Todas as Palavras da Sentença como Métrica para um Sumarizador Automático

Marcus V. C. Guelpeli
Departamento de Ciência da Computação
Universidade Federal Fluminense - UFF
Rua Passo da Pátria 156 - Bloco E
3º andar São Domingos - Niterói - RJ
CEP: 24210-240
mguelpeli@ic.uff.br

Flavia Cristina Bernardini
Laboratório de Documentação Ativa e Design Inteligente
Rua Passo da Pátria 156 - Bloco E
3º andar São Domingos - Niterói - RJ
CEP: 24210-240
flavia@addlabs.uff.br

Ana Cristina Bicharra Garcia
Departamento de Ciência da Computação
Universidade Federal Fluminense - UFF
Rua Passo da Pátria 156 - Bloco E
3º andar São Domingos - Niterói - RJ
CEP: 24210-240
bicharra@ic.uff.br

ABSTRACT
The purpose of this work is to present an automatic summarizer that uses as a metric the number of words into a sentence to define the text author's pragmatic profile. Using the number of words as a metric, the original text is classified according to its temporal measures and textual composition, which is based on its formality. Also, these features are parameters to the summary generation that indicate the compression level. This work uses traditional methodologies of automatic summarization and compares these results to results obtained with our proposal.

Key-Words
Automatic summarizer, pragmatic profile, automatic compression.

RESUMO
Este trabalho tem como meta apresentar um sumarizador automático que usa como métrica a quantidade de palavras dentro de uma sentença para definir o perfil pragmático do autor do texto. O trabalho usa metodologias tradicionais da área de sumarização automática e as compara com os resultados deste trabalho. Com o uso da palavra como métrica cria-se uma classificação no texto original em relação as suas medidas temporais e composição textual mediante a sua formalidade, criando assim parâmetros para determinar o nível de compressão para geração do sumário.

Palavras-Chave
Sumarizador Automático, Perfis Pragmáticos, Palavras, Compressão automática.

Categories and Subject Descriptors
I.2.7 [**Artificial Intelligence**]: Natural Language Processing– *Text analysis.*

General Terms
Human Factors, Measurement, Experimentation and Theory.

1. INTRODUÇÃO
Um sumário é um resumo que tem o objetivo de captar a idéia principal de determinado autor e representar esta em poucas linhas. [7] classifica os sumários em indicativos, informativos e sumários de críticas. Os sumários indicativos podem servir como indexadores, onde se descobre qual é a idéia do autor e havendo interesse sobre o tema, busca-se a leitura do texto original com maior riqueza de detalhes. Os sumários informativos são autocontidos, ou seja, detêm informações suficientes, não havendo a necessidade de remeter-se ao texto original. Os sumários de críticas avaliam e comparam o texto original com trabalhos relacionados a mesma temática[7].

Os processos automáticos de sumarização podem obter sumários (abordagem profunda) e extratos (abordagem superficial). Os sumários alteram o conteúdo e/ou as estruturas das frases originais do autor, juntando-se e reescrevendo-as com a finalidade de generalizar ou especificar as informações. Já os extratos seguem transposições das frases dos textos originais, e por algum tipo de método, escolhem as frases com maior relevância no texto e as colocam no extrato.

Segundo [15] os trabalhos recentes estão adotando metodologias híbridas, ou seja, o uso da abordagem superficial e profunda, variando os métodos de cada uma. Os trabalhos de sumarização automática enfatizam nos seus algoritmos a exclusão das stopwords para que possam realizar a fase da redução, como por exemplo, o algoritmo TextTiling usado por [5], [6], [14] e [11]. Neste trabalho todas as palavras serão mantidas como forma de preservar a idéia do autor.

A hipótese deste trabalho está fundamentada na palavra como métrica gramatical, onde agrega todas as formas e variações de palavras e suas ocorrências nas sentenças para realizar a sumarização. O trabalho ressalta a importância da palavra para mensurar o conhecimento gramatical do interlocultor, podendo assim classificar o texto e produzir o sumário mais próximo do perfil gramatical de cada usuário.

Este trabalho está organizado como segue. A Seção 2 a aborda adaptação do sumarizador ao perfil do usuário baseado na palavra escrita, usando as regras de Hovy. Seção 3 apresenta a metodologia, a simulação e o corpus utilizado. Na Seção 4 discute

os resultados obtidos com o Sumarizador Automático. Finalmente, a Seção 5 apresenta as conclusões, vantagens e desvantagens do método proposto e sugere trabalhos futuros.

