(Re)Estruturação de parágrafos em redações dissertativas/argumentativas com base em Padrões Problema-Solução e Grafos.

João Carlos Silva Nobre, Sérgio Roberto Matiello Pellegrino

Divisão de Ciência da Computação
Instituto Tecnológico de Aeronáutica (ITA) – São José dos Campos, SP - Brasil

{jcnobre, pell}@ita.br

Abstract. *This paper presents a proposal for (re) organization of paragraphs using Graphs Theory and the organization of Problem-Solution Patterns for essays discourse in Portuguese. The system receives a text and it must to build their paragraphs in order to observing the organization of problem-solution patterns associated to Graphs Theory. Preliminary results indicate that in 80% of essays the process had improved the presentation of the content.*

Resumo. *Este artigo apresenta uma proposta de (re)estruturação de parágrafo por meio da teoria dos grafos e a organização de padrões problema-solução em redações dissertativas em português. O sistema recebe como entrada um texto e deve reconstruir sua estrutura de parágrafos por meio da organização de padrões problema-solução apoiada na teoria dos grafos. Resultados preliminares indicam que em 80% das redações houve uma melhor apresentação de conteúdo.*

1. Introdução

O processamento de línguas naturais - PLN não é uma tarefa trivial e vem recebendo da comunidade acadêmica mais atenção. Analisar um texto sob a ótica do discurso pressupõe que certas barreiras do PLN tenham sido superadas, visto que há aumento da complexidade da modelagem e do tratamento computacional quando se envolve conhecimento nos níveis semântico, pragmático e discursivo, segundo Pardo (2005).

Promover a (re)estruturação dos parágrafos em redações de vestibular torna-se mais complexo, não apenas por se ater ao aspecto discursivo e morfossintáticos, como também porque são textos que apresentam, em tese, incorreções de natureza gramatical, que dificultam e reduzem o desempenho de sistemas concebidos para PLN.

Sazdijian (2007) evidencia que a análise do discurso pode prover um melhor entendimento das dificuldades encontradas por "alunos" ao produzirem suas redações. Seguindo essa linha de pensamento, o presente artigo aborda o uso de padrões problema-solução (Hoey, 2001), em redações de vestibular, visando apresentar uma proposta de organização textual, utilizando as mesmas idéias e construções sintáticas do texto original, com o objetivo de melhorar a apresentação daquelas idéias, propiciando uma melhor compreensão do texto, e, ainda a facilitação da leitura.

Os estudos de Costa (2005) e Newman (2004) corroboram a necessidade de análise das redes complexas a partir de suas características para identificar sua topologia, conhecer o nível de coesão existente entre os documentos, identificar documentos determinantes para manutenção das ligações e extrair conhecimento para tomada de decisão.

A forma de concepção do texto dissertativo permite a adoção da teoria dos grafos e análise de redes complexas para avaliar algumas características da sua estrutura textual, como exemplo a coesão e a estruturação dos parágrafos. Os resultados desse artigo comportam tal afirmação.

2. Embasamento Teórico

2.1 A Organização Problema-Solução

Segundo Hoey (2001), o texto pode ser definido como a evidência visível de uma interação com intenção auto-contida entre um ou mais autores ou leitores. Esta interação é vislumbrada como uma série de perguntas hipotéticas que o leitor faz ao escritor, cujas respostas deveriam suprir a expectativa do leitor tanto no nível frasal como no nível do discurso, podendo ser prefixada e, portanto, conhecida por ambos, leitor e escritor. Ele prefere explicar este fenômeno por meio do que chama popularmente de padrões de organização.

Esses padrões são principalmente caracterizados: por uma organização e não uma estrutura, em que certos elementos são mais freqüentes do que outros; pela não existência de combinações inadequadas, novamente em contraste com as estruturas; pelo fato de que eles estão limitados pela cultura, e, por último, pela sua popularidade, isto é, a grande freqüência com que alguns deles ocorrem.

