em dez. 2013.

SANTOS, J. B**. Automatizando o Processo de Estimativa de Revocação e Precisão de Funções de Similaridade** / Juliana Bonato dos Santos – Porto Alegre: Programa de Pós-Graduação em Computação, 2008. 61 f. il. Dissertação (mestrado) – Universidade

Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação. Porto Alegre, BR – RS, 2008. Orientador: Carlos Alberto Heuser; Co-orientadora: Viviane Moreira Orengo.

SCHONS, C. H. O volume de informações na internet e sua desorganização: reflexões e perspectivas. : **Inf. Inf,** Londrina vol. 12, n.1, 2007

StatPlus® (Versão trial) [Programa de Computador]. Analyst Soft Inc. Disponível em: <http://www.analystsoft.com/en/products/statplus/ >. Acesso em jan. 2014.

TAN, P. N.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining**. Addison-Wesley, 2006.

WIVES, L. K. **Utilizando conceitos como descritores de textos para o processo de identificação de conglomerados (clustering) de documentos** – Tese (doutorado) – Universidade Federal do Rio Grande do Sul.Programa de Pós-graduação em Computação, Porto Alegre, BR – RS, Brasil, 2004.

WIVES, L. K.; LOH, S. **Tecnologias de descoberta de conhecimento em informações textuais** (ênfase em agrupamento de informações). In: OFICINA DE INTELIGÊNCIA ARTIFICIAL (OIA), III, 1999, Pelotas, RS. Proceedings... Pelotas: EDUCAT, 1999. p28-48. (TUTORIAL)

ZOUBI, M. B.; RAWI, M.A. Efficient Approach for Computing Silhouette Coefficients. **Journal of Computer Science** Volume 4 Page No.: 252–255, 2008.

APÊNDICES

APÊNDICES A

APÊNDICES A – GRÁFICOS COM OS RESULTADOS DE *COEFICIENTE SILHOUETTE COESÃO E ACOPLAMENTO*

O Apêndice A traz os gráficos com os resultados de Coeficiente *Silhouette*, Coesão e Acoplamento obtidos pelos agrupamentos de texto nos idiomas espanhol e italiano. Os resultados foram divididos por idioma e subdividido por taxa de compressão[7].

1. Idioma Espanhol

1.1.Taxa de Compressão 50%

A Figura 22 indica que para o idioma espanhol, com taxa de compressão de 50%, houve bastante desequilíbrio nos resultados, pois os textos sumarizados com o sumarizador *OTS* obtiveram os resultados bem superiores aos textos sumarizados com os sumarizadores *Copernic*, *BLMSumm* e *Intellexer*, já que estes obtiveram resultados abaixo de 35%.

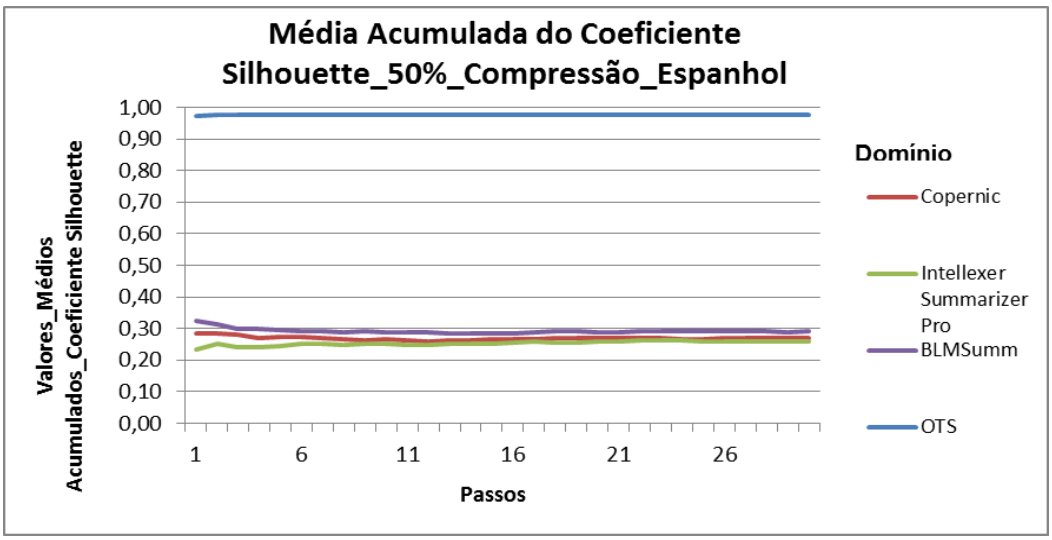

Figura 22. Média acumulada de Coeficiente *Silhouette* com taxa de compressão de 50% do idioma espanhol

[7] Os resultados alcançados pelo Idioma Espanhol e Italiano não conseguiram alcançar 50% (0,5) de *Coesão, Acoplamento* dessa forma, para melhor visualização dos resultados os gráficos gerados nas figuras22, 23, 25, 26, 28, 31, 32, 34, 35, 37, 38, 40, 41, 43 foram feitos com escala de 0,0 a 0,5 com variação de 0,1.

A Figura 23 (média acumulada de Coesão) apresenta os resultados da clusterização realizada com cada um dos sumarizadores automáticos. Observa-se que há um equilíbrio entre os resultados, mas o *BLMSumm*, possui uma ligeira vantagem.

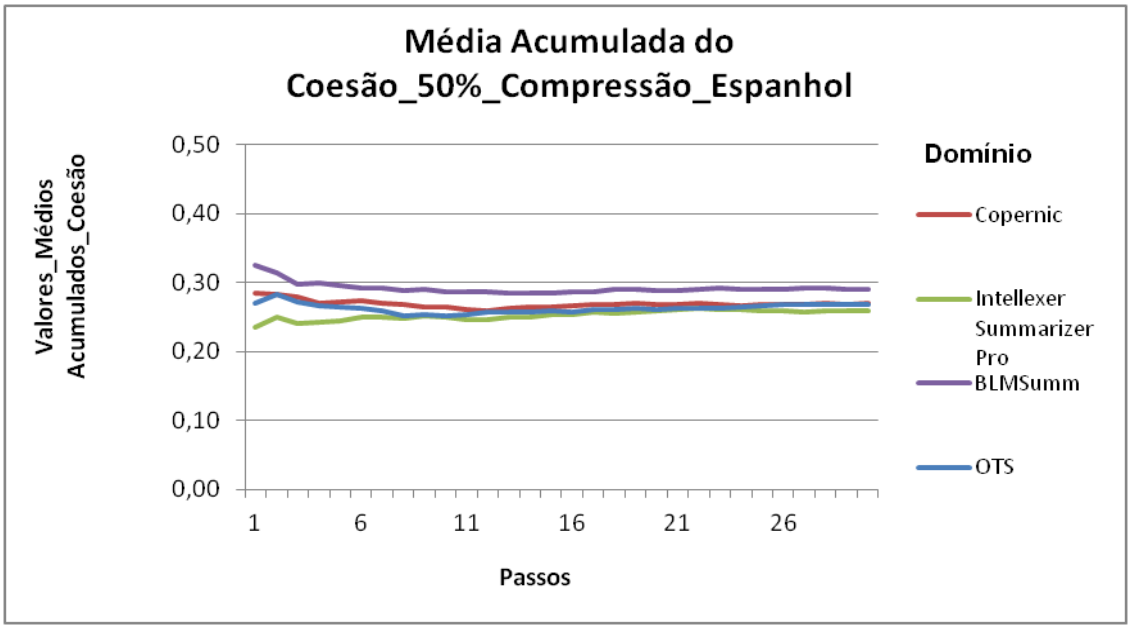

Figura 23. Média acumulada de Coesão com taxa de compressão de 50%

A Figura 24 (média acumulada de Acoplamento) apresenta os resultados da clusterização realizada com cada um dos sumarizadores automáticos. Observa-se que o *Copernic* apresentou os resultados mais significativos.

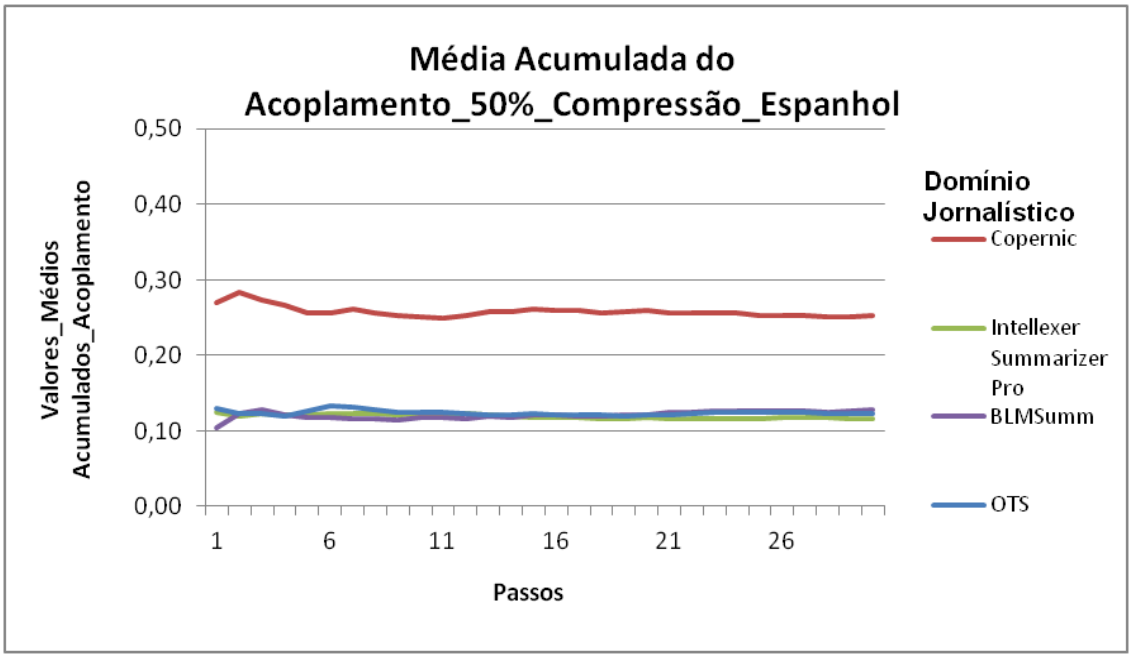

Figura 24. Média acumulada de Acoplamento com taxa de compressão de 50%

1.2. Taxa de Compressão 70%

A Figura 25 indica que para o idioma espanhol, com taxa de compressão de 70%, houve bastante equilíbrio nos resultados, mas, o sumarizador *Intellexer* obteve uma ligeira vantagem, seguido pelo *Copernic*, *OTS* e *BLMSumm*.

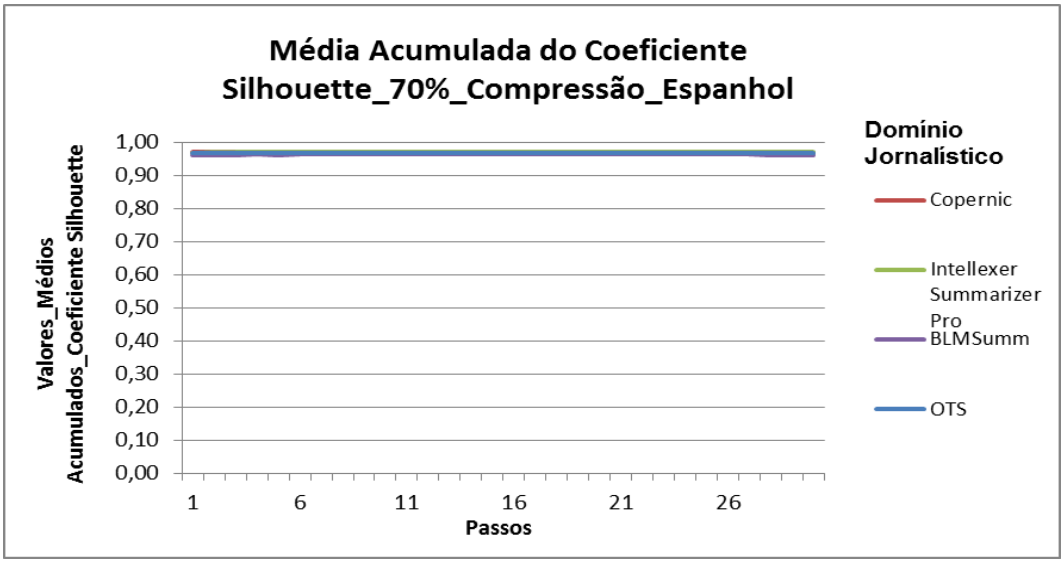

Figura 25. Média acumulada de Coeficiente *Silhouette* com taxa de compressão de 70% do idioma espanhol

É possível perceber que na Figura 26 (média acumulada de Coesão) o equilíbrio dos resultados persiste, entretanto é observado que o sumarizador *Copernic* e *BLMSumm* apresentam uma ligeira vantagem.

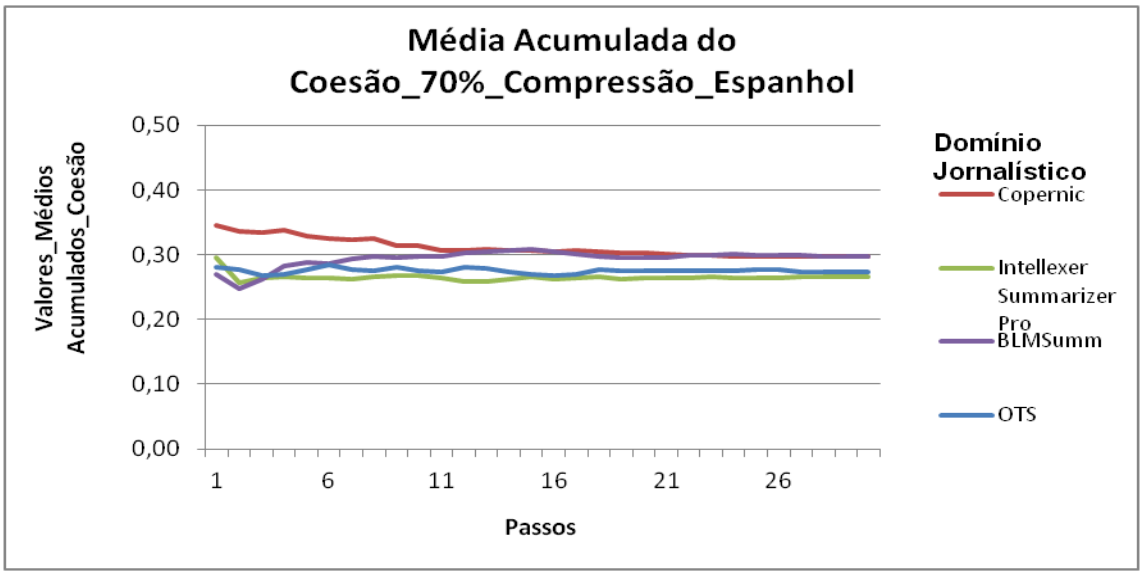

Figura 26. Média acumulada de Coesão com taxa de compressão de 70%

Na Figura 27 (Media Acumulada de Acoplamento) os resultados mais satisfatórios foram obtidos com os sumarizadores *OTS* e *BLMSumm*.

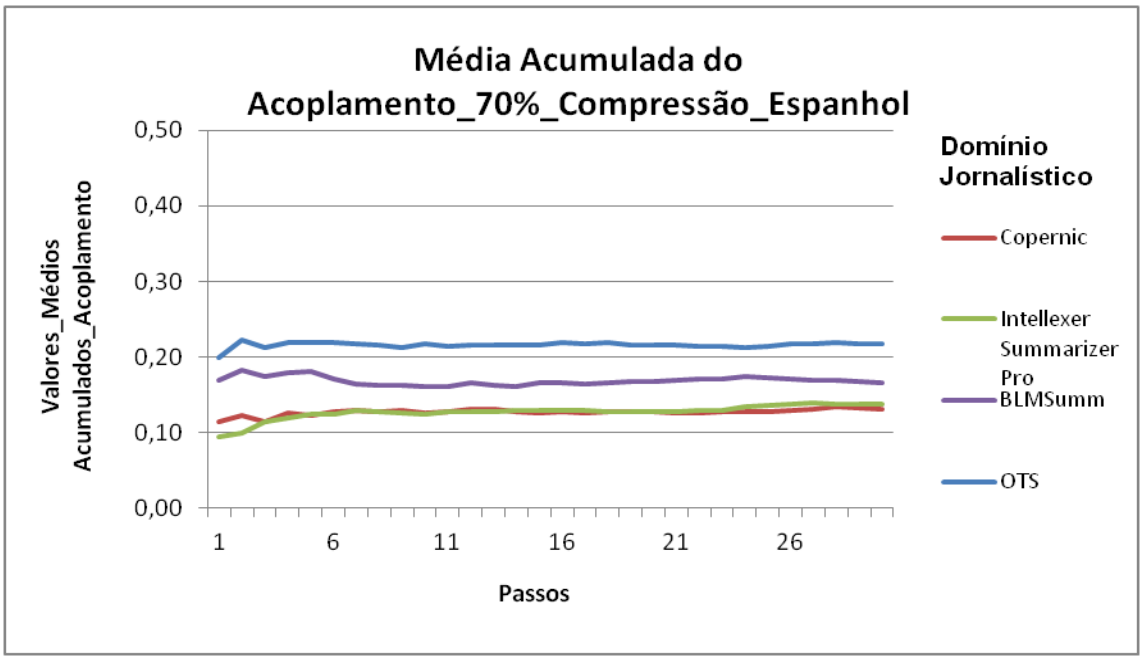

Figura 27. Média acumulada de Acoplamento com taxa de compressão de 70%

1.3.Taxa de Compressão 80%

É possível perceber na Figura 28 que para o idioma espanhol, com taxa de compressão de 80%, o *Intellexer* apresentou os resultados mais satisfatórios e o *OTS* os valores mais baixos. O *Copernic* obteve o segundo lugar.