2. ADAPTAÇÃO DO SUMARIZADOR AO PERFIL DO USUÁRIO

Um modelo do usuário é uma representação explícita de propriedades de um interlocultor em particular, que permite que um sistema adapte diversos aspectos de seu desempenho e de suas funcionalidades às necessidades individuais deste usuário. A necessidade de adequar o sumário ao perfil do interlocultor já é estudada desde o início da área de Sumarização Automática (SA). [12] propõe no seu trabalho, os métodos de identificação de segmentos relevantes calculando a significância de cada sentença em um texto-fonte por seu peso e, então, selecionam aquelas com maior peso (acima de um limite mínimo) para compor o extrato, incorporando os parâmetros clássicos para sua identificação e seleção, este tipo de metodologia é denominada hoje como abordagem estatística ou superficial na sumarização.[12] aborda também a importância de atribuir valores maiores as frases que tenham palavras e que pertençam ao âmbito de interesse do usuário.

[8] propõe, na abordagem profunda o uso de perfis pragmáticos, ou seja, os tamanhos dos sumários serão influenciados diretamente pelos objetivos do usuário. [8] estabelece algumas métricas denominadas pelo autor como características de estilo. Ele estabelece uma relação temporal na preparação do texto original e classifica como escasso, pouco, suficiente ou ilimitado. O autor também aborda o tipo de escrita textual, onde se baseia no uso de algumas regras para classificar textos quanto ao grau de formalidade, classificando os como coloquial, normal ou formal.

Tabela 1. Representa a interação das regras de estilo e implicações no conteúdo do sumário segundo[8].

Tempo / Formalidade	Escasso	Pouco	Suficiente	Ilimitado
Coloquial	Sumarização alta apenas tópico principal	Sumarização média	Sumarização média tópico principal, detalhes desconhecidos	Sumariza - ção baixa tópico principal, detalhes relevantes
Normal	Sumarização média tópico principal, poucos detalhes	Sumarização média tópico principal	Sumarização média tópico principal, detalhes importantes	Sumariza - ção baixa tópico principal, detalhes relevantes
Formal	Sumarização média tópico principal.	Sumarização média tópico principal	Sumarização baixa tópico principal, detalhes importantes	Sumariza- ção baixa tópico principal e correlatos, detalhes relevantes

Na Tabela 1, os textos com características formais tendem a ter frases mais longas, ou seja, com o uso de um número maior de palavras, enquanto textos mais coloquiais tendem a ter um número menor de palavras [8].

Em trabalhos mais recentes [9] e [10] usa a questão da compressão em resumos para sentenças que compartilham de um certo grau sobreposição de informações. Neste trabalho a idéia da compressão é automática. Ela seria usada para garantir que o tamanho do sumário seria coerente com o grau de formalidade do texto, tendo, como conseqüência direta a determinação automática do tamanho de compressão do sumário.

3. METODOLOGIA

O sumarizador proposto usa em sua forma de composição de sumários a extração e a transposição das sentenças, respeitando a sua posição no texto original, que é característica da abordagem superficial, mas adota também a abordagem profunda quando usa as regras de [8] para classificar o texto original, baseando-se no perfil pragmático do usuário de acordo com a Figura 1. O sumarizador usa o texto-fonte obtido do Corpus Temário, onde é utilizada a taxonomia quanto à sua formalidade e temporalidade (fase de análise) baseada nas regras de estilo de [8]. Com isso aplica-se o algoritmo Perfil que determina o grau de compressão automático baseado no grau de formalidade e temporalidade do texto que será usado para obter o sumário, refletindo assim o perfil pragmático do usuário (fase de redução), salientando não existência de qualquer tipo de interferência humana. O sumário realiza a extração e transposição das sentenças, respeitando a sua posição no texto original compondo o sumário com as frases com maior freqüência de palavras determinada pelo algoritmo perfil (fase síntese).