Um dos mais comuns é o padrão problema-solução, proposto em 1983, o qual é caracterizado pelos seguintes elementos: (1) uma situação anterior opcional, que fornece um contexto para o padrão; (2) o problema ou aspecto de uma situação que exige uma resposta; (3) a resposta ao problema e (4) um resultado positivo ou de Avaliação.

Padrões Textuais têm sido descritos com a finalidade objetiva de computar a forma como cláusulas ou grupos de cláusulas se relacionam entre si no discurso. Em outras palavras, a verdadeira natureza do modelo é o senso de ordem percebida por um leitor. Portanto, o padrão problema-solução é apenas uma das várias possibilidades de organização de texto (Hoey, 2001).

O padrão problema-solução - PPS surge como um resultado do processamento mental do escritor, respondendo a uma série de perguntas previsíveis que refletem o relacionamento das frases do texto. Não obstante a ausência de uma ordem pré-definida para o aparecimento das respostas, a principal característica do padrão problema-solução é a sinalização léxica.

McCarthy (1996), ao dissertar sobre padrões textuais, argumenta que os estudos mostram a prevalência de três tipos de padrões comuns que são classificados como problema-solução, hipotético-real, geral-específico.

2.2. O Parágrafo Dissertativo

O parágrafo é um grupo de frases que tratam de um mesmo tópico, por isso, o desenvolvimento deve ser completo e coerente, isto é, cada frase deve se referir, direta ou indiretamente, ao tópico frasal. Como unidade mínima do texto, ele deve apresentar: uma frase contendo a idéia principal (frase nuclear) e uma ou mais frases que explicitem tal idéia.

O parágrafo pode processar-se de diferentes maneiras: (1) Enumeração - caracteriza-se pela exposição de uma série de coisas, uma a uma; (2) Comparação - a frase nuclear pode-se desenvolver mediante comparação, que confronta idéias, fatos, fenômenos e apresenta-lhes as semelhanças ou dessemelhanças; (3) Causa e Conseqüência - a frase nuclear, muitas vezes, encontra no seu desenvolvimento um segmento causal, fato motivador, e, em outras situações, um segmento indicando conseqüências, fatos decorrentes; (4) Tempo e Espaço - muitos parágrafos dissertativos marcam temporal e espacialmente a evolução de idéias, processos; e (5) Explicitação - num parágrafo dissertativo, pode-se conceituar, exemplificar e aclarar as idéias para torná-las mais compreensíveis.

2.3. Teoria dos Grafos – Redes Complexas

Um grafo é um conjunto de vértices e um conjunto de arestas, em que cada aresta conecta um par de vértices. Portanto, um grafo é definido por G = {V, A}, onde V é o conjunto de vértices e A o conjunto de arestas definidas entre dois vértices.

O grafo pode ser orientado e não orientado. Um grafo orientado determina a direção das arestas entre os vértices e, os não orientados não apresentam direção, e a aresta é definida nos dois sentidos.

Um caminho é definido pelas arestas que deverão ser percorridas entre um par de vértices. As arestas podem ser ponderadas, introduzindo o conceito de distância, que se refere ao comprimento entre o par de vértices. O significado da distância está diretamente relacionado ao conceito apresentado ao grafo.

É possível verificar em um grafo que os vértices podem não ser totalmente interligados, introduzindo o conceito de grau. O grau de um vértice é determinado em função do número de arestas a ele conectadas.

Segundo Costa (2005), as redes complexas são caracterizadas por representar um conjunto de elementos em que ligações dependem da característica do estudo que se deseja construir.

As redes possuem propriedades próprias, que são suportadas pela teoria de grafos e se apresentam de três formas: (a) *directed* - ligações estabelecidas em uma única direção, também chamadas de cíclicas, e são representadas por um grafo orientado; (b) *undirected* - ligações estabelecidas em duas direções, também chamadas de acíclicas, que são representadas por um grafo não orientado, e (c) *bipartite* – existência de várias ligações com propriedades diferentes entre pares de vértices.

A análise topológica de redes complexas consiste em mensurar as propriedades estruturais envolvidas na rede, como a conectividade (como e com qual vértice estabelecem-se as ligações) e, a centralidade (Costa, 2005) (qual vértice possui a melhor

conexão ou a maior influência). De acordo com Newman (2004), cada propriedade é utilizada para caracterização topológica.