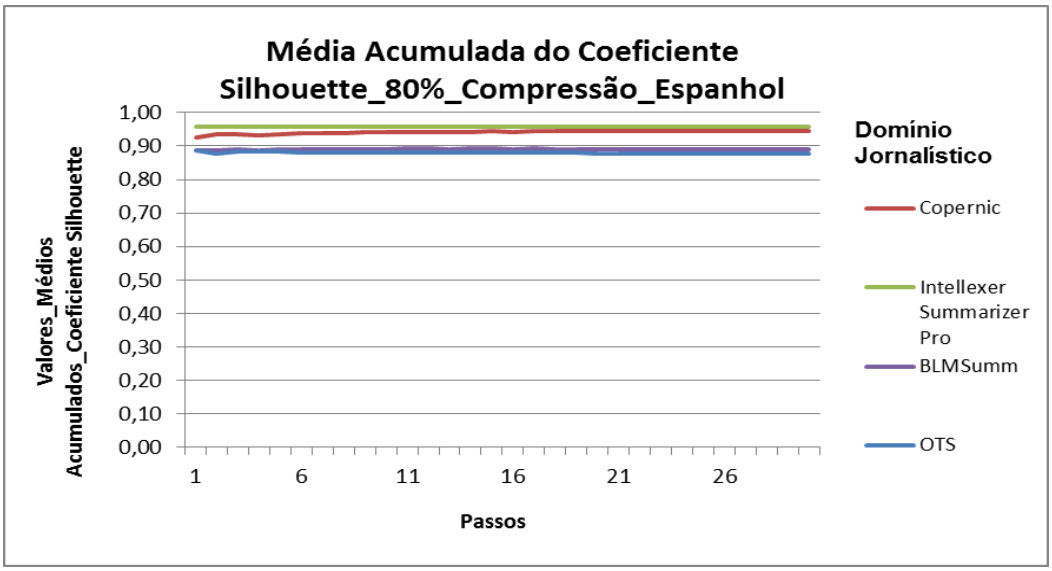

Figura 28. Média acumulada de Coeficiente *Silhouette* com taxa de compressão de 80% do idioma espanhol

É possível perceber que na Figura 29 (média acumulada de Coesão) que o sumarizador *Copernic* apresentou o resultado mais satisfatório.

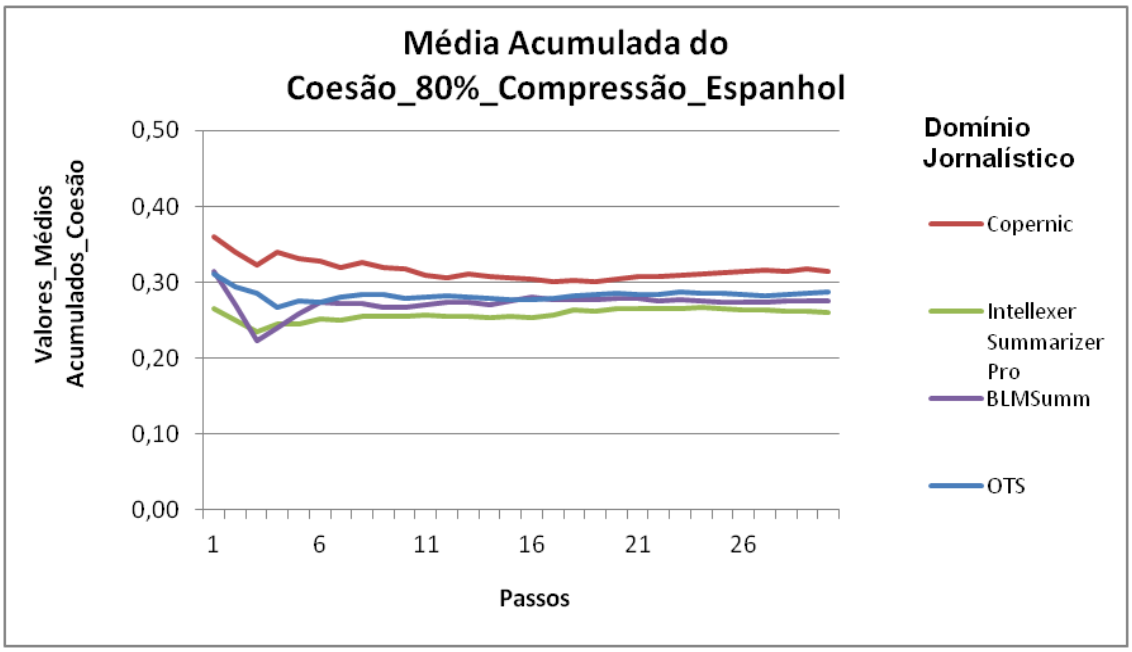

Figura 29. Média acumulada de Coesão com taxa de compressão de 80%

É possível perceber na figura 30 (média acumulada de acoplamento) que o sumarizador *BLMSumm* apresenta um pico de variação muito altos, mas depois decai. Pode-se observar que os valores mais equilibrados são os *Copernic*.

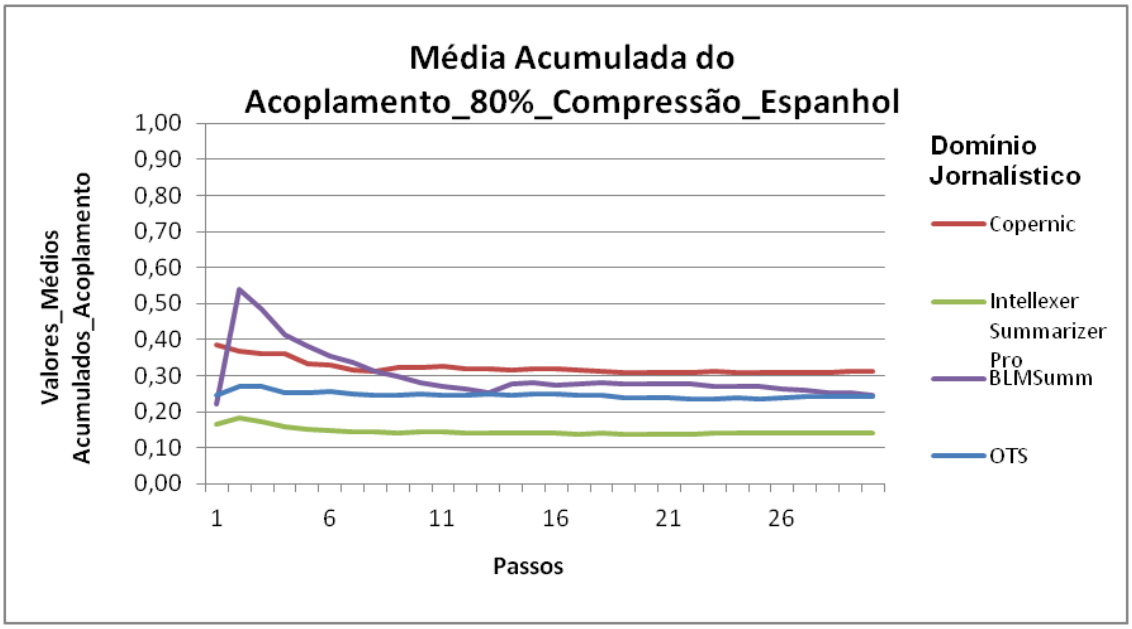

Figura 30. Média acumulada de Acoplamento com taxa de compressão de 80%

1.4. Taxa de Compressão de 90%

Na Figura 31 observa-se que no idioma espanhol, os resultados mais satisfatórios, quando a taxa de compressão aplicada foi de 90%, foram obtidos pelo sumarizador *OTS*. O pior resultado foi obtido pelo *Intellexer*.

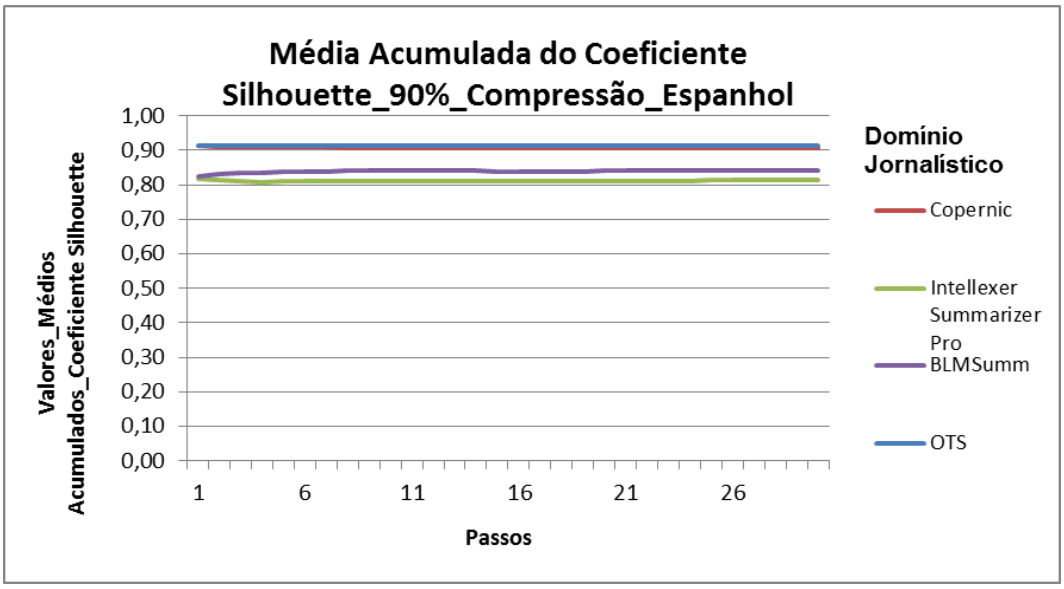

Figura 31. Média acumulada de Coeficiente *Silhouette* com taxa de compressão de 90% do idioma espanhol

É possível perceber que na Figura 32 (média acumulada de Coesão) que o sumarizador *Copernic* apresentou o resultado mais estável.

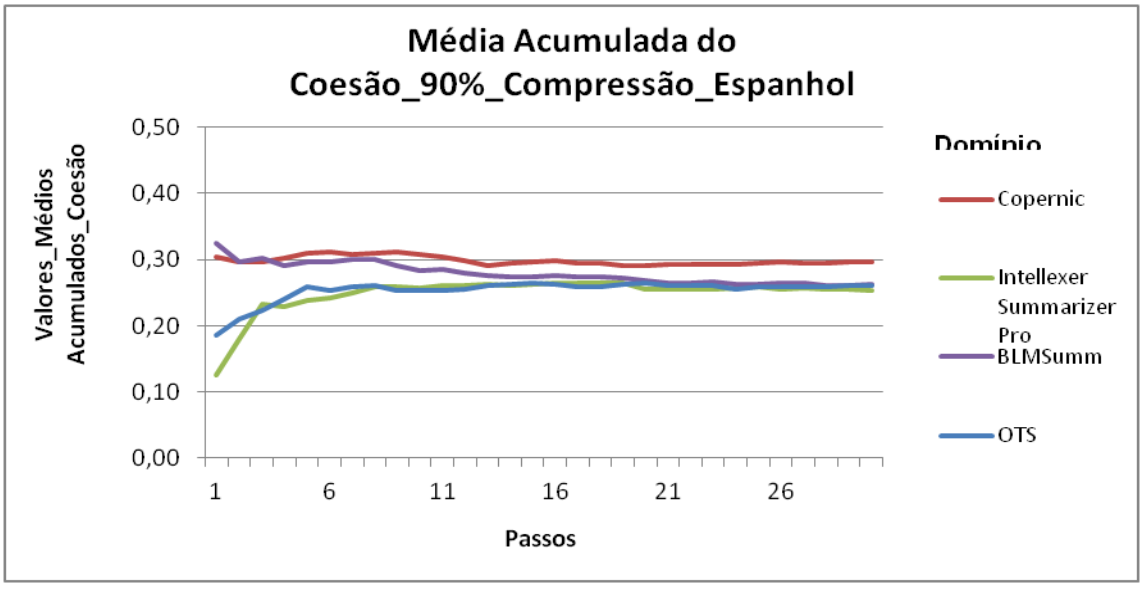

Figura 32. Média acumulada de Coesão com taxa de compressão de 90%

Na Figura 33 (Média Acumulada de Acoplamento) o sumarizador que apresentou os resultados mais aceitáveis foi o *Intellexer*. Os outros sumarizadores apresentaram equilíbrio em seus resultados.

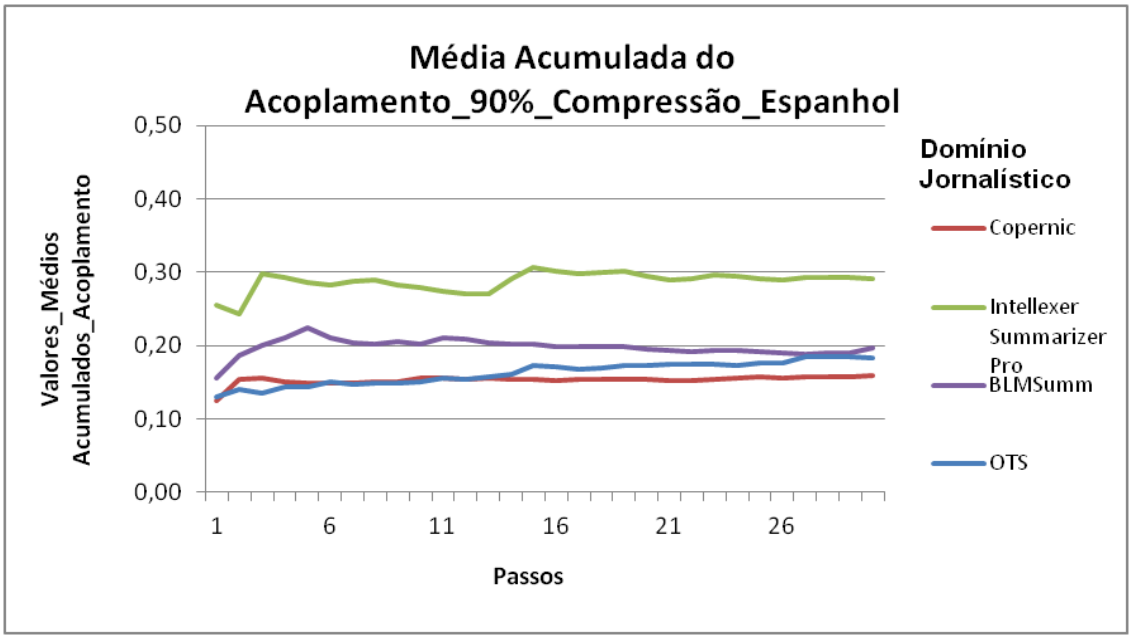

Figura 33. Média acumulada de Acoplamento com taxa de compressão de 90%

2. Idioma Italiano

2.1. Taxa de Compressão 50%

Na Figura 34 é possível observar que para o idioma italiano, com taxa de compressão de 50%, houve bastante equilíbrio entre os resultados e o *Copernic* apresentou uma ligeira vantagem sobre os outros.

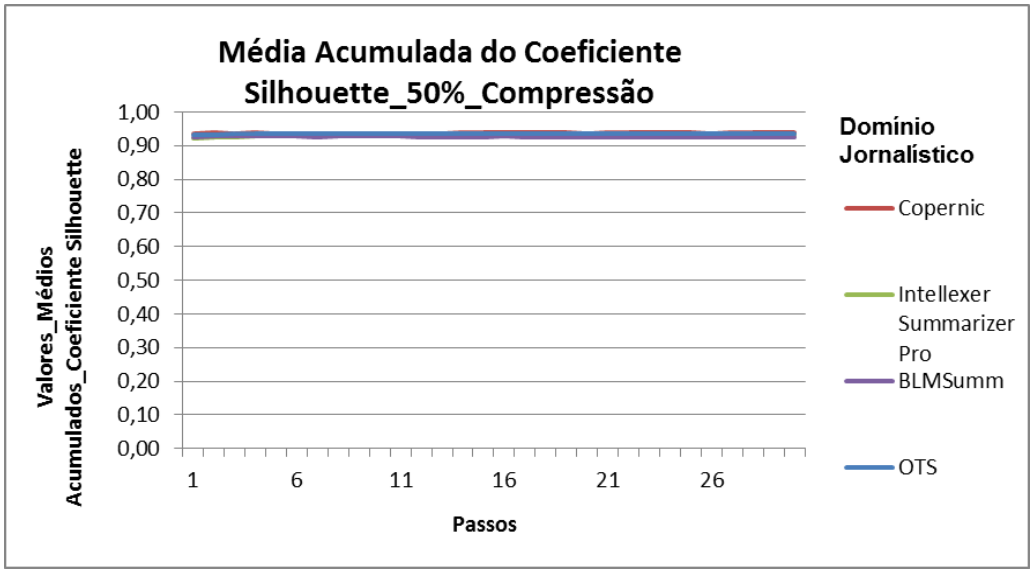

Figura 34. Média acumulada de Coeficiente *Silhouette* com taxa de compressão de 50% do idioma italiano

A Figura 35 (média acumulada de Coesão) apresenta os resultados da clusterização realizada com cada um dos sumarizadores automáticos. Observa-se que há um ligeiro equilíbrio entre os resultados, mas o *BLMSumm* e o *Intellexer*, apresentam os valores mais altos.

