# Extração Automática de Expressões Indicativas para a Classificação de Textos Científicos

**Danilo Machado Junior[1], Valéria Delisandra Feltrim[1]**

[1]Departamento de Informática – Universidade Estadual de Maringá (UEM)
Maringá – PR – Brasil

{danilo.junior, valeria.feltrim}@din.uem.br

***Abstract.*** *This work aims the construction of an automatic system to extract formulaic expressions used as classifying features of Portuguese scientific texts' sentences. The underling classification task consists of setting one out of seven rhetorical categories used by the* AZPort *classifier for each sentence of a scientific abstract. The bag-of-clusters approach was adopted to extract formulaic expressions to be used as classifying features and different heuristics were implemented to increase the quality of the system output. In order to avoid irrelevant clusters, the statistic measure Information Gain was also applied. The accomplishment of this work will help to overcome some bottlenecks in the adaptation of the* AZPort *classifier for other research areas.*

***Resumo.*** *Este trabalho dedica-se à construção de um sistema automático para extração de expressões utilizadas como atributos de classificação de sentenças em textos científicos. Tal classificação consiste na atribuição às sentenças de um resumo científico um dos sete papéis retóricos identificados pelo classificador AZPort.Para a geração das expressões utilizou-se a abordagem bag-of-clusters e foram empregadas diferentes heurísticas para a filtragem das expressões geradas.Utilizou-se a medida estatística Information Gain para avaliar a relevância das expressões. A conclusão deste projeto auxiliará a adaptação do AZPort para diferentes domínios evitando-se o gargalo da anotação manual de expressões.*

## 1. Introdução

O AZPort (Feltrim et al., 2004) é um classificador estatístico de estruturas retóricas implementado para resumos acadêmicos em português. Tal classificador atribui uma categoria retórica para cada sentença do resumo apresentado, com base em um conjunto de atributos extraídos das sentenças do texto. Um desses atributos, chamado de "expressão", detecta a presença de expressões indicativas do papel retórico da sentença com base em uma lista de expressões gerada manualmente, por meio da observação e anotação de corpus. Uma expressão pode ser considerada indicativa do papel retórico da sentença quando ela (ou suas variações) ocorre com certa freqüência em sentenças da mesma categoria retórica. Vale ressaltar que esse é um dos atributos de maior poder de distinção entre as possíveis categorias a serem atribuídas às sentenças.

A geração manual da lista de expressões consiste na classificação manual das sentenças do corpus e o reconhecimento e anotação de expressões mais freqüentes nas sentenças. Esse processo de reconhecimento e anotação das expressões representa um

gargalo na adaptação do classificador para outras áreas devido ao tempo e trabalho demandados. Com o objetivo de sanar este gargalo e facilitar a adaptação do classificador para áreas além da Ciência da Computação, abordada atualmente, este trabalho dedica-se a construção de um sistema para a geração automática da lista padrão de expressões baseada em corpus cujas sentenças já estejam classificadas quanto ao papel retórico. Este sistema deve ainda avaliar a relevância das expressões geradas e possuir mecanismos para eliminar as expressões irrelevantes ao processo de classificação.

## 2. Materiais e Métodos

Para a realização deste trabalho foi feita uma revisão bibliográfica dos possíveis métodos de extração de termos compatíveis com o contexto descrito e o modelo *bag-of-clusters*, proposto por Anthony & Lashkia (2003), foi escolhido como a melhor opção. Esse método extrai conjuntos de palavras das sentenças previamente classificadas e as agrupa em *clusters* ou n-gramas, respeitando sua ordem. Cada *cluster* é considerado um atributo de classificação para as sentenças em que é encontrado. Como cada *cluster* pode estar presente em mais de um tipo de sentença, deve-se criar mecanismos para filtrar os *clusters* menos relevantes, ou seja, aqueles que aparecem em muitos tipos de sentenças e, portanto, apresentam pouco poder de classificação. Tais *clusters* constituem o chamado "ruído".

O sistema para implementação do modelo *bag-of-clusters* foi construído na linguagem de programação C++ utilizando-se o ambiente de desenvolvimento Borland C++ Builder 6.0. O sistema foi dividido em dois módulos distintos: o primeiro é constituído por ferramentas de pré-processamento do corpus anotado utilizado como entrada, e o segundo módulo consiste nos algoritmos de identificação e geração de *clusters*, além das ferramentas para avaliação e filtragem dos dados gerados.

A primeira ferramenta do módulo de pré-processamento constitui um organizador que elimina as etiquetas XML utilizadas para anotação no corpus que não sejam relevantes ao processo de extração de expressões. As etiquetas não relevantes ao processo são constituídas por delimitadores de sentenças, parágrafos e resumos, já as relevantes são etiquetas que especificam o papel retórico da sentença. Por fim, o organizador agrupa as sentenças do mesmo papel retórico.

Com a finalidade de aumentar a freqüência dos *clusters*, a ferramenta Lemma-br, desenvolvida pela equipe do Núcleo Interinstitucional de Lingüística Computacional (NILC), foi acoplada ao sistema de pré-processamento com a finalidade de colocar os verbos na forma infinitiva. O Lemma-br faz uso do etiquetador morfossintático MXPOST (Ratnaparkhi, 1996), que também é utilizado pelo sistema e pode ser útil em uma futura busca de *clusters* sinônimos.

A última ferramenta incorporada ao módulo de pré-processamento tem a finalidade de remover palavras freqüentemente encontradas em sentenças de todos os tipos retóricos. Tais palavras, conhecidas como *stopwords*, podem constituir uma fonte de ruído para as expressões geradas.