3. (Re)estruturação de Parágrafos

Esta seção apresenta o (re)estruturador de parágrafos, destacando os procedimentos na execução do processo. Antes, porém, as subseções 3.1 e 3.2 descrevem a construção do corpus de trabalho e algoritmo de (re)estruturação a partir da análise do corpus, respectivamente.

Os textos que estarão sob análise apresentam diversas incorreções de natureza ortográfica, sintática, semântica e pragmática, o que dificulta sobremaneira o desempenho de qualquer sistema concebido para analisar textos de forma sistemática.

Diante disso, procurar indícios norteadores à (re)estruturação dos parágrafos pode ser mais indicado do que procurar por "padrões fechados", haja vista alguns "autores" dessas redações, em tese, cometerem erros no emprego de conjunções, verbos, preposição, advérbios, sendo os mais comuns, os de pontuação.

3.1. Construção do Corpus

O modelo de (re)estruturação de parágrafos proposto é baseado em conhecimento, abrangendo as informações morfológicas, sintáticas, semânticas. As bases de conhecimento são produzidas a partir da análise de um corpus de redação de vestibular, como se descreve a seguir.

Um corpus de 60 redações de vestibular, cujo tema principal é a violência, chamado CorpusR (CORPUS de Redação), contendo 16.717 *tokens*, 735 sentenças e 369 parágrafos, montado a partir de textos das redações de vestibular, do ano de 2008. Nesse corpus, efetuou-se marcações retóricas manualmente e, posteriormente, usou-se parte do corpus marcado para automatizar o processo de identificação dos Padrões Problema-Solução, obtendo-se uma precisão de 91,7 como medida de avaliação do sistema. Contudo, optou-se por utilizar o corpus marcado manualmente, a fim de evitar possíveis distorções no resultado final.

As anotações das informações de natureza morfológicas, sintáticas do corpus foram executadas por meio do *parser* Palavras (Bick, 2000). O arquivo resultante desse processamento possui informações que para serem melhor utilizadas, tem sua estrutura alterada, por meio de um aplicativo desenvolvido, o qual gera arquivos em formato XML, segundo a estrutura proposta por Bruckschen, et al. (2008).

3.2. Algoritmo de (Re)estruturação

A (re)estruturação dos parágrafos apóia-se na teoria dos grafos e na organização do PPS, a qual é similar a organização tradicional de texto. Ao se comparar as duas estruturas tem-se: (1) introdução que é composta por situação e problema; (2) desenvolvimento formado pela resposta ou solução; e conclusão que corresponde a avaliação. Os inter-relacionamentos entre cada parte do texto é avaliado por meio da teoria dos grafos.

A partir do conhecimento apontado pelas marcações do corpus, elaborou-se o algoritmo para (re)estruturação, descrito na Figura 1.

```
Para cada texto
        Criar lista vazia de sentenças para cada tipo de padrão
        Computar grau, densidade  e gerar dados dos grafos
        Para cada parágrafo
                Verificar o PPS de cada frase
                Se PPS do parágrafo >= 3 # ([s]+p+r/s+[a]), [elemento opcional]
                        Manter frases juntas e posicionamento do parágrafo no texto.
                Caso contrario
                        Verificar quantidade de elos com as frases da respectiva lista de padrão
                        Encaixar a frase na melhor posição da lista de padrão
        Para cada lista de padrão   #na ordem → s – p – r/s  – a
                Criar parágrafos  # máximo de palavras 60 por parágrafos
                Verificar  e eliminar elementos de contradição entre cada frase
        Para cada parágrafo com PPS >= 3
                Verificar quantidade de elos com os  demais parágrafos
                Encaixar o parágrafo na melhor posição da estrutura atual, levando em conta a posição
                  original
        Para novo_texto
                Verificar e eliminar elementos de contradição entre cada parágrafo
Fim para cada texto
```

Figura 1. Algoritmo de (re)estruturação de parágrafos

3.3. O (Re)estruturador

O (Re)estruturador é baseado em informações morfossintáticas fornecidas pelo *parser* Palavras e o corpus marcado manualmente com PPS. Dessa maneira, as redações são primeiramente processadas pelo *parser*, de modo a obter todo o conhecimento morfossintático necessário à entrada do (Re)estruturador. A Figura 2 apresenta o esquema de funcionamento.