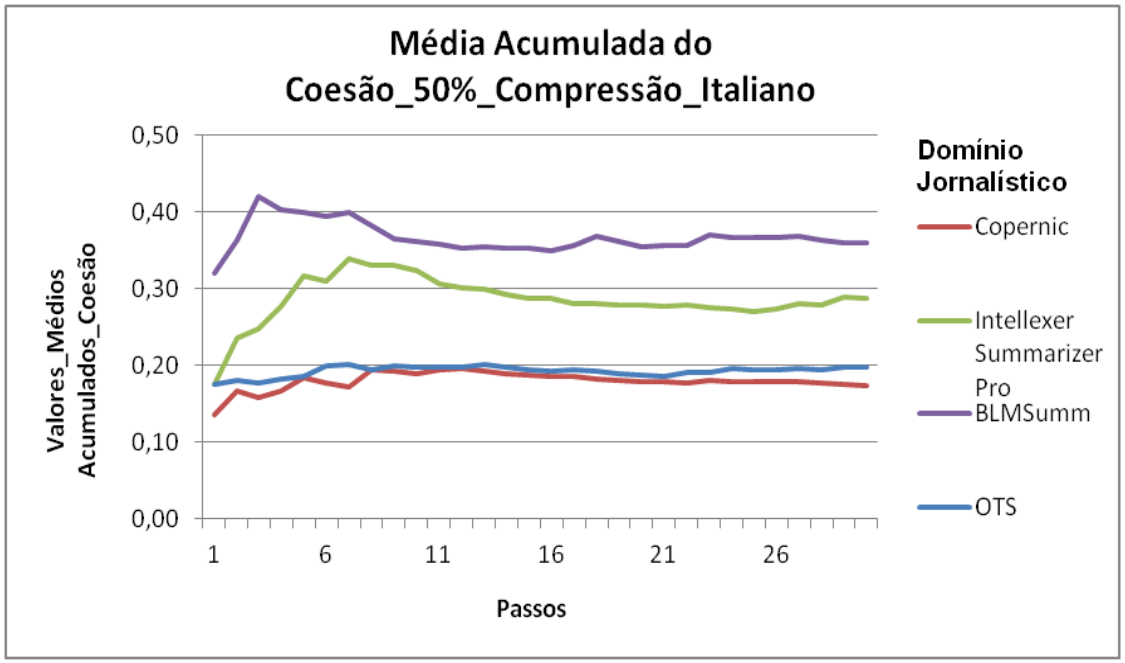

Figura 35. Média acumulada de Coesão com taxa de compressão de 50%

Nota-se na Figura 36 (Média Acumulada Acoplamento) que os valores mais satisfatórios pertencem aos sumarizadores *BLMSumm* e *Intellexer*. O valor mais baixo é do *Copernic*.

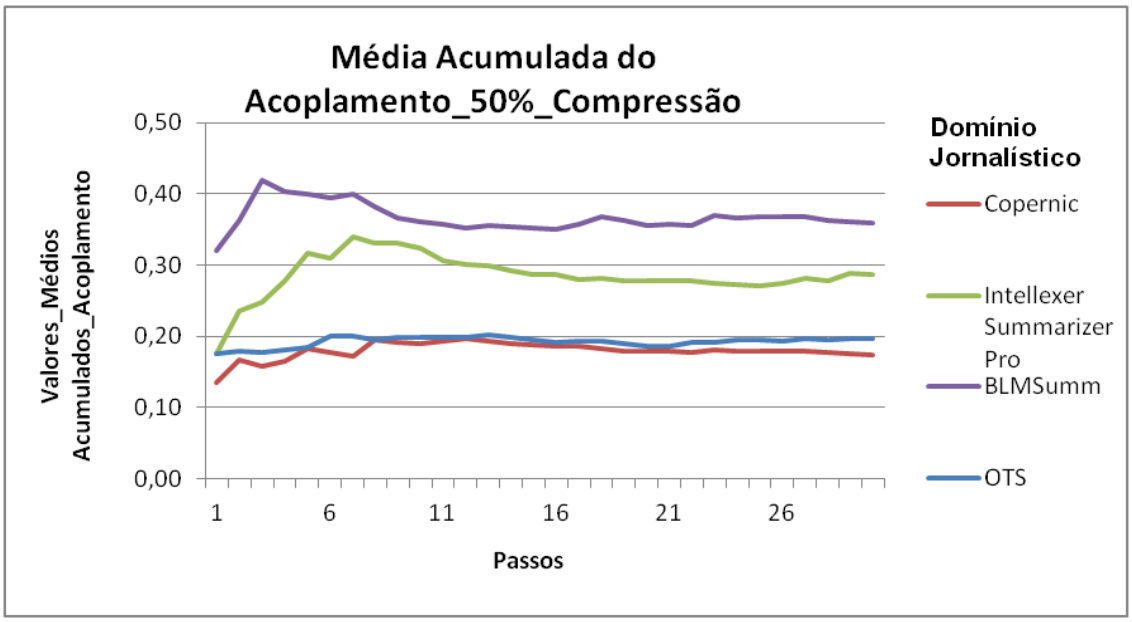

Figura 36. Média acumulada de Acoplamento com taxa de compressão de 50%

2.2.Taxa de Compressão 70%

Na Figura 37 é possível observar que para o idioma italiano, com taxa de compressão de 70%, houve bastante equilíbrio entre os resultados e o OTS apresentou os melhores valores, seguido pelo *Copernic*.

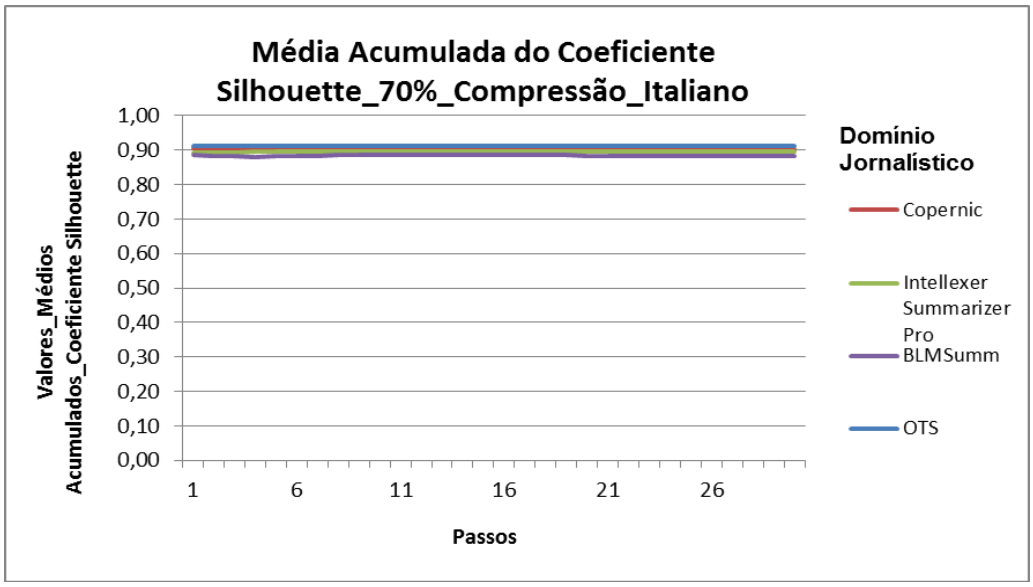

Figura 37. Média acumulada de Coeficiente *Silhouette* com taxa de compressão de 70% do idioma italiano

Nota-se na Figura 38 (Média Acumulada Coesão) que os resultados estão balanceados. É possível perceber que o *Copernic* obteve uma ligeira vantagem sobre os outros sumarizadores.

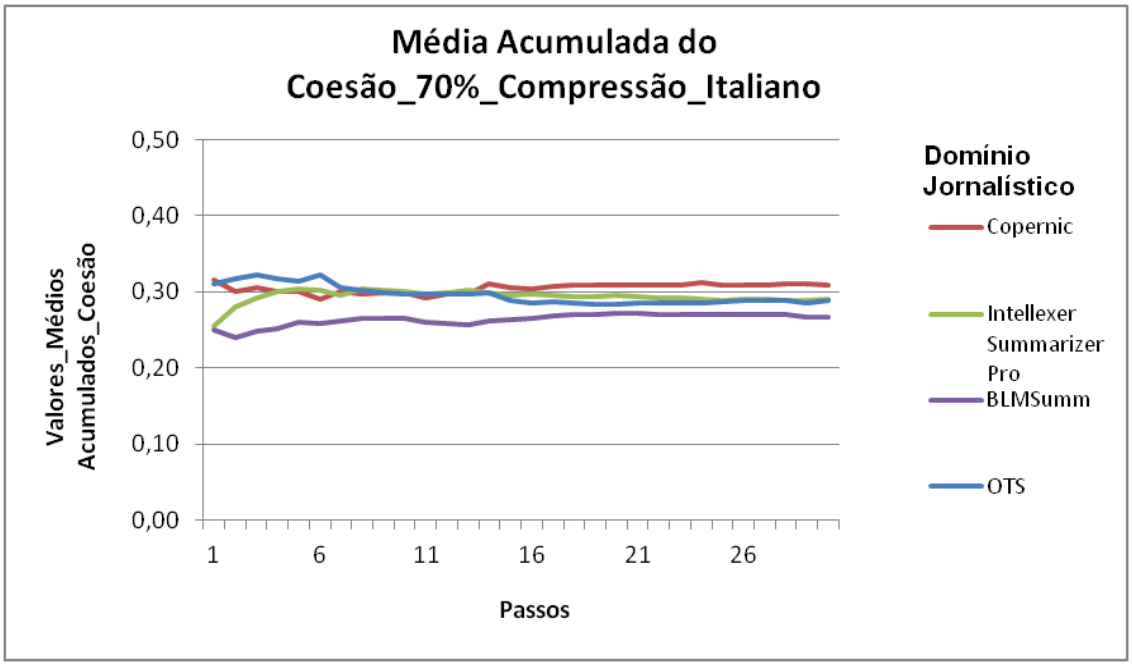

Figura 38. Média acumulada de Coesão com taxa de compressão de 70%

A Figura 39 (média acumulada de Acoplamento) apresenta os resultados da clusterização realizada com cada um dos sumarizadores automáticos. Observa-se que há um ligeiro equilíbrio entre os resultados, mas o *OTS* e o *BLMSumm*, apresentam os valores mais altos.

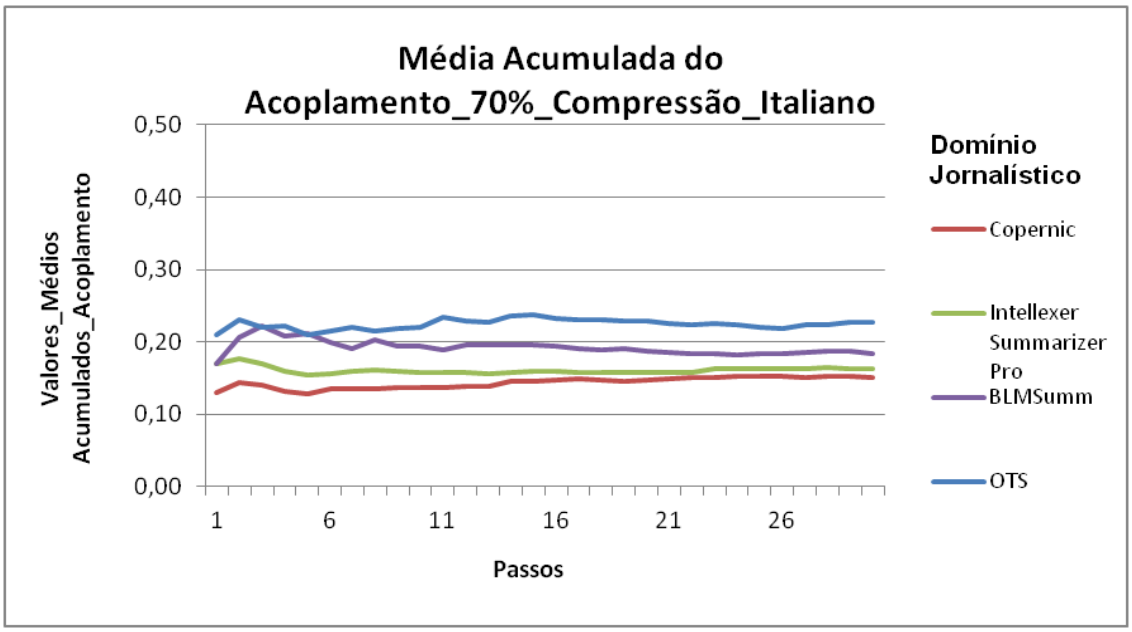

Figura 39. Média acumulada de Acoplamento com taxa de compressão de 70%

2.3.Taxa de Compressão 80%

Para uma taxa de compressão de 80%, no idioma italiano, a Figura 40 mostra que o sumarizador que apresentou os melhores valores foi o *Copernic*. O *Intellexer* teve os resultados mais baixos.

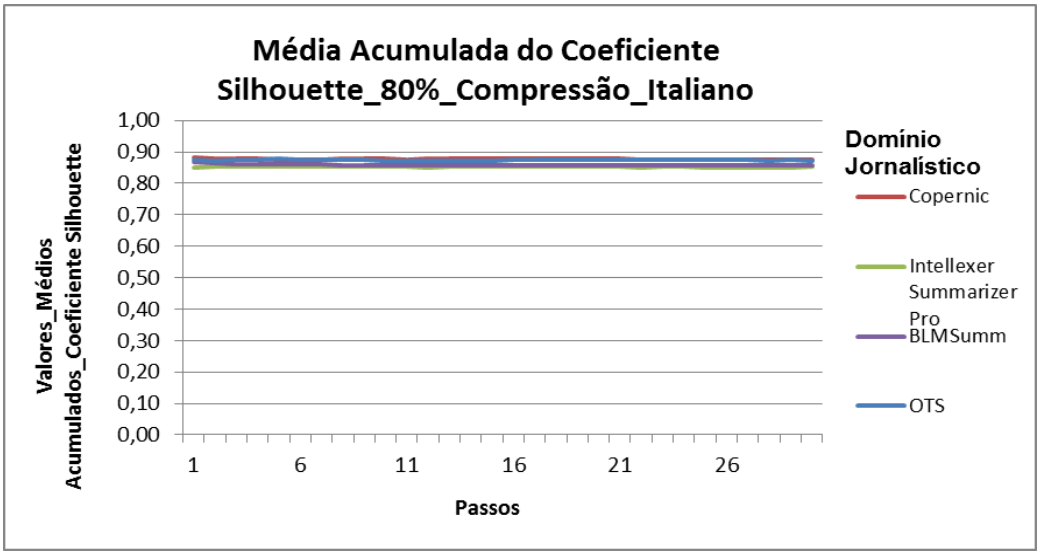

Figura 40. Média acumulada de Coeficiente *Silhouette* com taxa de compressão de 80% do idioma italiano

A Figura 41 (média acumulada de Coesão) apresenta os resultados da clusterização realizada com cada um dos sumarizadores automáticos. Observa-se que houve um equilíbrio entre os resultados, entretanto o *Copernic* e o *Intellexer* apresentaram os melhores valores.

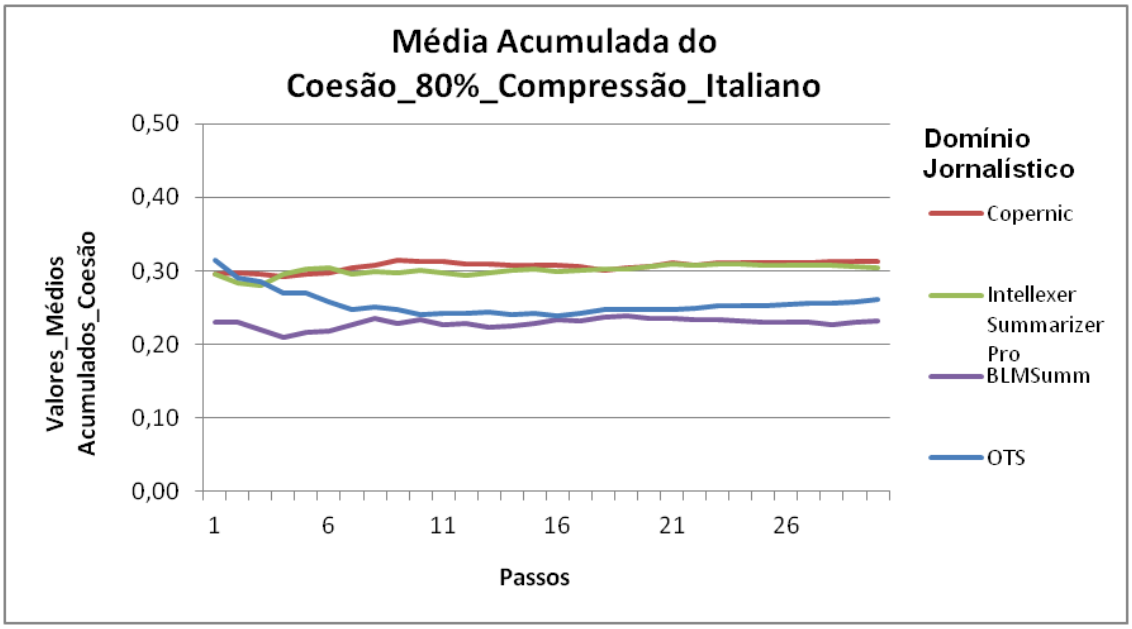

Figura 41. Média acumulada de Acoplamento com taxa de compressão de 80%

Na Figura 42 (Média Acumulada de Acoplamento) o sumarizador que apresentou os resultados mais significativos foi o OTS. Os outros sumarizadores apresentaram equilíbrio em seus resultados.