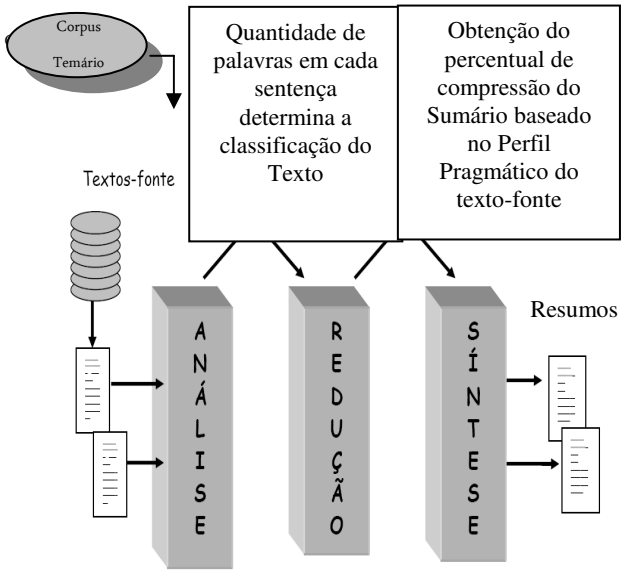

Figura 1. Estrutura do Sumarizador Automático usando algoritmo Perfil.

Outro fato a destacar do trabalho é manutenção das "stopwords" usadas na fase de análise. A retirada das "stopwords" do texto serve para diminuir o volume de processamento. Segundo estudos de [19] estas palavras são relevantes, porém não existem estudos

comparativos que comprove tal eficiência ou não. Neste trabalho as stopwors são mantidas seguindo a proposta original deste trabalho que é manter todas as palavras da sentença como métrica qualitativa na composição do sumário, sendo assim, baseando-se nos trabalhos: [13], [4] e [1] que enfatizam a importância da palavra para representação do conhecimento e evolução sintática e semântica do indivíduo, medido pela M.L.U.-m e depois adaptada pela M.L.U.-w. Esta métrica representa a evolução gramatical, o número de vocábulos do indivíduo. Este indicador acompanha o aumento do comprimento do texto, reflete sua evolução medida em palavras. Baseado nas indicações teóricas acredita-se que as denominadas "stopwords" sejam representativas da evolução gramatical do indivíduo, sendo imprescindíveis para a determinação do conhecimento lexical pragmático abordado no uso da língua em diferentes contextos e a importância da palavra em toda sua forma gramatical. O estudo comparativo realizado neste trabalho usa a metodologia que adota as "stopwords" (algoritmo TextTiling) e o algoritmo Perfil [3], que não adota a retirada destas palavras.

4. EXPERIMENTOS

Seguindo a estratégia do uso de extratos ideais, este trabalho adotou o corpus Temário, [16] e [17] para realizar os testes dos sumários gerados. Este corpus é composto por 100 textos, que são classificados por Sumários e Textos Fontes. Os textos fontes são divididos por texto fonte com origem e título, texto fonte com título e texto fonte sem título. Já os sumários são classificados por: Extratos Ideais Automáticos, Sumários Manuais e Sumários Manuais Marcados. No diretório Texto fonte com origem e título encontra-se uma subdivisão com textos de dois jornais de grande circulação no Brasil Folha de São Paulo e Jornal do Brasil. No diretório da Folha de São Paulo nos textos fontes da seção Especial Mundo e Opinião. Já na seção do Jornal do Brasil encontram-se os textos fontes classificados por Internacional e Política. Para realizar os experimentos foram usadas as seções grifadas conforme exibido na Figura 2.

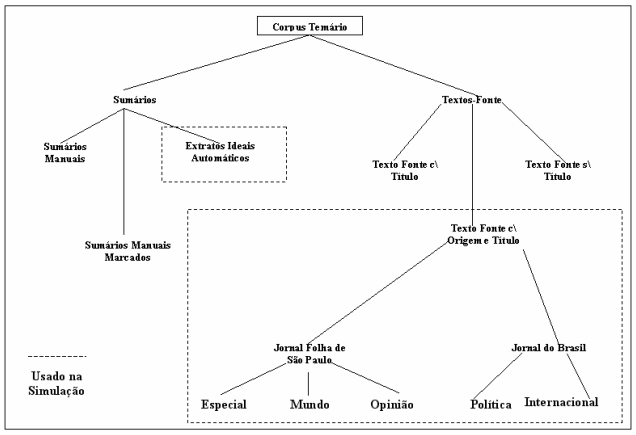

Figura 2. Taxonomia do Corpus Temário.