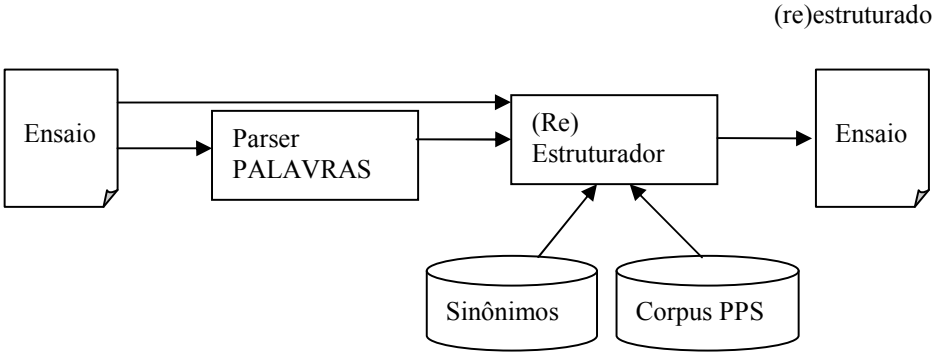

Figura 2. Esquema do processo de (Re)Estruturação

Durante o processo de (Re)estrutura, o sistema cria uma rede complexa para a coleção de parágrafos e outra para a coleção de frases, definidas por G(P, A) e G(S, A) onde os parágrafos "P", ou sentenças "S", representam os vértices e "A" as arestas formadas a partir de lista de substantivos e adjetivos, quando este é antecedido de substantivo, para cada parágrafo ou sentença do ensaio. Efetua-se a contagem dos elos que relacionam as frases e os parágrafos, ou seja, computa-se o peso do relacionamento por meio da quantidade de substantivos e/ou adjetivos e seus sinônimos, por intermédio da base de sinônimos Tep2 (Maziero et al., 2008), que são comuns a cada frase e a cada parágrafo. A partir dos valores computados geram-se automaticamente o grau de entrada

e saída de cada vértice, a densidade da rede e os arquivos com a lista de adjacência para visualização dos grafos no software Pajek 1.20, conforme ilustrado na Figura 3.

Após efetuar os cálculos, passa-se a efetiva organização dos parágrafos. De forma a melhorar a compreensão do método, utilizar-se-á o texto da Figura 4 para explicar o processo de (re)estruturação. Ressalte-se que os dados entre colchete significam, respectivamente, o posicionamento da frase e o PPS identificado.

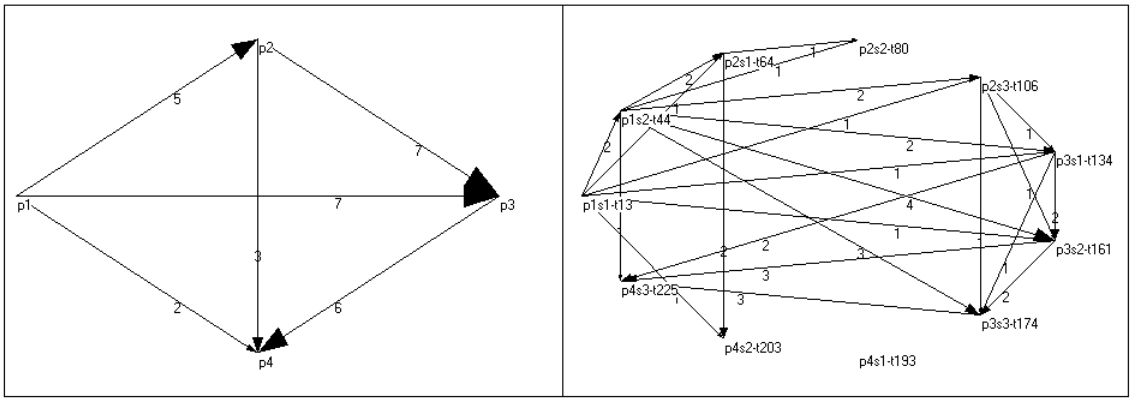

Figura 3. Grafos da representação de parágrafos (A) e sentenças (B)