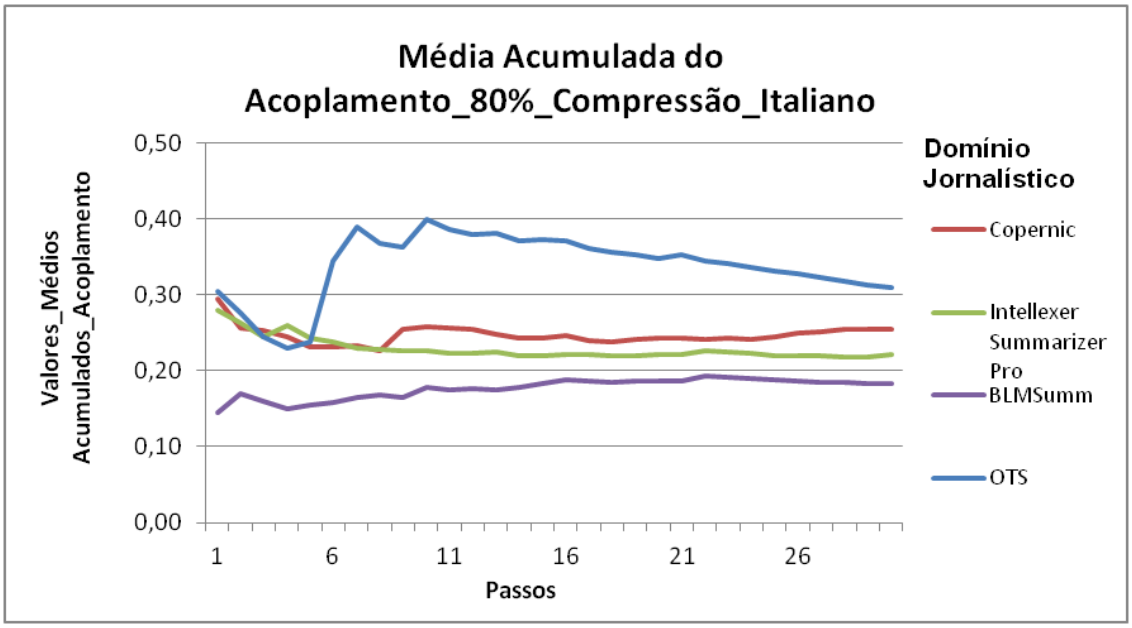

Figura 42. Média acumulada de Acoplamento com taxa de compressão de 80%

2.4. Taxa de Compressão 90%

A Figura 43 mostra que o sumarizador automático que apresentou os melhores resultados para uma taxa de compressão de 90% foi o *BLMSumm*. O *OTS* obteve os valores mais baixos.

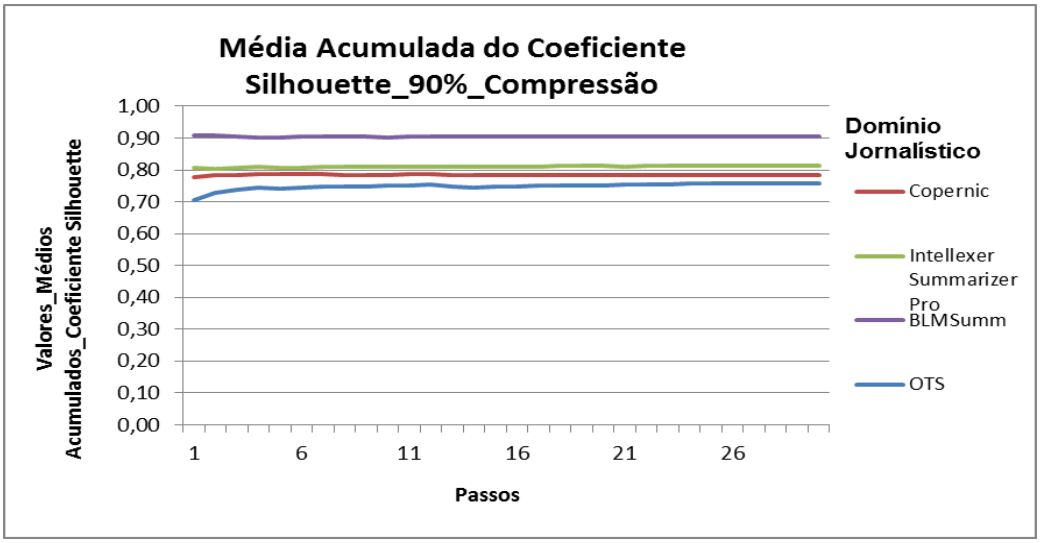

Figura 43. Média acumulada de Coeficiente *Silhouette* com taxa de compressão de 90% do idioma italiano

Na Figura 44 (Média Acumulada de Coesão) o sumarizador que apresentou os resultados mais aceitáveis foi o OTS, seguido pelo *Intellexer*. O resultado mais baixo foi do sumarizador *BLMSumm*.

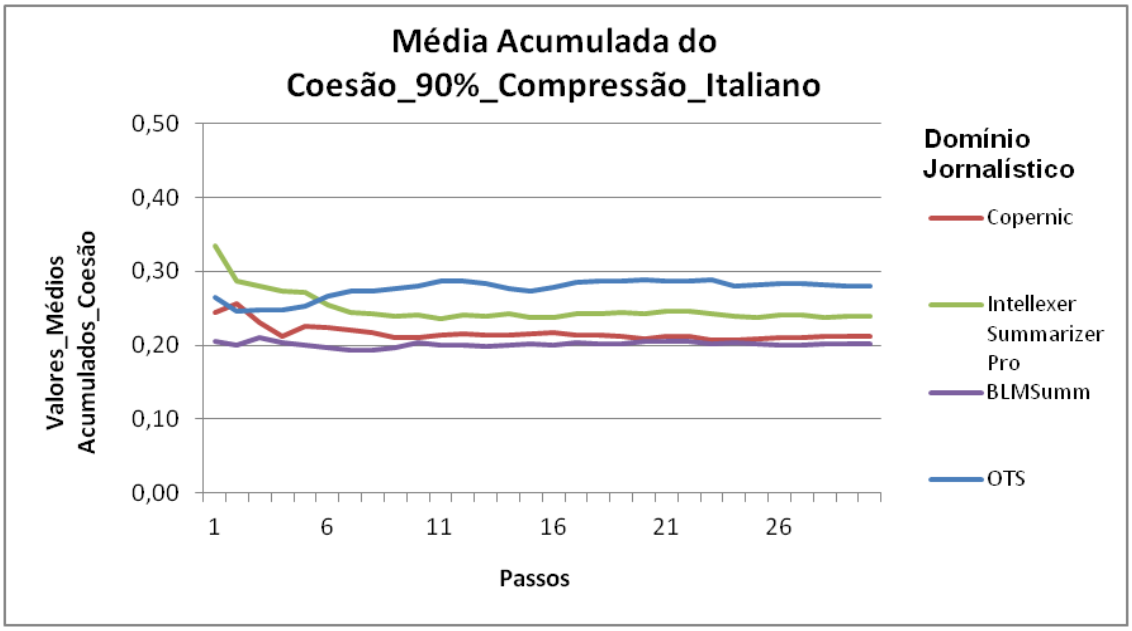

Figura 44. Média acumulada de Coesão com taxa de compressão de 90%

A Figura 45 (média acumulada de Acoplamento) apresenta os resultados da clusterização realizada com cada um dos sumarizadores automáticos. Observa-se que houve um desequilíbrio entre os resultados, pois o *OTS* apresentou valores extremamente altos em relação aos outros sumarizadores.

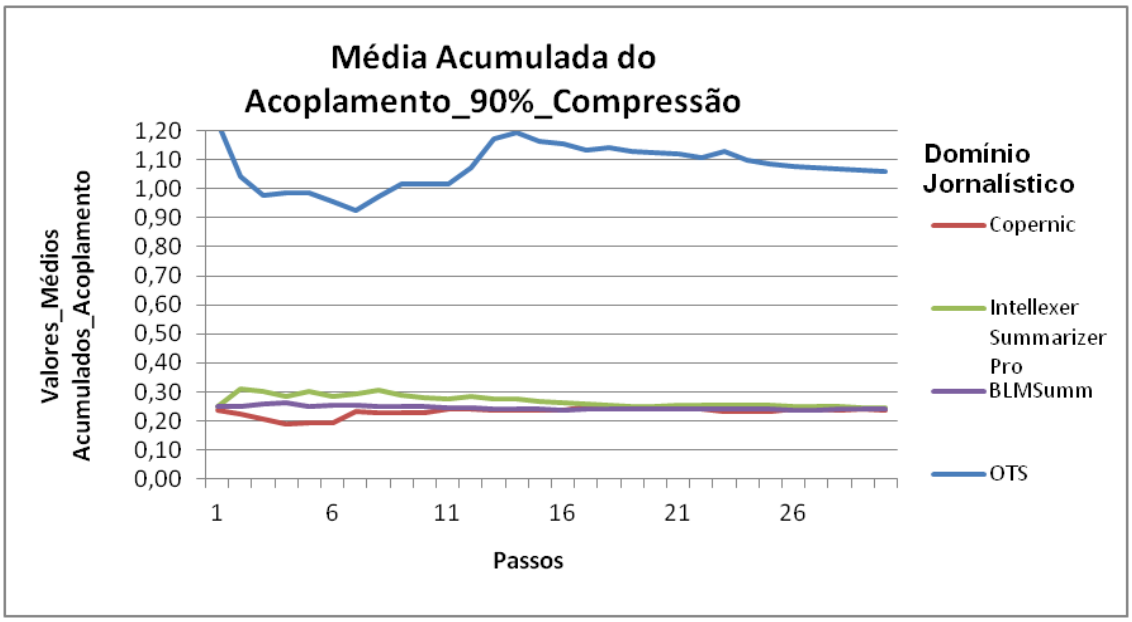

Figura 45. Média acumulada de Acoplamento com taxa de compressão de 90%

APÊNDICE B

APÊNDICES B – GRÁFICOS COM OS RESULTADOS DE *F-MEASURE,RECALL* E *PRECISION*

O Apêndice A traz os gráficos com os resultados de *F-Measure, Recall* e *Precision* obtidos pelos agrupamentos de texto nos idiomas espanhol e italiano. Os resultados foram divididos por idioma e subdividido por taxa de compressão.

1. Idioma Espanhol

1.1. Taxa de Compressão 50%

A Figura [8] 46 indica que para o idioma espanhol, com taxa de 50%, houve um sensível equilíbrio nos resultados, mas o sumarizador *BLMsumm* obteve os melhores resultados. Observa-se ainda que, de forma geral, os resultados obtidos foram baixos, já que, os sumarizadores utilizados não conseguiram alcançar 30%, ou seja, 0,2 de *F-measure.*

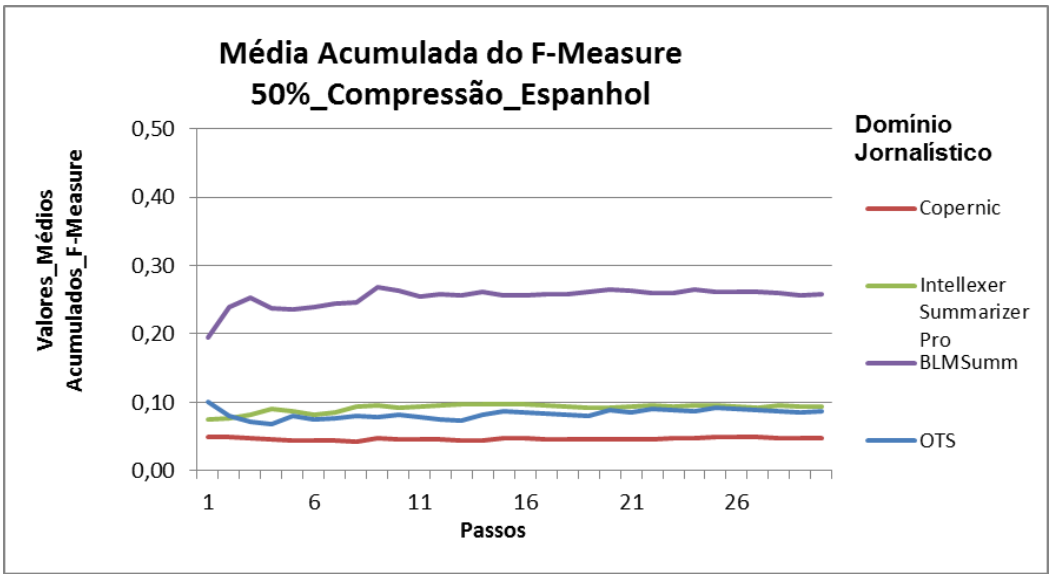

Figura 46. Média acumulada de *F-Measure* com taxa de compressão de 50% do idioma espanhol

A Figura 47 (média acumulada de *Recall*) apresentam os resultados dos agrupamentos de texto para cada um dos sumarizadores automáticos. Observa-se que o resultado do sumarizador *BLMSumm* é o mais satisfatório.

[8] Os resultados alcançados pelo Idioma Espanhol e Italiano não conseguiram alcançar 50% (0,5) de *F-Measure,* dessa forma, para melhor visualização dos resultados os gráficos gerados nas figuras 45,46, 48,49, 51,52, 54,55, 57,58,60, 61, 63 e 64 foram feitos com escala de 0,0 a 0,5 com variação de 0,1.

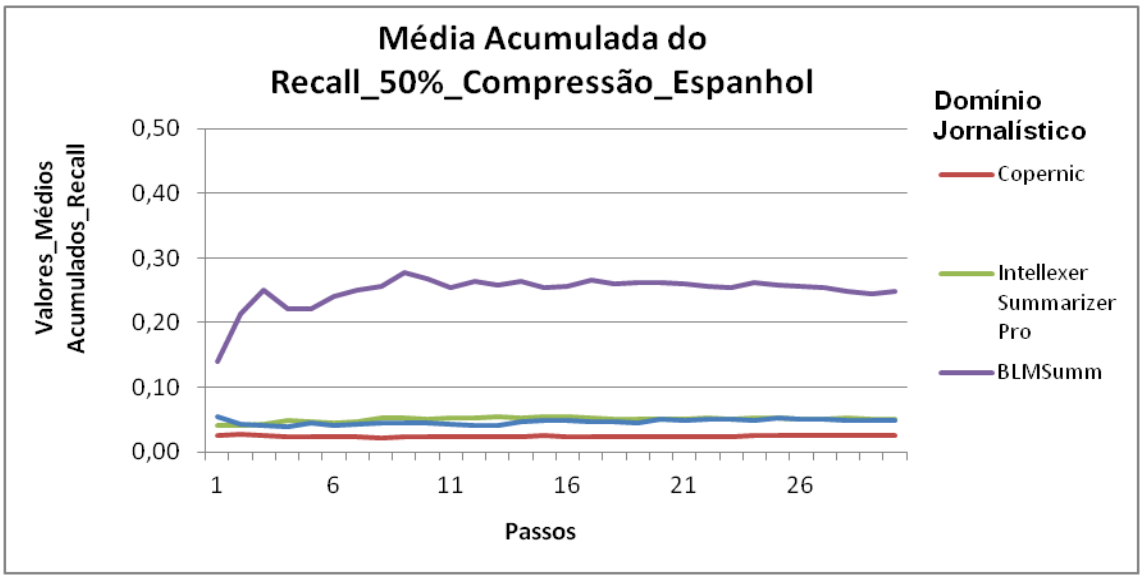

Figura 47. Média acumulada de *Recall* com taxa de compressão de 50%

Na Figura 48 (Média Acumulada de *Precision*) os sumarizadores*Intellexer*, *Copernic* e *OTS* apresentam o mesmo resultado. Já o *BLMSumm* apresenta o pior.