4.1 Resultados Obtidos

Este trabalho analisa o comportamento Sumarizador Automático - SA usando o algoritmo TextTiling formalizada por [20] e o comportamento do SA utilizando o algoritmo Perfil, que é proposto neste trabalho. Estes valores foram comparados com o extrato ideal, que é o melhor tipo de sumário de referência para a avaliação de sistemas de SA, pois, por conter somente sentenças do texto-fonte, pode ser comparado mais facilmente com um sumário automático, já que este também se origina do mesmo texto-fonte. No corpus Temário o extrato ideal foi composto pelas sentenças do texto-fonte julgadas por humanos como essenciais para compor um sumário do texto. Foram utilizadas as seguintes medidas de avaliação, nas quais são considerados os aspectos qualitativos e o conteúdo de um sumário:

- *Recall*(R): número de sentenças do sumário automático presente no sumário de referência / número de sentenças do sumário automático;
- *Precision*(P): número de sentenças do sumário automático presente no sumário de referência / número de sentenças do sumário de referência;
- *Medida-F*(F-Measure): $\frac{PxR}{P+R}x2$, a medida F combina as métricas de *recall* e de *precicion*. O resultado da medida F é um indicativo de que, quanto mais próximo de 1, melhor é o sumário e resultados mais próximos de 0, demonstra que os sumários são ruins, ou seja, longe do proposto pelo extrato ideal.

Outra medida usada para avaliação foi compressão(C_r – compression rate) e o número de sentenças do texto original mantidas no sumário(NS_{org}), C_r é dado pelo número de sentenças eliminadas para compor o sumário e NS_{org} é dada pelo número de sentenças do texto original mantidas no sumário. C_r varia de 0 e 1, quanto menor o valor de C_r maior é o sumário- o desejado é que seja mais próximo de 1. Na Tabela 2 são representados os resultados das médias e do desvio padrão obtido nas aplicações das medidas em todas as seções do corpus.

Observa-se que os resultados obtidos no algoritmo Perfil são significativamente na média melhores em todas as seções e usando qualquer métrica.

Para validar as médias obtidas na tabela 2 usou-se a análise de variância, que é o método estatístico usado para testar diferenças entre médias de amostras e entre combinações lineares das médias. É assim chamada pelo fato das variâncias serem usadas para medir diferenças entre médias. Classicamente, o teste-t é a ferramenta utilizada para investigar diferenças entre duas médias sendo que a análise de variância também é utilizada para esta finalidade. Os parâmetros da tabela 3 foram analisados pelo teste "t" de Student, considerando variâncias diferentes. As duas amostras (independentes) foram coletadas na simulação.

Este trabalho adotou a hipótese nula $H_0 = \sigma_1^2 = \sigma_2^2$ onde as duas variâncias são iguais, ou seja, $H_1 = \sigma_1^2 \neq \sigma_2^2$ tendo como hipótese alternativa .

O resultado obtido na Tabela 3 mostra que à hipótese nula foi aceita nas seções Internacional usando a medida precision, na seção Mundo usando as medidas C_R e NS_{orig} e na seção Especial usando as medidas C_R e NS_{orig}, sendo assim não houve diferença significativa entre as variâncias, determinando assim a igualdade estatística entre suas médias. Este resultado foi obtido no grau de significância de 5% .

Nos outros casos, o que equivale 80% dos resultados representados na Tabela 3, a hipótese nula foi rejeitada, sendo

assim prevaleceu a hipótese alternativa onde as variâncias são diferentes, estabelecendo assim uma diferença significativa entre as médias obtidas pelos algoritmos TextTiling e Perfil, neste caso o grau de significância também foi 5%.

Pode-se observar ainda que hipótese nula foi rejeitada no grau de significância de 1% nas seguintes seções, com suas respectivas medidas: Opinião(Recall, F-Mesure e C_R), Política(Recall, *Precision* e *F-Mesure*), Especial(*Recall* e *F-Mesure*) e Internacional(*Recall*).

5. CONCLUSÃO

Neste trabalho foi proposta uma métrica que usa toda a palavra dentro da sentença para classificar os textos dentro do perfil pragmático, baseado nas Regras de [8]. Os resultados obtidos são animadores, pois foram comparados com a metodologia amplamente conhecida na literatura e os resultados foram animadores em relação a metodologia proposta. Mantendo todas as palavras dentro da sentença, este trabalho preserva as *stopwords*, ao contrário dos trabalhos desta área que removem as *stopwords* para diminuir o volume de processamento, esta metodologia contrapõe-se aos trabalhos justamente no momento em que é mantida toda a palavra para gerar o sumário e os resultados apresentados são satisfatórios e motivadores.