A idéia básica da (re)estruturação é criar uma seqüência lógica para o texto, de forma a expor a situação e problema, as respostas ou soluções aos problemas identificados e, por fim, apresentar a avaliação das propostas de solução ou resposta ao problema.

Inicialmente, as sentenças encontram-se agrupados por tipo de PPS. O primeiro PPS a ser manipulado é o padrão "Situação", que ocorre duas vezes no texto exemplo, nas posições p2s2 e p4s1. Contudo, somente o p2s2 pode ser movimentado livremente, haja vista não haver no parágrafo de origem mais que dois tipos do PPS. Assim sendo, a sentença p2s2 deve ser a primeira.

Para compor os parágrafos referentes ao problema, busca-se aquela sentença com maior número de elos com o parágrafo ou frase anterior, ou seja, o parágrafo tipo "situação" do PPS, segundo o exemplo.

A próxima sentença a ser agregada ao parágrafo tipo "problema" será a de maior elo com a sentença anterior, no caso, a primeira sentença do parágrafo tipo "problema" do PPS. A agregação de sentenças ao parágrafo é limitada a sessenta palavras.

O mesmo princípio é adotado aos parágrafos dos tipos "resposta/solução" e "avaliação" do PPS.

Para reposicionar os parágrafos que possuem mais de dois tipos de PPS, observa-se a quantidade de ligações com os outros parágrafos e a posição original. Importante pontuar que, no texto exemplo, o quarto parágrafo foi mantido, pois, em tese, seria o parágrafo conclusivo.

Efetua-se uma varredura no texto à procura de elementos que indiquem contradição. Por exemplo, ao agrupar as frases identificadas por p3s1 e p1s2, constata-se que cada frase inicia-se por conjunções coordenadas adversativas, gerando um

conflito de idéias no texto. Para sanar o problema, optou-se por excluir a segunda conjunção.

As informações obtidas são utilizadas durante o processo de inserção das frases e planejamento da disposição dos parágrafos, conforme mencionado no algoritmo, Figura 1.

A influência da televisão na mente de uma criança é um fato [p1s1, Problema]. No entanto, o exemplo de moderação e inteligência dos pais é muito mais importante, na formação da personalidade mirim, que a programação infantil violenta transmitida pela televisão [p1s2, Resposta].

Não é possível negar os efeitos nocivos que programas televisivos podem provocar na formação do caráter de uma criança [p2s1, Problema]. É cientificamente verificado que indivíduos de até sete anos de idade possuem caráter facilmente modelável [p2s2, Situação]. Características violentas podem, facilmente, ser agregadas à mente de quem, nessa faixa etária, submete-se constantemente a esse tipo de programação [p2s3, Problema].

Apesar disso, a responsabilidade maior sobre o desenvolvimento infantil pertence aos pais, sobre os quais repousa um profundo senso de respeito e confiança dos filhos [p3s1, Resposta]. O amor, carinho, aconselhamento, equilíbrio e bom exemplo de conduta dos pais expandirão a inteligência emocional e a capacidade de crítica e análise [p3s2, Resposta]. O procedimento dos pais é bem mais determinante que a programação televisiva [p3s3, Resposta].

A realidade é que o objetivo principal das redes de comunicação, em geral, é o lucro [p4s1, Situação]. Programas de violência atraem crianças e jamais serão retirados [p4s2, Problema]. Cabe aos pais das aos filhos a boa referência de comportamento, vigiando e analisando, sempre, suas próprias condutas [p4s3, Resposta].

Figura 4. Redação original e na integra.