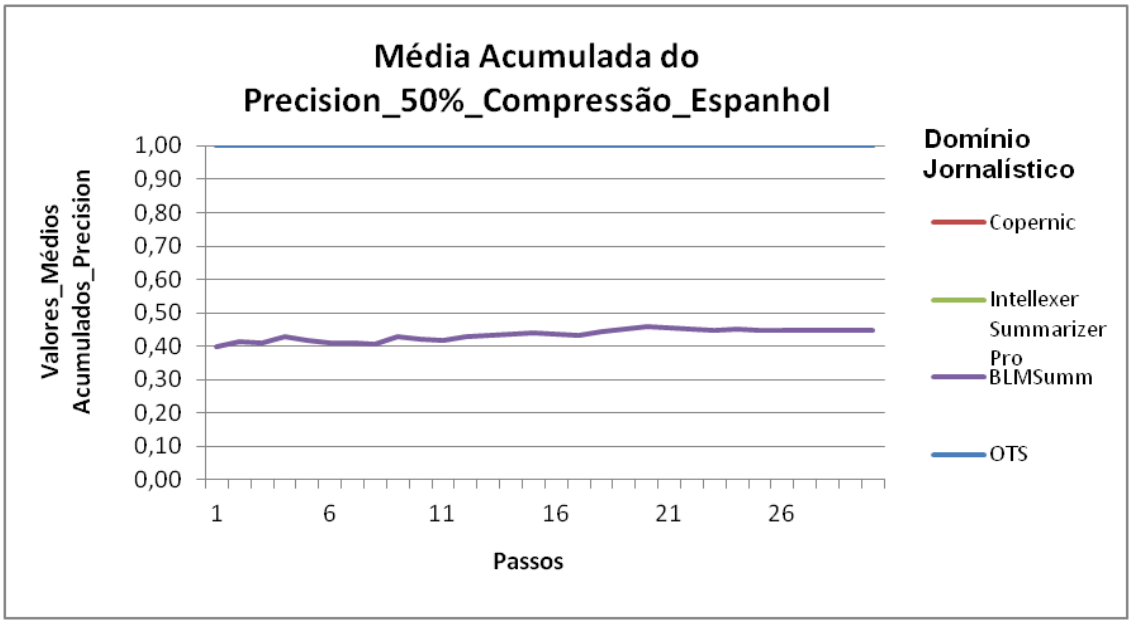

Figura 48. Média acumulada de *Precision* com taxa de compressão de 50%

1.2. Taxa de Compressão 70%

É possível observar na Figura 49 que para o idioma espanhol, com taxa de compressão de 70%, o *BLMSumm* apresentou os resultados mais satisfatório e o *Copernic* os valores mais baixos. O *Intellexer* obteve o segundo lugar.

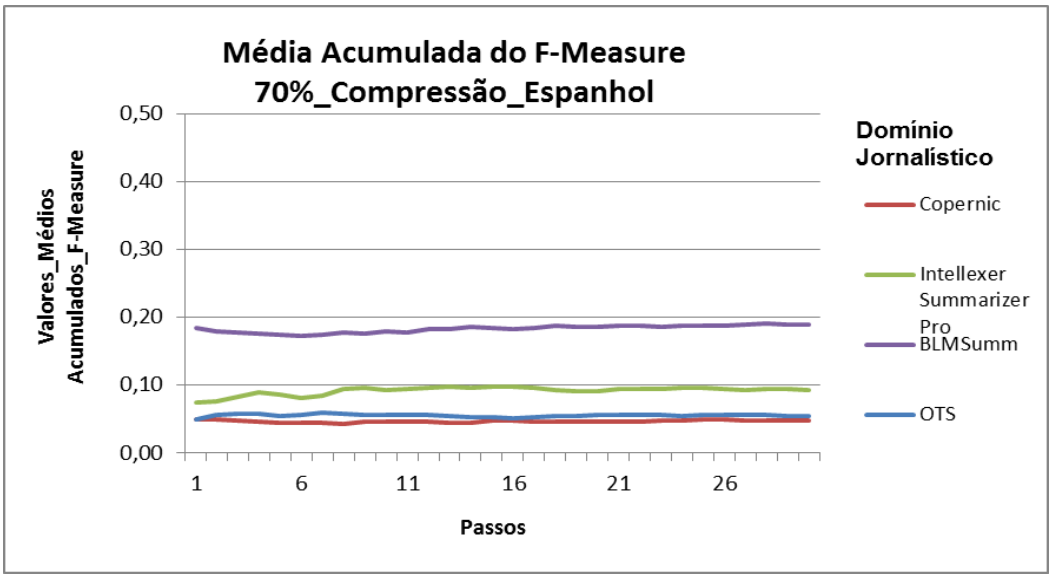

Figura 49. Média acumulada de *F-Measure* com taxa de compressão de 70% do idioma espanhol

É possível perceber que na Figura 50 (média acumulada de *Recall*) que o sumarizador *BLMSumm* apresentou os valores mais altos, e o Intellexer apresentou os valores mais baixos.

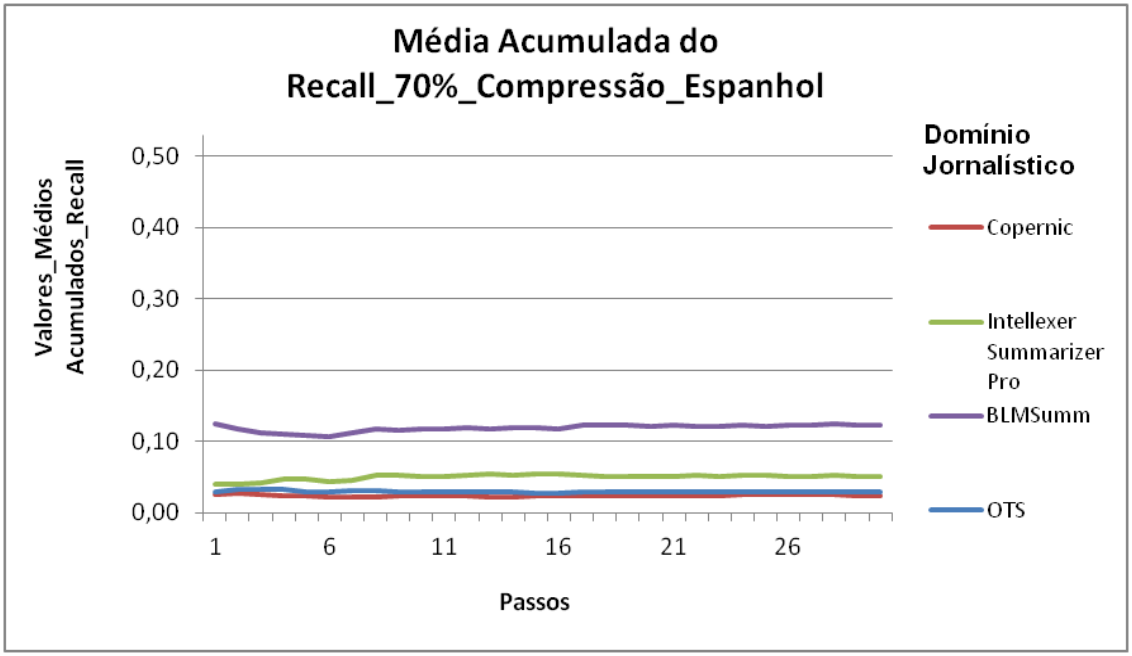

Figura 50. Média acumulada de *Recall* com taxa de compressão de 70%

Na Figura 51 (Média Acumulada de *Precision*) os sumarizadores *Intellexer*, *Copernic* e *OTS* apresentam o mesmo resultado. Já o *BLMSumm* apresenta o resultado mais baixo.

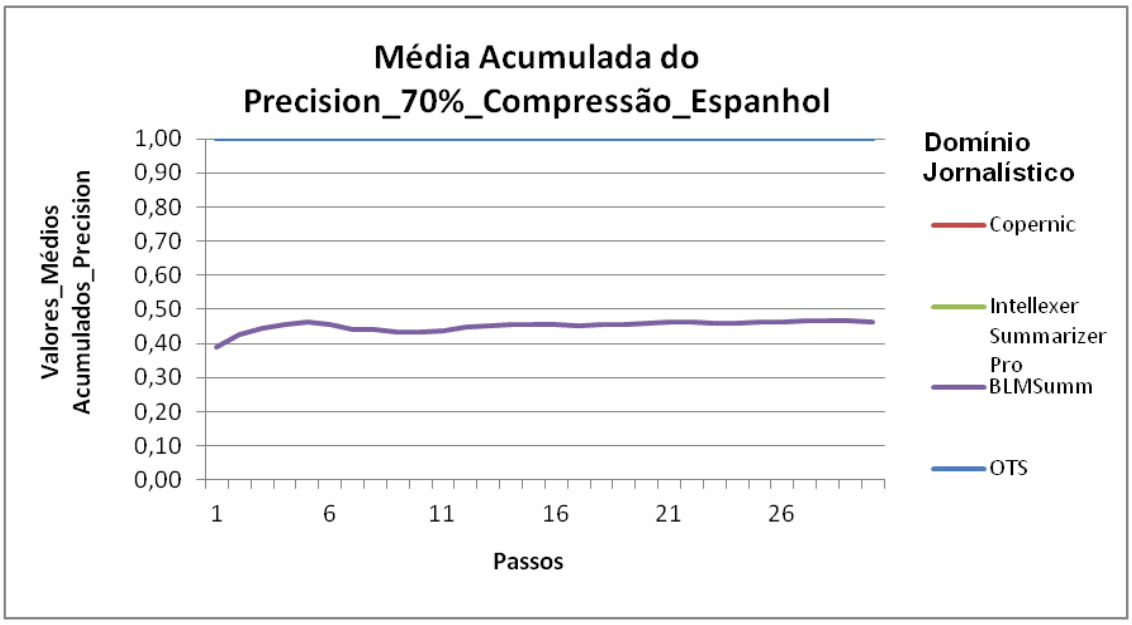

Figura 51. Média acumulada de *Precision* com taxa de compressão de 70%

1.3.Taxa de Compressão 80%

Na Figura 52 observa-se que no idioma espanhol os melhores resultados, quando a taxa de compressão foi de 80%, foram obtidos pelo sumarizador *OTS*. O resultado mais baixo foi obtido pelo *Intellexer*.

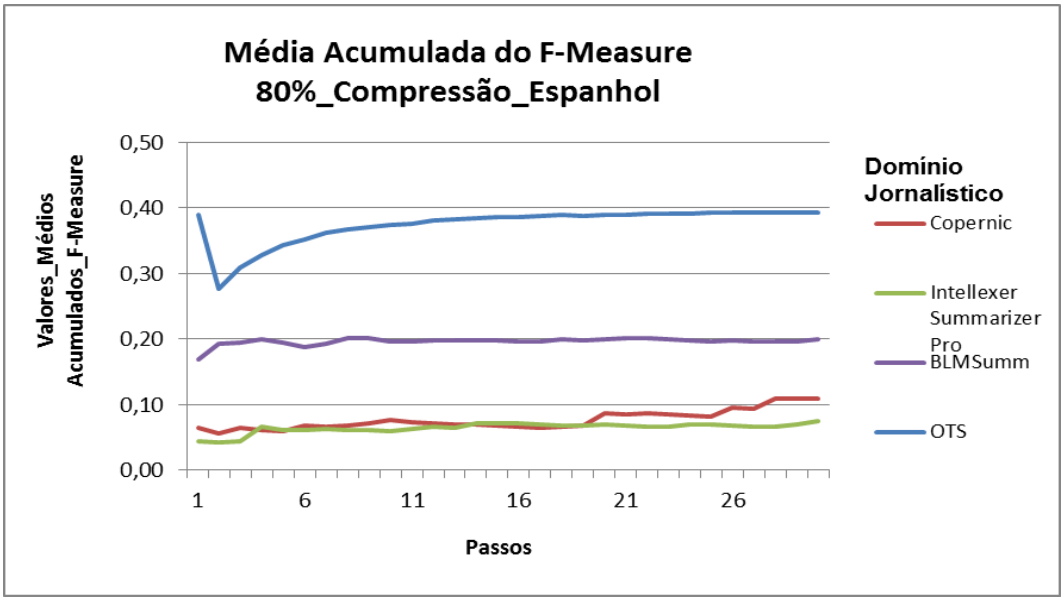

Figura 52. Média acumulada de *F-Measure*com taxa de compressão de 80% do idioma espanhol

É possível perceber que na Figura 53 (média acumulada de *Recall*) que o sumarizador OTS apresentou os valores mais altos, e o Copernic apresentou os valores mais baixos.

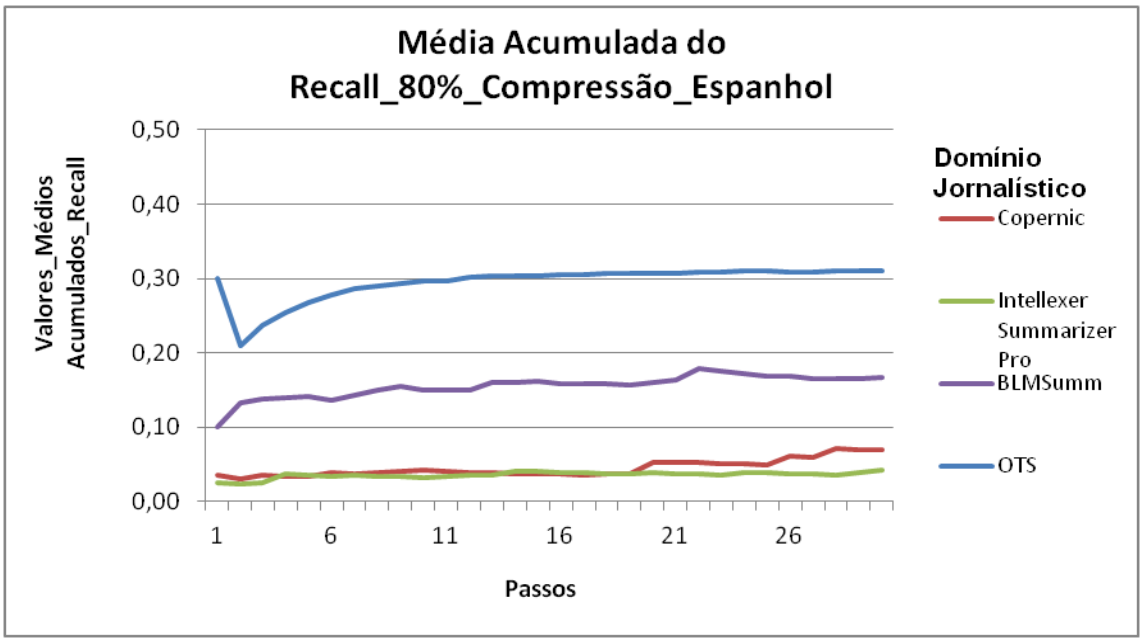

Figura 53. Média acumulada de *Recall* com taxa de compressão de 80%

Na Figura 54 (Média Acumulada de *Precision*) os sumarizadores*Intellexer*, *Copernic* e *OTS* apresentam o mesmo resultado. Já o *BLMSumm* apresenta os valores mais baixos.

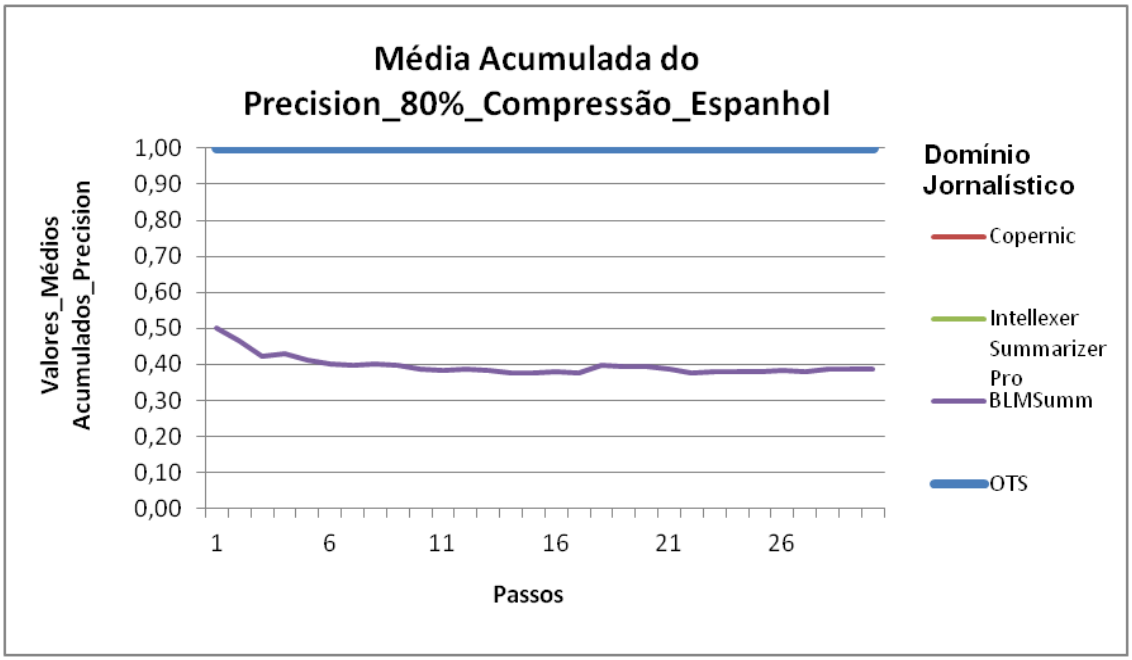

Figura 54. Média acumulada de *Precision* com taxa de compressão de 80%

1.4.Taxa de Compressão 90%

Na Figura 55 é possível perceber que para o idioma espanhol, com taxa de 90%, o *BLMSumm* apresentou os melhores resultados seguido pelo *Intellexer*. O *Copernic* apresentou os piores resultados.