O trabalho pode ainda ser ampliado em relação aos parâmetros dos percentuais de compressão que são fixos, já que estes podem ser aprendidos, no decorrer da interação, com cada autor. Outra abordagem seria a comparação dos resultados do algoritmo Perfil preservando e removendo as *stopwords*. Como proposta para trabalhos futuros pretende-se implementar as métricas de [8] usadas neste trabalho, introduzindo Aprendizagem Autônoma, usando o conceito Cadeias Ocultas de Markov (*Hidden Markov Models* - HMM) [18] e criar perfis para os usuários [2], usando aprendizado por reforço para modelagem autônoma. Esta idéia poderia ser entendida e usada na Internet como forma de sumarização para notícias as quais o usuário mais se interesse.

6. REFERÊNCIAS BIBLIOGRÁFICAS

[1] Brown, R. (1973). "A first language", UNIVERSITY Press, Cambridge. http://bowland files.lancs.ac.uk/chimp/langac/LECTURE2/2brown.htm, acessado em março de 2008.

[2] Guelpeli, M.V.C.; Ribeiro, C.; Omar, N. (2004). "Aprendizado por Reforço para um Sistema Tutor Inteligente sem Modelo Explícito do Aprendiz", Revista Brasileira de Informática na Educação- RBIE – SBC Volume 12 – Número 2 pág. 69-77 - ISSN 1414-5685- mês de Julho a Dezembro de 2004-Rio de Janeiro-Brasil.

[3] Guelpeli, M.V.C.; Garcia A.C.B.(2007)."Automatic Summarizer Based on Pragmatic Profiles" International Conference WWW/Internet 2007- IADIS- Volume II pág. 149-153- ISBN: 978-972-8924-44-7 - mês de Outubro de 2007- Vila Real-Portugal .

[4] Franco, M.G.; Reis, M.J; Gil, T.M.S.(2003). "Comunicação, Linguagem e Fala" Ministério da Educação de Portugal, Lisboa.

[5] Hearst, M. A. (1993)."TextTiling: A quantitative approach to discourse segmentation." Technical Report Sequoia 93/24, Computer Science Division, University of California, Berkeley.

[6] Hearst, M. A.(1997)."TextTiling: Segmenting text into multi–paragraph subtopic passages" Computational Linguistics, vol. 23, no. 1 pp. 33–64, 1997. Disponível em:http://ucrel.lancs.ac.uk/acl/J/J97/J97-1003.pdf , acessado em 06 de Maio de 2007.

[7] Hutchins, J. (1987). "Summarization: Some problems and Methods." In: Jones. Meaning: The frontier of informatics. Cambridge. London, pp. 151-173.

[8] Hovy, E. (1988). "Generating Natural Language under Pragmatic Constraints." Lawrence Erlbaum Associates Publishers, Hillsdale, New Jersey.

[9] Hovy, E.; C.Y. Lin; L. Zhou. (2005). "A BE-based multi-document summarizer with sentence compression." To appear in Proceedings of Multilingual Summarization Evaluation (ACL 2005), Ann Arbor, MI.

[10] Hovy, E.; Kim, S.M.(2005). "Automatic Detection of Opinion Bearing Words and Sentences." In Proc. of IJCNLP-05, 2005.

[11] Larocca, J. N.1, Santos,A. D. S, Kaestner ,C. A.A. e Freitas A. A. (2000). "Generating Text Summaries through the Relative Importance of Topics. ".Lecture Notes in Computer Science Springer Berlin / Heidelberg Volume 1952/2000, ISSN0302-9743 (Print) 1611-3349 (Online) pp 300, 2000, Brazil.

[12] Luhn, H. P.(1958). "The automatic creation of literature abstracts.".IBM Journal of Research and Development, 2, pp. 159-165.

[13] Magalhães, T.M.V.(2006)."O Sistema Pronominal Sujeito e Objeto na Aquisição do Português Europeu e do Português Brasileiro". Tese de Doutorado, UNICAMP.