4. Experimentos

Com o propósito de verificar se a mudança no posicionamento das sentenças nas redações facilitou a sua compreensão e leitura, foram sorteadas seis redações para uma avaliação mais minuciosa num primeiro momento.

Compôs-se uma banca com três examinadores, com no mínimo especialização em Língua Portuguesa, para avaliar todas as redações, sendo que para cada redação foram apresentadas três versões: (A) texto original; (B) texto (re)estruturado pelo sistema, conforme algoritmo; e (C) texto (re)estruturado pelo sistema usando a estratégia de manter cada tipo do PPS em parágrafos diferentes.

A Tabela 1, resume o resultado do experimento de seis redações de um total de 60 que foram avaliadas, sendo que cada redação foi classificada atribuindo-se-lhes valores de “1” a “3”, segundo os quais “1” representa a melhor e “3” a pior estruturação de parágrafos.

Tabela 1. Demonstrativo da Avaliação das 6 redações

	R 1			R 2			R 3			R 4			R 5			R 6		
	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
Avaliador 1	2	1	3	1	1	3	2	1	3	1	2	3	2	1	3	1	1	3
Avaliador 2	2	1	3	2	1	3	2	1	3	1	2	3	1	2	3	2	1	3
Avaliador 3	2	1	3	2	1	3	2	1	3	1	2	3	2	1	3	2	1	3

Das seis redações, apenas uma não obteve melhora na leitura e compreensão das idéias. Aos inspecionar os motivos, verificou-se que o texto possuía muitas deficiências de natureza ortográfica, semântica e pragmática.

Os especialistas humanos concordaram que 48 redações se apresentam melhor estruturadas, ou seja, houve alguma melhora em 80% das redações (re)estruturadas pelo sistema.

5. Conclusões

A abordagem apresentada neste artigo é simples. Requer corpus com anotação morfossintática e retórica, sendo que a anotação retórica pode ser crítica à automatização do processo, visto depender da identificação a ser efetuada manualmente por um especialista. Entretanto, com a conclusão do módulo de identificação de PPS, que se encontra em fase de avaliação, a lacuna será preenchida.

Os resultados preliminares indicam que a abordagem empregada permite melhorar a estruturação do texto em 95% dos casos e que 80% tornam-se melhores, mesmo que os textos possuam incorreções de natureza léxica, sintática e semântica.

As características estruturais do grafo representativo da redação se mostram aplicáveis e satisfatórias, contudo, deve-se ampliar a investigação para determinar se há outras características de relacionamento intra e/ou inter-frases que permitam melhorar o desempenho.

References

Bick, E. (2000) The Parsing System PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. PhD Thesis, Arhus University, Arhus.

Bruckschen, M.; Muniz, F.; Souza, J.G.C.; Thiesen, J.; Fuchs, K. I.; Muniz, M.; Gonçalves, P.N.; Vieira, R.; Aluísio, S.M. (2008). Anotação Lingüística em XML do Corpus PLN-BR. Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional NILC - NILC-TR-08-09.

Costa, L. da F., Rodrigues F. A., Travieso, G. e Villas Boas, P.R..,(2005) "Characterization of Complex Networks: A survey of measurements". Instituto de Física de São Carlos, Universidade de São Paulo.

Hoey, M. (2001) Textual Interaction: An Introduction to Written Discourse Analysis. London: Routledge.

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.

McCarthy, M (1996) Discourse Analysis for Language Teachers. Cambridge: Cambridge Universitsy Press.

Newman, M. E. J.. "The Structure and Function of complex networks". University of Michigan. Department of Physics, University of Michigan, Ann Arbor, MI 48109, U.S.A.< http://www-personal.umich.edu/~mejn/courses/2004/cscs535/review.pdf> Acessado em 01.agosto.2009.

Pardo, Thiago A. S. (2005) Métodos para Análise Discursiva Automática. Tese de Doutorado. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos.

Sazdijian, Anaid Bertezlian (2007) As redações do SARESP: O texto argumentativo e a análise das Três Pontas. Dissertação de Mestrado. Departamento de Letras, Pontifícia Universidade de São Paulo, São Paulo.