Após o pré-processamento, o texto está pronto para o processo de extração e seleção de *clusters* do segundo módulo. O usuário deve fornecer o número mínimo e

máximo de palavras por n-grama e então a lista de *clusters* é gerada. Esta lista contém, além dos *clusters*, as classes nas quais eles ocorrem, suas freqüências totais e o valor da *Information Gain* de cada um. Esse valor é explicado a seguir.

O usuário então é questionado sobre quais heurísticas devem ser aplicadas. Tais heurísticas foram implementadas com o objetivo de reduzir o ruído da lista de expressões por meio da remoção de *clusters* com pouco ou nenhum poder de classificação. As heurísticas permitem definir um limite mínimo de freqüência e um limite máximo de classes para cada *cluster*, além de um limite máximo da entropia das sentenças no qual o *cluster* ocorre. Tais limites são definidos pelo usuário e os *clusters* que não se enquadram nas exigências são eliminados da lista.

Para medir a relevância das expressões na lista gerada, é calculada a *Information Gain* (Yang & Pedersen, 1997; Mitchell, 1997; Manning & Schutze, 1999) de cada uma delas. A IG será utilizada para se definir um limiar de corte seguro para as expressões, eliminando-se aquelas que não representam bons atributos para o classificador.

## 3. Resultados Parciais e Discussão

Utilizou-se como entrada do sistema um corpus com cerca de 370 sentenças. Foram geradas listas com diferentes números de palavras por n-grama, alternando-se a aplicação de ferramentas de pré-processamento e diferentes combinações das heurísticas do sistema. Essas listas encontram-se em processo de teste no AZPort.

Para se comparar os resultados da IG calculada pelo sistema, o corpus foi adaptado para ser processado pela ferramenta *Mover* que segue o modelo de Anthony & Lashkia (2003). A lista de *clusters* obtida pelo sistema e suas medidas IG se mostraram semelhantes à lista obtida pelo Mover, salvas as alterações provocadas pela aplicação das ferramentas de pré-processamento, as quais provocam um sensível aumento no IG *score* de alguns *clusters* produzidos pelo novo sistema.

Espera-se que os sucessivos testes com o AZPort permitam a escolha de um limiar seguro e eficiente para filtrar as expressões relevantes ao processo de classificação.

## 4. Conclusões e Trabalhos Futuros

O sistema mostrou-se versátil na geração de *clusters* de diferentes tamanhos definidos pelo usuário de forma rápida e eficiente. De fato, um corpus com cerca de 370 sentenças pode ser processado em poucos segundos, ao contrário da anotação manual. As heurísticas e ferramentas de pré-processamento reduzem de forma significativa o ruído da saída do sistema. Entretanto, ainda não se definiu um limiar de corte baseado da *Information Gain* das expressões. Todavia, com o aperfeiçoamento da extração automática, será possível adaptar o AZPort para diferentes contextos com maior rapidez.

Com o intuito de aumentar a lista de expressões e ainda identificar *clusters* sinônimos, aumentando a freqüência e IG *score* destes, podemos citar como trabalho futuro a criação de uma ferramenta para buscar palavras sinônimas na base de dados lexical do thesaurus eletrônico para português (Dias-da-Silva et al., 2000; Dias-da-Silva

et al., 2003; Maziero et al., 2008). Essa ferramenta deve ainda possuir um sistema de desambiguação de termos de forma a evitar a inserção de ruído na lista de expressões.

Ainda como trabalho futuro, podemos destacar a inserção de outras medidas estatísticas para a definição de um limiar de corte para a lista de expressões. Este trabalho se tornará especialmente necessário caso os testes de classificação no AZPort com as listas limitadas pela *Information Gain* se mostrem inconclusivos ou insatisfatórios. Por fim, um novo classificador baseado apenas no atributo expressão deve ser construído.

## Referências


Anthony, L., Lashkia, G.V. (2003). Mover: A machine learning tool to assist in thereading and writing of technical papers. IEEE Transactions on Professional Communication, 46(3), p.185–193.

Dias-da-Silva, B.C.; Moraes, H.R.; Oliveira, M.F.; Hasegawa, R.; Amorim, D.A.; Paschoalino, C.; Nascimento, A.C. (2000). Construção de um thesaurus eletrônico para o português do Brasil. Processamento Computacional Do Português Escrito e Falado (PROPOR), Vol. 4, pp. 1-10.

Dias-da-Silva, B.C.; Moraes, H.R. (2003). A construção de um thesaurus eletrônico para o português do Brasil. ALFA, Vol. 47, N. 2, pp. 101-115.

Feltrim, V.D., Pelizzoni, J.M., Teufel, S., Nunes, M.G.V., Aluísio, S.M. (2004). Applying Argumentative Zoning in an automatic critiquer of academic writing. In Proceedings of the 17th Brazilian Symposium on Artificial Intelligence (SBIA 2004), São Luis-MA, Brazil. Lecture Notes in Artificial Intelligence, 3171, Springer, p. 214-223.

Manning, C., Schütze, H. (1999). Foundations of Statistical Natural Language Processing. Cambridge, Massachusetts: MIT Press.

Maziero, E.G.; Pardo, T.A.S.; Di Felippo, A.; Dias-Da-Silva, B.C. (2008). A Base de Dados Lexical e a Interface Web do TeP 2.0 - Thesaurus Eletrônico para o Português do Brasil. VI Workshop em Tecnologia da Informação e da Linguagem Humana (TIL), pp. 390-392.

Mitchell, T. (1997). Machine Learning. New York: McGraw-Hill.

Ratnaparkhi, A. (1996). A maximum entropy part-of-speech tagger. In Proceedings of the Empirical Methods in Natural Language Processing Conference.

Yang, Y., Pedersen, J.O. (1997). A comparative study on feature selection in text categorization. In Machine Learning: Proceedings of the Fourteenth International Conference (ICML'97), 412-420.