[14] Mittal,V. O.,Kantrowitz,M.,Goldstein,J.,Carbonell,J. G.(1999)"Selecting Text Spans For Document Summaries: Heuristics And Metrics" ,In *Aaai/Iaai* (1999), Pp. 467-473.

[15] Pardo, T.A.S.; Espina, A.P.; Rino, L.H.M.; Martins, C.B.;(2001). "Introdução à Sumarização Automática." Tech. Report RT-DC 002/2001, Departamento de Computação, Universidade Federal de São Carlos. Abril. 38p.

[16] Pardo, T.A.S.; Rino, L.H.M.; Martins (2003). "TeMário: Um Corpus para Sumarização Automática de Textos." Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional, NILC - ICMC-USP, Outubro de 2003.

[17] Pardo, T.A.S.; Rino, L.H.M.; Martins (2006) "A Coleção TeMário e a Avaliação de Sumarização Automática." Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional, NILC - ICMC-USP, Janeiro de 2006.

[18] Rabiner, L. R.; Juang, B.H. (1986), "An introduction to hidden Markov models", IEEE ASSP Magazine, Vol. 3(1), pp. 4-16.

[19] Riloff, E.; Wiebe, J.; Phillips, W. (2005). "Exploiting Subjectivity Classification to Improve Information

Extraction", Proceedings of the 20th National Conference on Artificial Intelligence (AAAI-05) .

[20] Salton, G.; Buckley, C. (1988). "Term-weighting Approaches in Automatic Text Retrieval." Information Processing and Management 24, pp. 513-523.

Tabela 2. Resultado das Médias e do Desvio Padrão da aplicação algoritmo TextTiling e Perfil no corpus Temário com todas as medidas.

Jornais	Seções		Recall		Precision		F-Mesure		C_R		NS_{org}	
			Text Tiling	Perfil	Text Tiling	Perfil	Text Tiling	Perfil	Text Tiling	Perfil	Text Tiling	Perfil
Folha SP	Mundo	Med	25.72	37.51	33.33	46.24	28.17	39.94	61.65	64.00	38.35	36.00
		DP	13.69	13.38	15.65	15.31	13.91	13.18	5.62	5.06	5.62	5.06
	Opinião	Med	23.62	35.80	28.78	37.87	25.48	35.94	60.40	66.40	39.6	34
		DP	8.70	14.68	10.77	16.17	8.88	14.44	6.74	5.71	6.74	6.48
	Especial	Med	25.72	37.72	33.45	44.11	25.84	40.43	59.90	63.80	39.85	36.20
		DP	13.58	13.96	17.03	14.89	13.43	14.00	7.56	4.34	7.56	4.34
JB	Inter nacional	Med	36.16	52.90	38.91	46.92	33.95	45.85	59.05	67.20	40.28	32.80
		DP	21.57	22.52	17.07	17.27	12.89	13.81	12.14	9.64	12.14	9.64
	Política	Med	25.30	40.07	33.61	47.18	28.05	42.27	59.35	64.90	40.13	34.65
		DP	10.80	16.37	10.06	13.88	9.38	13.52	10.10	3.65	10.10	4.82

Tabela 3. Comparativo entre os algoritmos TextTiling e Perfil usando o Teste t de variância Diferente.

Jornais	Seções		Recall		Precision		F-Mesure		C_R		NS_{orig}	
			t_{cal}	t_{tab}	t_{cal}	t_{tab}	t_{cal}	t_{tab}	t_{cal}	t_{tab}	t_{cal}	t_{tab}
Folha SP	Mundo	5%	***2.588***	***2.024***	***2.481***	***2.024***	***2.655***	***2.024***	1.389	2.024	1.389	2.024
	Opinião	5%	***3.193***	***2.024***	***2.092***	***2.024***	***2.758***	***2.024***	***3.038***	***2.024***	***2.678***	***2.024***
	Especial	5%	***2.757***	***2.024***	***2.106***	***2.024***	***3.361***	***2.024***	2.001	2.024	2.001	2.024
JB	Inter	5%	***2.400***	***2.024***	1.474	2.024	**2.817**	**2.024**	***2.351***	***2.024***	***2.351***	***2.024***
	Política	5%	***3.368***	***2.024***	***3.541***	***2.024***	***3.865***	***2.024***	***2.310***	***2.024***	***2.397***	***2.024***