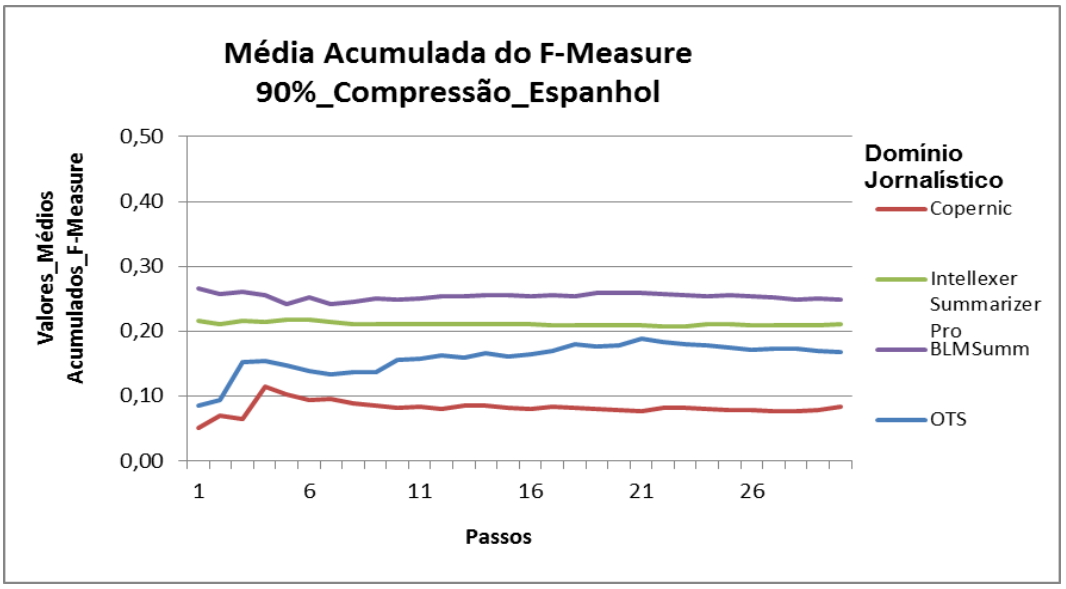

Figura 55. Média acumulada de *F-Measure* com taxa de compressão de 90% do idioma espanhol

É possível perceber que na Figura 56 (média acumulada de *Recall*) que o sumarizador *BLMSumm* e o *Intellexer* estão em equilíbrio e apresentam os valores mais altos, e o *Copernic* apresentou os valores mais baixos.

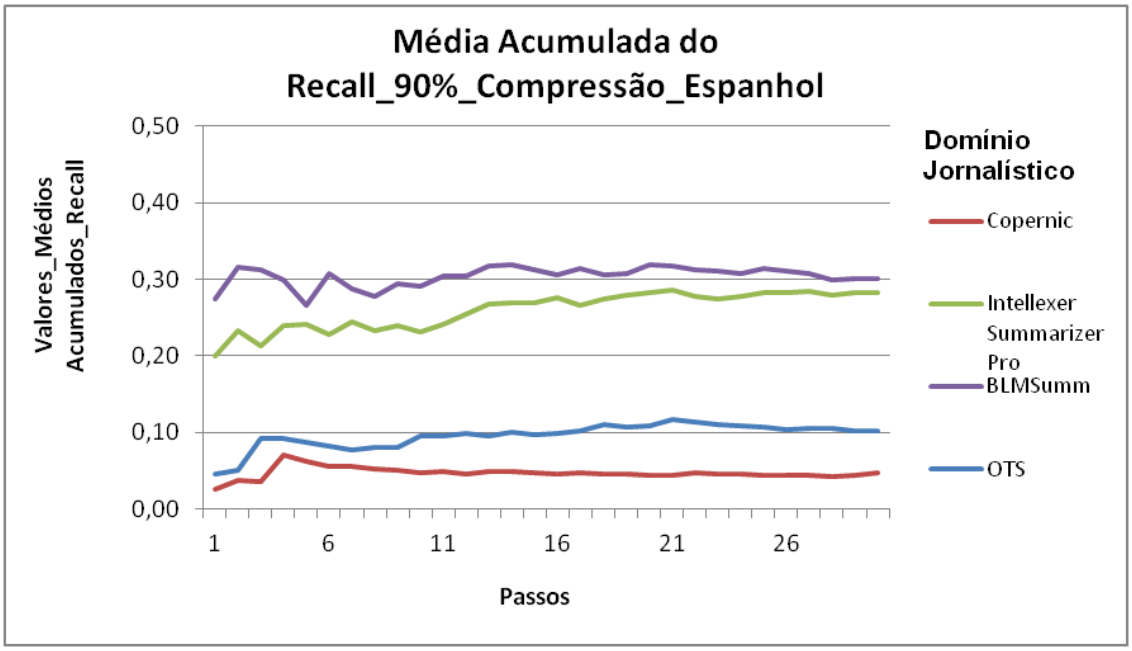

Figura 56. Média acumulada de *Recall* com taxa de compressão de 90%

É possível perceber que na Figura 57 (média acumulada de *Precision*) que o sumarizador *OTS* e *Copernic* apresentaram os valores mais altos, e o *BLMSumm* apresentou os valores mais baixos.

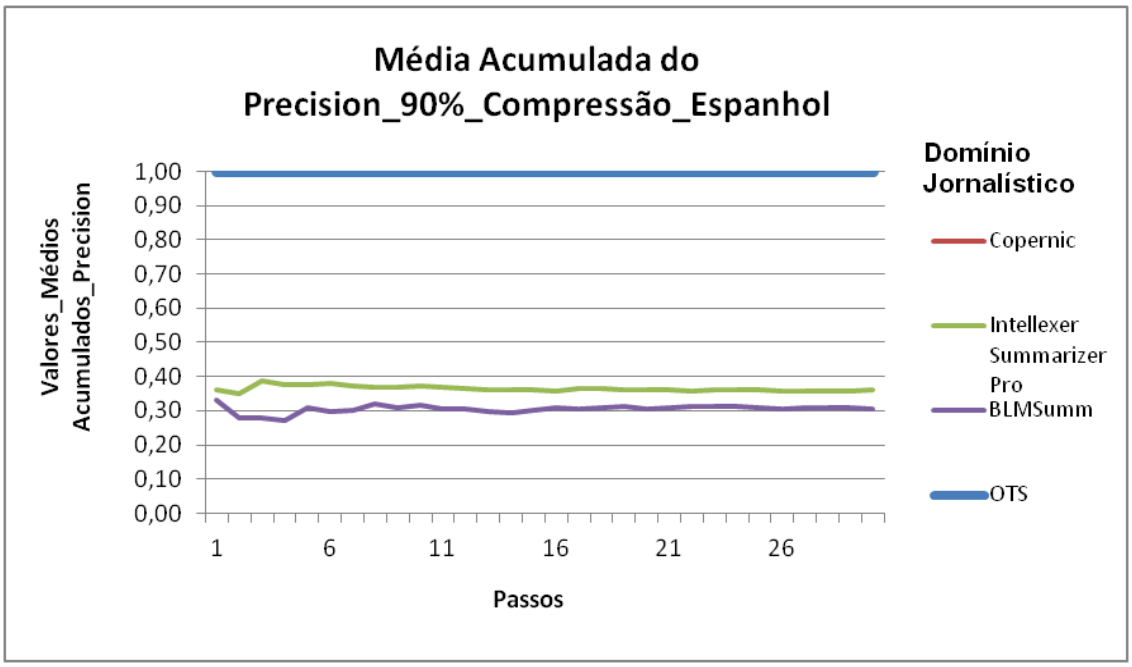

Figura 57. Média acumulada de *Precision* com taxa de compressão de 90%

2. Idioma Italiano

2.1. Taxa de Compressão 50%

Na Figura 58 é possível observar que para o idioma italiano, com taxa de compressão de 50%, houve bastante equilíbrio entre os resultados e o *OTS* apresentou os melhores valores.

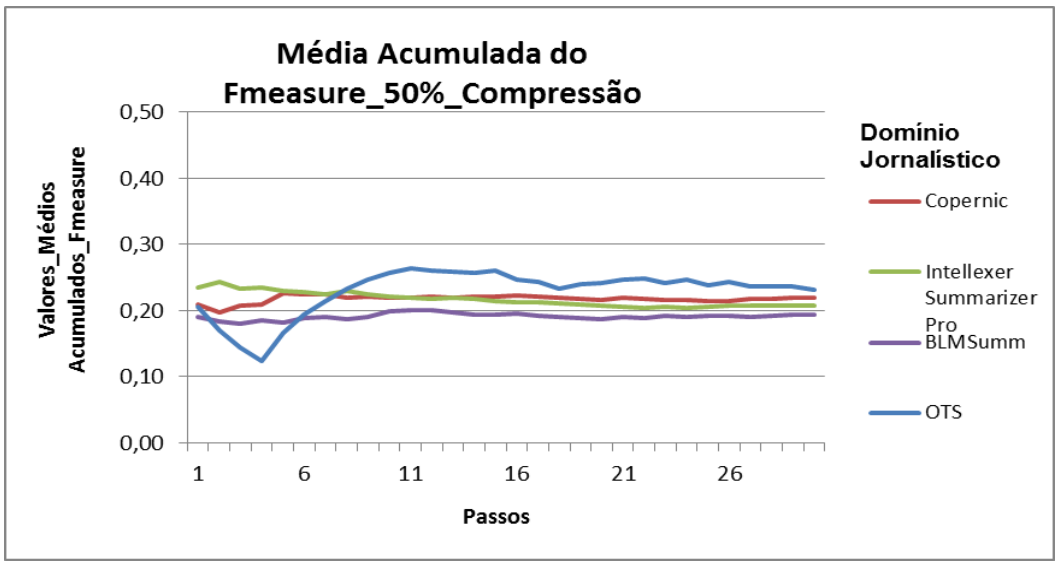

Figura 58. Média acumulada de *F-Measure* com taxa de compressão de 50% do idioma italiano

É possível perceber que na Figura 59 (média acumulada de *Recall*) que o sumarizador *Intellexer* e o *Copernic* estão em equilíbrio e apresentam os valores mais altos, e o *BLMSumm* apresentou os valores mais baixos.

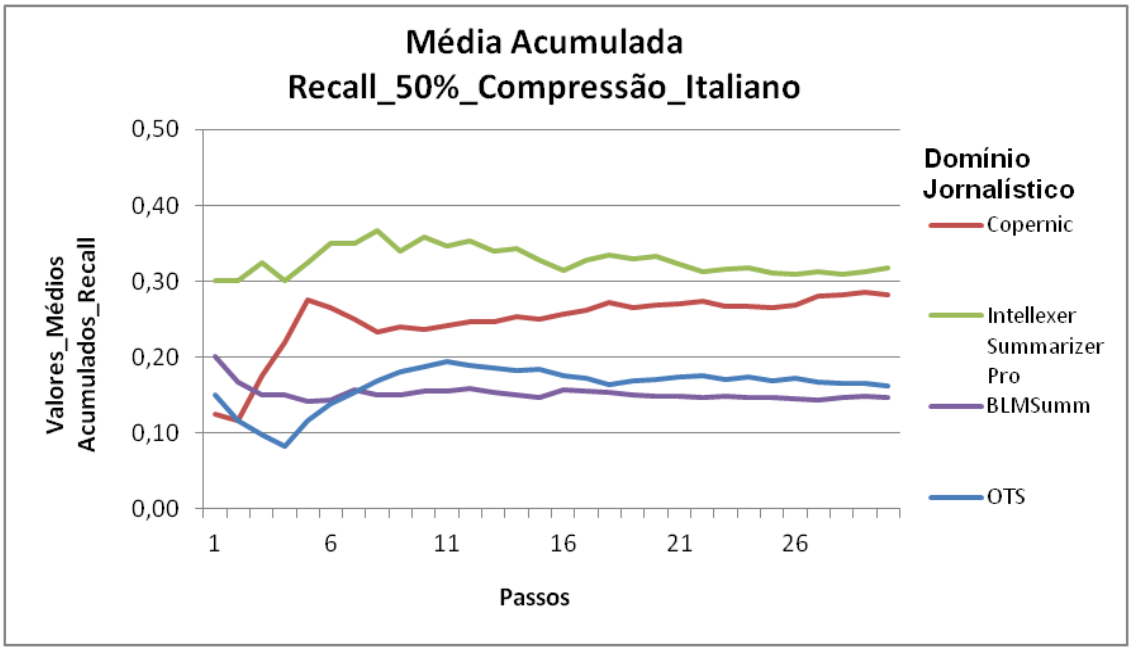

Figura 59. Média acumulada de *Recall* com taxa de compressão de 50%

É possível perceber que na Figura 60 (média acumulada de *Precision*) que o sumarizador OTS apresentou os valores mais satisfatórios, e o *Copernic*, *BLMSumm* e o *Intellexer* apresentaram os valores equilibrados.

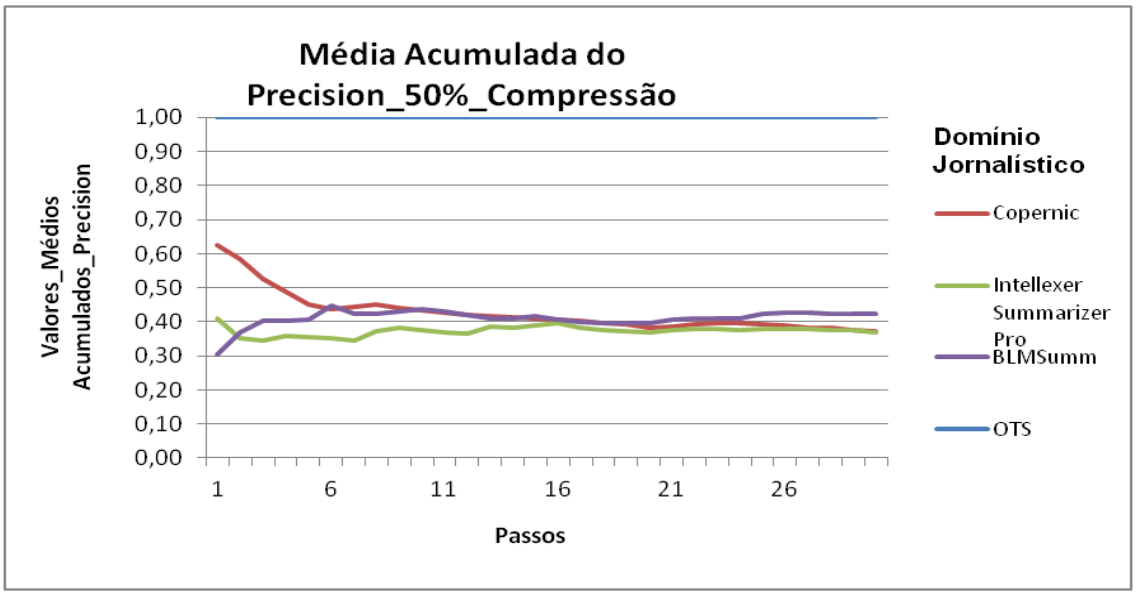

Figura 60. Média acumulada de *Precision* com taxa de compressão de 50%

2.2. Taxa de Compressão 70%

Na Figura 61 observa-se que no idioma italiano com taxa de compressão de 70% o sumarizador *OTS* apresentou o melhor resultado seguido pelo *Copernic*. O *Intellexer* apresentou os valores mais baixos.

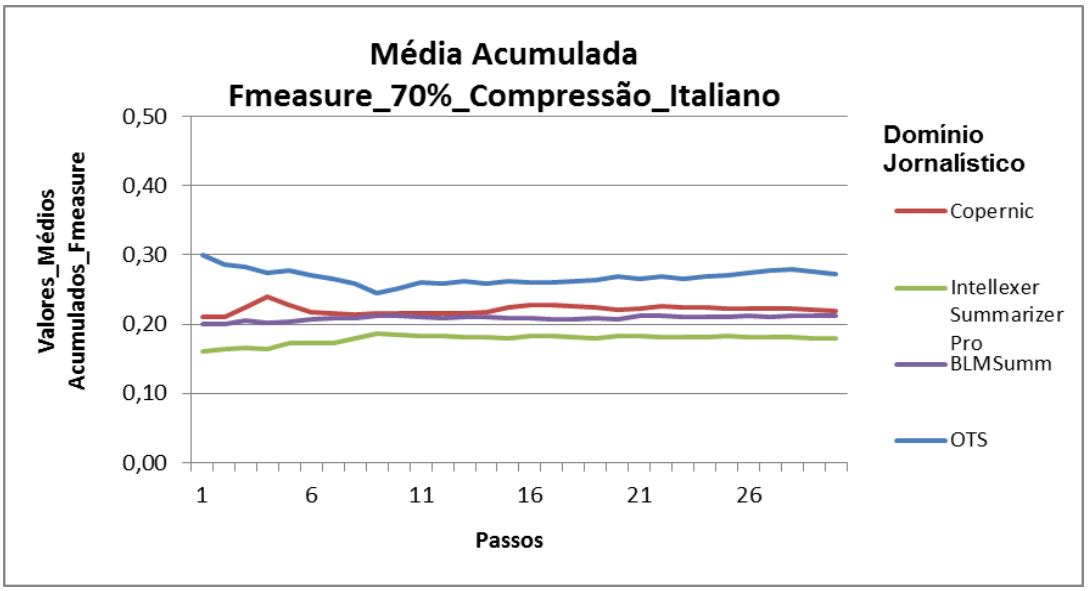

Figura 61. Média acumulada de *F-Measurecom*compressão de 70% do idioma italiano

É possível perceber que na Figura 62 (média acumulada de *Recall*) que o sumarizador *BLMSumm* e o *Copernic* estão em equilíbrio e apresentam os valores mais altos. O *Intellexer* apresentou os valores mais baixos.

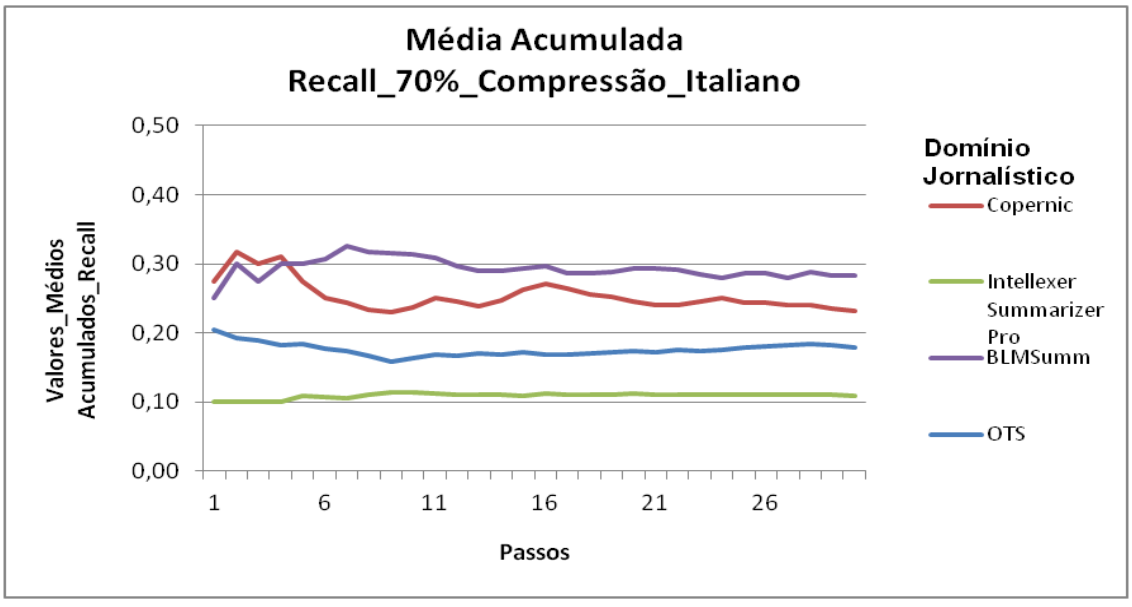

Figura 62. Média acumulada de *Recall* com taxa de compressão de 70%

A Figura 63 (média acumulada de *Precision*) apresentam os resultados dos agrupamentos de texto para cada um dos sumarizadores automáticos. Constata-se que o resultado do sumarizador *OTS* é o mais satisfatório, seguido pelo *Intellexer*, *Copernic* e *BLMSumm*.

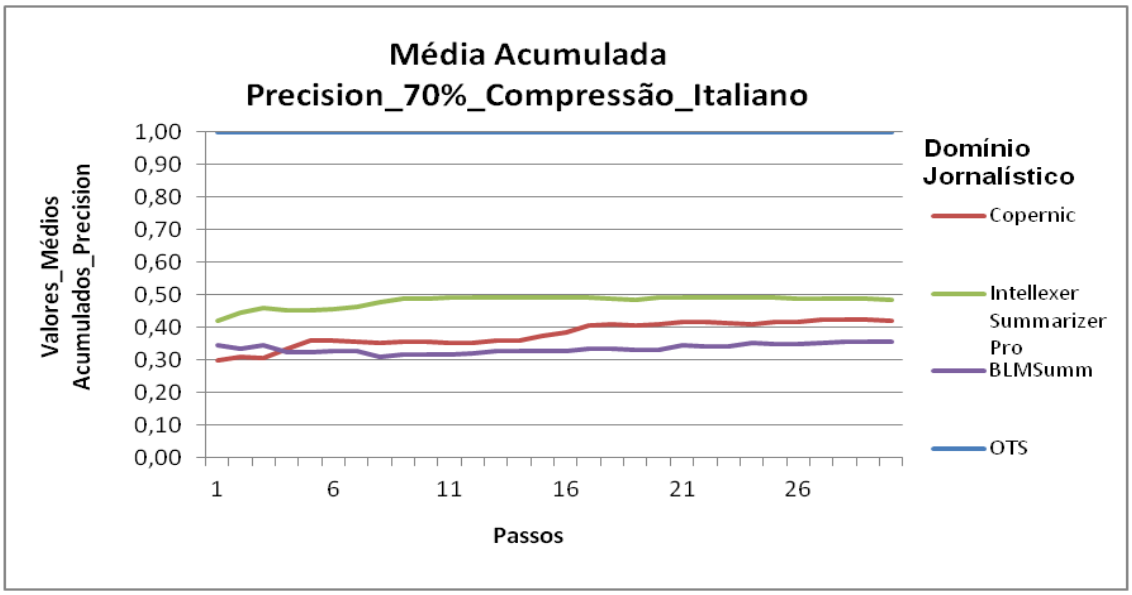

Figura 63. Média acumulada de *Precision* com taxa de compressão de 70%

2.3.Taxa de Compressão 80%

Com uma taxa de compressão de 80%, no idioma italiano, a Figura 64 mostra que o sumarizador que apresentou os melhores valores foi o *OTS*. O *BLMSumm* obteve os resultados mais baixos.

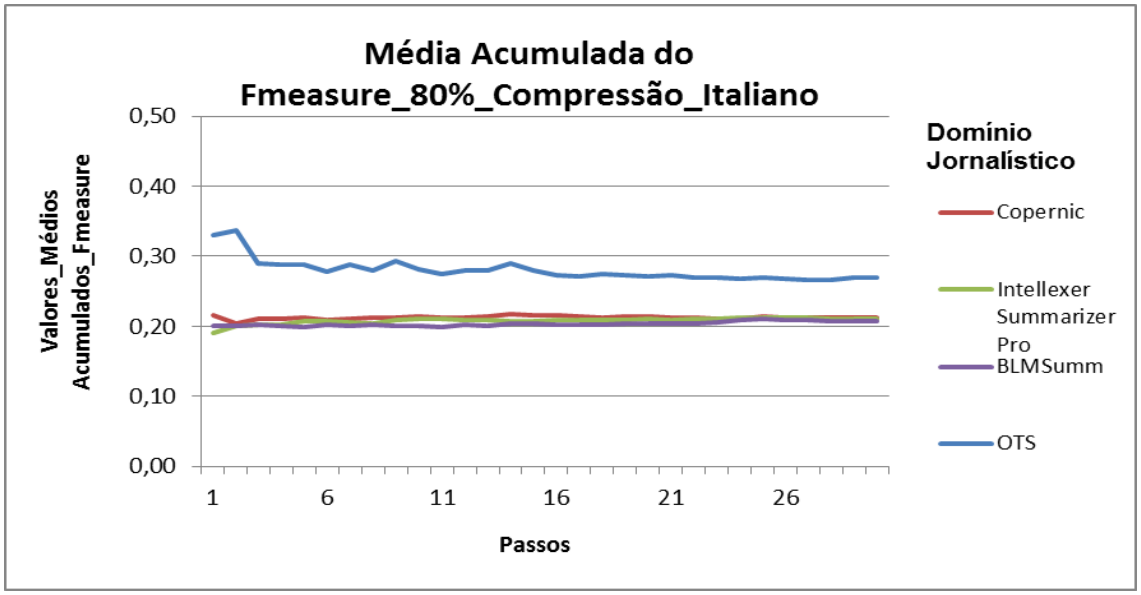

Figura 64. Média acumulada de *F-Measure* com taxa de compressão de 80% do idioma italiano

A Figura 65 (média acumulada de *Recall*) apresentam os resultados dos agrupamentos de texto para cada um dos sumarizadores automáticos. Nota-se que o resultado do *Copernic* é o mais satisfatório. E os valores mais baixos são *BLMSumm* e *OTS*.

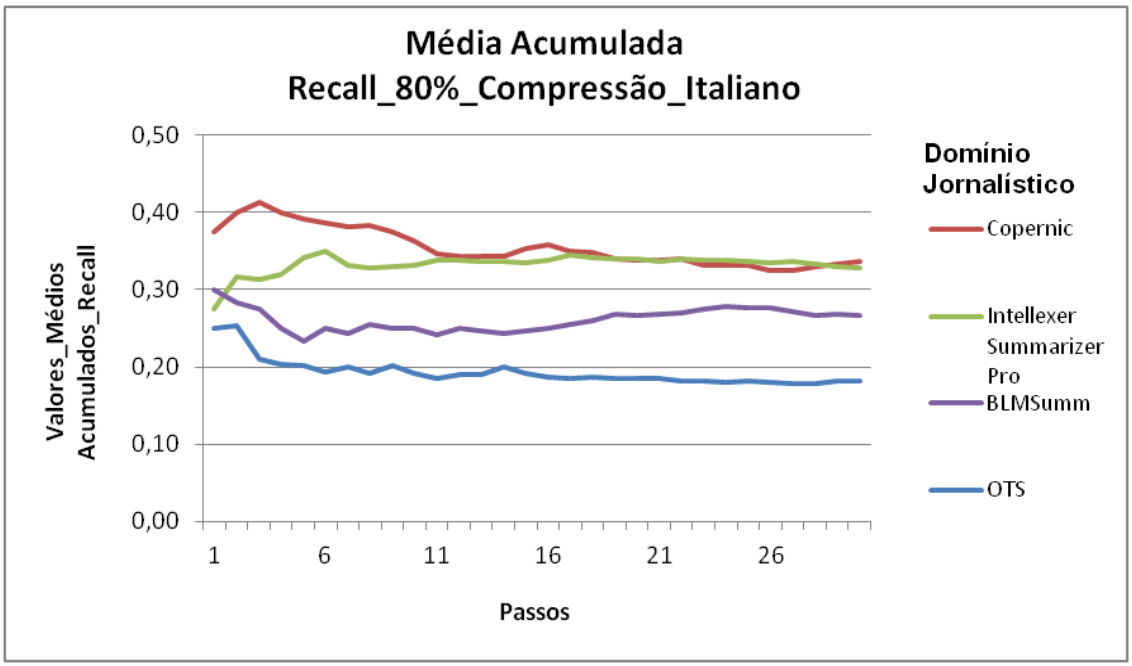

Figura 65. Média acumulada de *Recall* com taxa de compressão de 70%

Na Figura 66 (Média Acumulada de *Precision*) o sumarizador *OTS* apresenta os valores mais satisfatórios. O *BLMSumm*, o *Intellexer* e o *Copernic* apresenta os valores equilibrados.

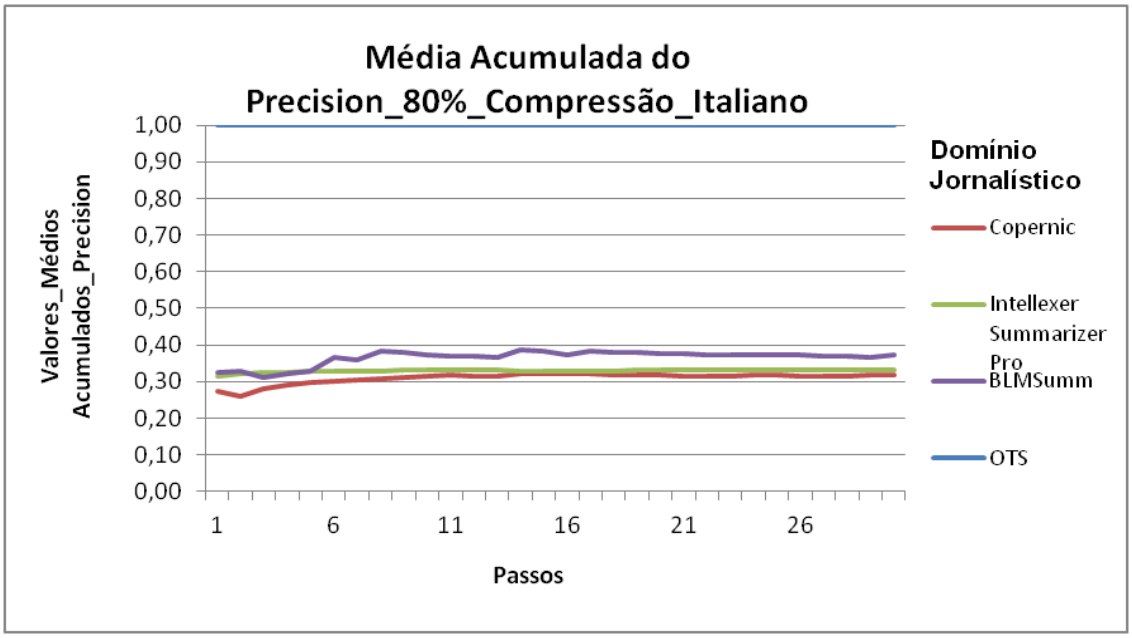

Figura 66. Média acumulada de *Precision* com taxa de compressão de 80%

2.4.Taxa de Compressão 90%

A Figura 67 mostra que o sumarizador automático que apresentou os melhores resultados para uma taxa de compressão de 90% foi o *OTS*, seguido pelo *Copernic*. O *BLMSumm* apresentou os valores mais baixos.

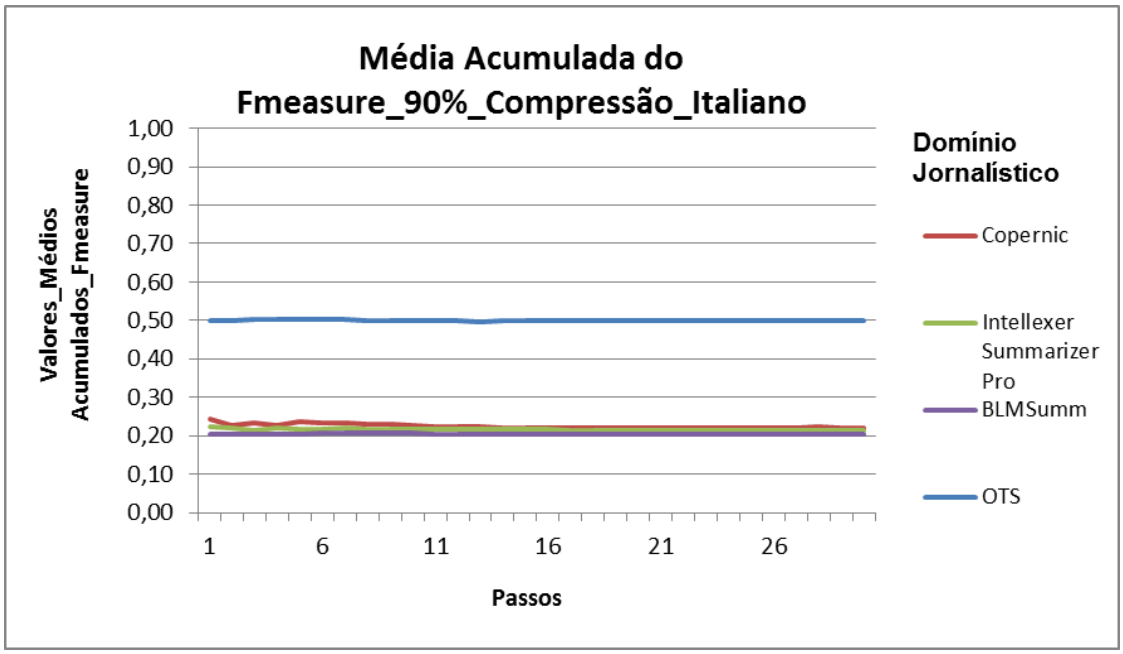

Figura 67. Média acumulada de *F-Measure* com compressão de 90% do idioma italiano

Nota-se na Figura 68 (Média Acumulada *Recall*) que os resultados estão balanceados. É possível perceber que o *BLMSumm* obteve uma ligeira vantagem sobre os outros sumarizadores.

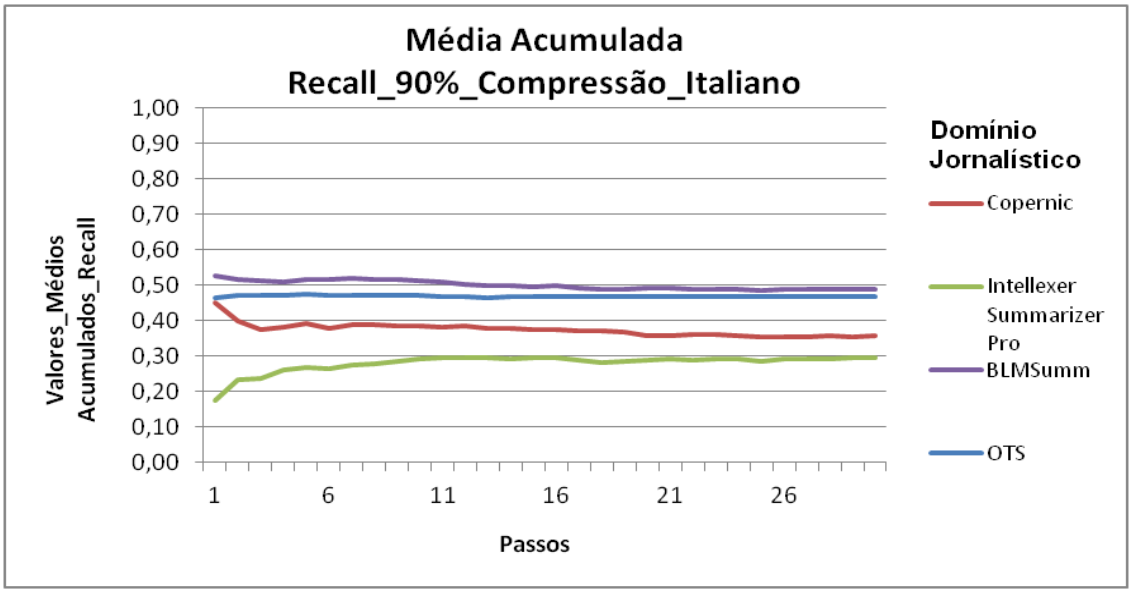

Figura 68. Média acumulada de *Recall* com taxa de compressão de 90% do idioma italiano

É possível perceber que na Figura 69 (média acumulada de *Precision*) que o sumarizador *OTS* apresenta resultados mais significativos. Seguido pelo sumarizador *BLMSumm* e *Intellexer*. O resultado mais baixo pertence ao sumarizador *Copernic*.

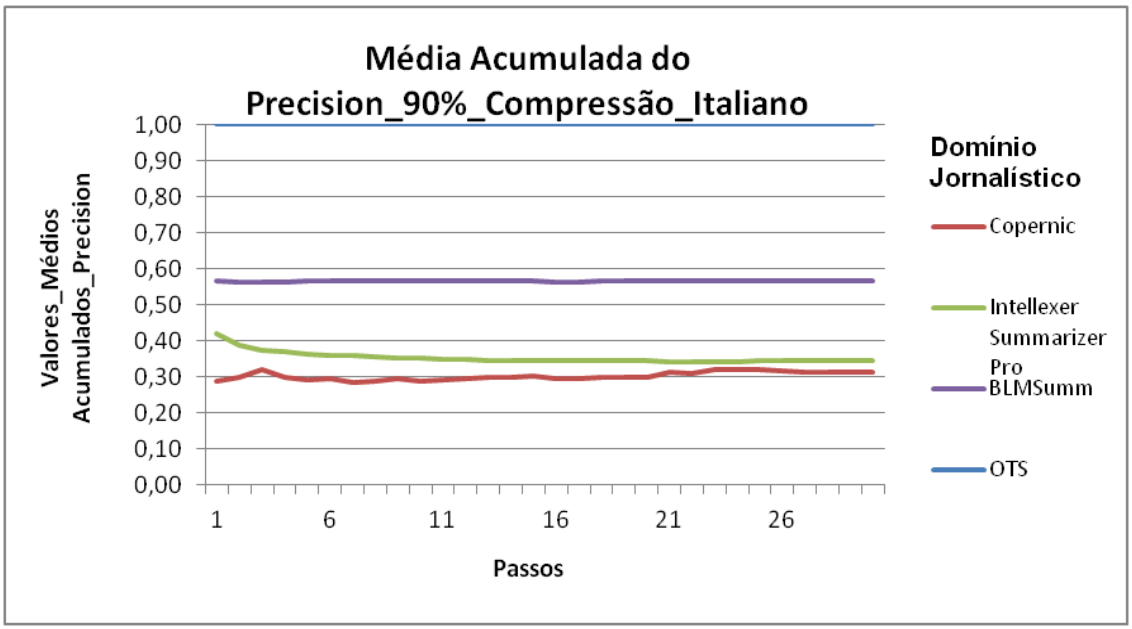

Figura 69. Média acumulada de Precision com taxa de compressão de 90%

APÊNDICE C

APÊNDICE C – SOFTWARE COM OS TESTES ESTATISTICOS

Existe vários *software* estatísticos tais como: *Statistic*, *Statgrafics*, *SPSS*, *Minitab*, *SAS*, *SPHINX*, *WINKS*, entre outros. Entretanto são software geralmente de alto custo e envolvem um aprendizado específico de usabilidade

Neste trabalho foi utilizado para realizar os testes estatísticos dos experimentos de comprovação da hipótese, o *software* Staplus® (http://www.analystsoft.com/en/products/statplus/). Foi utilizada uma versão *Trial*, este software foi escolhido porque contém os testes estatísticos adotados neste trabalho, são eles ANOVA de Friedman e Coeficiente de Concordância de Kendall.

As tabelas apresentadas neste apêndice são dividas por idioma e subdividas em taxas de compressão. Todas as tabelas comprovam a hipótese alternativa e rejeitam a hipótese nula, em virtude de terem obtidos um resultado de ordem médio e coeficiente de concordância de Kendall superior a 0,5.

1. Idioma Espanhol

1.1. Taxa de Compressão 50%

Tabela 18. Teste Estatístico da métrica Coeficiente *Silhouette* – Idioma Espanhol com Compressão de 50%

Comparando amostras múltiplas relacionadas			
N	30	*Graus de liberdade*	3
qui-quadrado	90	*p-nível*	0
Coef. de concordância de Kendall	1	*Ordem médio*	1
	Ordem médio	**Soma de ordens**	**Média**
Copernic	2	60	0,2692
Intellexer Summarizer Pro	1	30	0,253
BLMSumm	3	90	0,2919
OTS	4	120	0,9774

Tabela 19. Teste Estatístico da métrica *F-Measure* – Idioma Espanhol com Compressão de 50%

Comparando amostras múltiplas relacionadas			
N	30	*Graus de liberdade*	3
qui-quadrado	87,76	*p-nível*	0
Coef. de concordância de Kendall	0,9751	*Ordem médio*	0,9743
	Ordem médio	**Soma de ordens**	**Média**
Copernic	1	30	0,0467
Intellexer Summarizer Pro	2,9333	88	0,0919
BLMSumm	4	120	0,254
OTS	2,0667	62	0,0831

1.2.Taxa de Compressão 70%

Tabela 20. Teste Estatístico da métrica Coeficiente *Silhouette* – Idioma Espanhol com Compressão de 70%

Comparando amostras múltiplas relacionadas			
N	30	*Graus de liberdade*	3
qui-quadrado	87,4214	*p-nível*	0
Coef. de concordância de Kendall	0,9713	*Ordem médio*	0,9704
	Ordem médio	**Soma de ordens**	**Média**
Copernic	3,05	91,5	0,9684
Intellexer Summarizer Pro	3,9333	118	0,9707
BLMSumm	1	30	0,963
OTS	2,0167	60,5	0,9672

Tabela 21. Teste Estatístico da métrica *F-Measure* – Idioma Espanhol com Compressão de 70%

Comparando amostras múltiplas relacionadas			
N	30	*Graus de liberdade*	3
qui-quadrado	89,709	*p-nível*	0
Coef. de concordância de Kendall	0,9968	*Ordem médio*	0,9967
	Ordem médio	**Soma de ordens**	**Média**
Copernic	1,0167	30,5	0,0467
Intellexer Summarizer Pro	3	90	0,0919
BLMSumm	4	120	0,1832
OTS	1,9833	59,5	0,0553

1.3.Taxa de Compressão 80%

Tabela 22. Teste Estatístico da métrica Coeficiente *Silhouette* – Idioma Espanhol com Compressão de 80%

Comparando amostras múltiplas relacionadas			
N	30	*Graus de liberdade*	3
qui-quadrado	90	*p-nível*	0
Coef. de concordância de Kendall	1	*Ordem médio*	1
	Ordem médio	**Soma de ordens**	**Média**
Copernic	3	90	0,9411
Intellexer Summarizer Pro	4	120	0,9586
BLMSumm	2	60	0,8907
OTS	1	30	0,8797

Tabela 23. Teste Estatístico da métrica *F-Measure* – Idioma Espanhol com Compressão de 80%

Comparando amostras múltiplas relacionadas			
N	30	*Graus de liberdade*	3
qui-quadrado	83,56	*p-nível*	0
Coef. de concordância de Kendall	0,9284	*Ordem médio*	0,926
	Ordem médio	**Soma de ordens**	**Média**
Copernic	1,7667	53	0,077
Intellexer Summarizer Pro	1,2333	37	0,0648
BLMSumm	3	90	0,1966
OTS	4	120	0,3749

1.4.Taxa de Compressão 90%

Tabela 24. Teste Estatístico da métrica Coeficiente *Silhouette* – Idioma Espanhol com Compressão de 90%

Comparando amostras múltiplas relacionadas			
N	30	*Graus de liberdade*	3
qui-quadrado	90	*p-nível*	0
Coef. de concordância de Kendall	1	*Ordem médio*	1
	Ordem médio	**Soma de ordens**	**Média**
Copernic	1	30	0,0816
Intellexer Summarizer Pro	3	90	0,2109
BLMSumm	4	120	0,2533
OTS	2	60	0,1587

Tabela 25. Teste Estatístico da métrica *F-Measure* – Idioma Espanhol com Compressão de 90%

Comparando amostras múltiplas relacionadas			
N	30	*Graus de liberdade*	3
qui-quadrado	90	*p-nível*	0
Coef. de concordância de Kendall	1	*Ordem médio*	1
	Ordem médio	**Soma de ordens**	**Média**
Copernic	3	90	0,9077
Intellexer Summarizer Pro	1	30	0,8114
BLMSumm	2	60	0,839
OTS	4	120	0,9129

2. **Idioma Italiano**

2.1.Taxa de Compressão 50%

Tabela 26. Teste Estatístico da métrica Coeficiente *Silhouette* – Idioma Espanhol com Compressão de 50%

Comparando amostras múltiplas relacionadas			
N	30	*Graus de liberdade*	3
qui-quadrado	84,68	*p-nível*	0
Coef. de concordância de Kendall	0,9409	*Ordem médio*	0,9389
	Ordem médio	**Soma de ordens**	**Média**
Copernic	3,9667	119	0,9374
Intellexer Summarizer Pro	1,8667	56	0,9309
BLMSumm	1,1333	34	0,9269
OTS	3,0333	91	0,936

Tabela 27. Teste Estatístico da métrica F-Measure – Idioma Espanhol com Compressão de 50%

Comparando amostras múltiplas relacionadas			
N	30	*Graus de liberdade*	3
qui-quadrado	51,16	*p-nível*	0
Coef. de concordância de Kendall	0,5684	*Ordem médio*	0,5536
	Ordem médio	**Soma de ordens**	**Média**
Copernic	2,9	87	0,2178
Intellexer Summarizer Pro	2,5667	77	0,217
BLMSumm	1,1333	34	0,1911
OTS	3,4	102	0,2289

2.2.Taxa de Compressão 70%

Tabela 28. Teste Estatístico da métrica Coeficiente *Silhouette* – Idioma Espanhol com Compressão de 70%

Comparando amostras múltiplas relacionadas			
N	30	*Graus de liberdade*	3
qui-quadrado	90	*p-nível*	0
Coef. de concordância de Kendall	1	*Ordem médio*	1
	Ordem médio	**Soma de ordens**	**Média**
Copernic	3	90	0,9036
Intellexer Summarizer Pro	2	60	0,8948
BLMSumm	1	30	0,884
OTS	4	120	0,912

Tabela 29. Teste Estatístico da métrica *F-Measure* – Idioma Espanhol com Compressão de 70%

Comparando amostras múltiplas relacionadas			
N	30	*Graus de liberdade*	3
qui-quadrado	90	*p-nível*	0
Coef. de concordância de Kendall	1	*Ordem médio*	1
	Ordem médio	**Soma de ordens**	**Média**
Copernic	3	90	0,221
Intellexer Summarizer Pro	1	30	0,1784
BLMSumm	2	60	0,2088
OTS	4	120	0,2682

2.3.Taxa de Compressão 80%

Tabela 30. Teste Estatístico da métrica *F-Measure* – Idioma Espanhol com Compressão de 80%

Comparando amostras múltiplas relacionadas			
N	30	*Graus de liberdade*	3
qui-quadrado	88,84	*p-nível*	0
Coef. de concordância de Kendall	0,9871	*Ordem médio*	0,9867
	Ordem médio	**Soma de ordens**	**Média**
Copernic	3,9667	119	0,877
Intellexer Summarizer Pro	1	30	0,8524
BLMSumm	2	60	0,8582
OTS	3,0333	91	0,874

Tabela 31. Teste Estatístico da métrica *F-Measure* – Idioma Espanhol com Compressão de 80%

Comparando amostras múltiplas relacionadas			
N	30	*Graus de liberdade*	3
qui-quadrado	84,6162	*p-nível*	0
Coef. de concordância de Kendall	0,9402	*Ordem médio*	0,9381
	Ordem médio	**Soma de ordens**	**Média**
Copernic	2,8833	86,5	0,2124
Intellexer Summarizer Pro	2,05	61,5	0,2081
BLMSumm	1,0667	32	0,2036
OTS	4	120	0,2804

2.4.Taxa de Compressão 90%

Tabela 32. Teste Estatístico da métrica Coeficiente *Silhouette* – Idioma Espanhol com Compressão de 90%

Comparando amostras múltiplas relacionadas			
N	30	*Graus de liberdade*	3
qui-quadrado	90	*p-nível*	0
Coef. de concordância de Kendall	1	*Ordem médio*	1
	Ordem médio	**Soma de ordens**	**Média**
Copernic	2	60	0,7842
Intellexer Summarizer Pro	3	90	0,8105
BLMSumm	4	120	0,9054
OTS	1	30	0,7486

Tabela 33. Teste Estatístico da métrica *F-Measure* – Idioma Espanhol com Compressão de 90%

Comparando amostras múltiplas relacionadas			
N	30	*Graus de liberdade*	3
qui-quadrado	90	*p-nível*	0
Coef. de concordância de Kendall	1	*Ordem médio*	1
	Ordem médio	**Soma de ordens**	**Média**
Copernic	3	90	0,225
Intellexer Summarizer Pro	2	60	0,2161
BLMSumm	1	30	0,2042
OTS	4	120	0,5002

ANEXOS

ANEXO A
ESCOLHA DA TÉCNICA TESTE ESTATÍSTICO A PARTIR DO NÚMERO DE AMOSTRAS

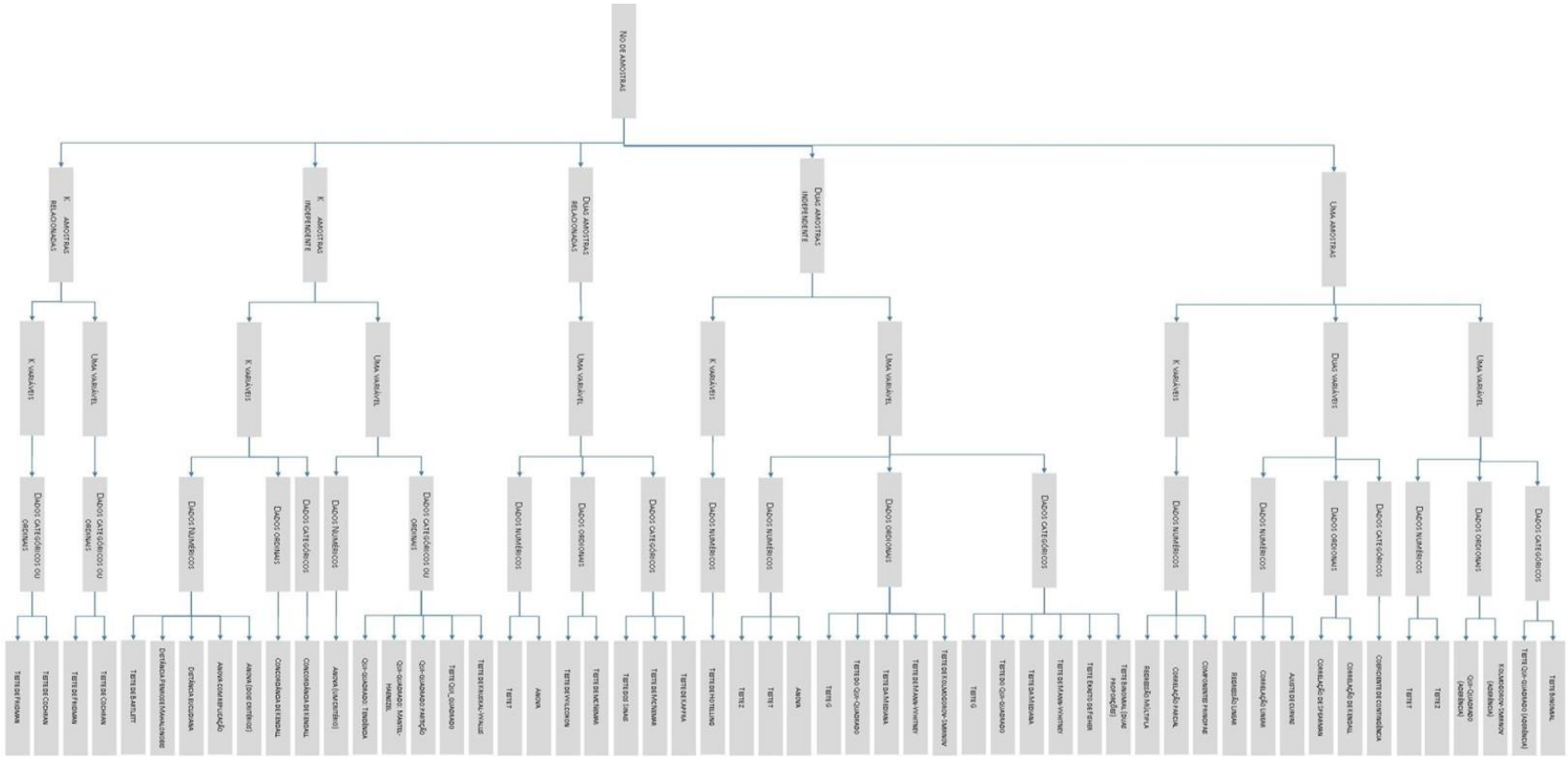

Figura 70. Diagrama para escolha da técnica teste estatístico a partir do número de amostras (CALLEGARI-JACQUES, 2007).