Proceedings of Symposium in Information and Human Language Technology. Natal, RN, Brazil, November 4–7, 2015.

Análise Automática de Coerência Textual em Resumos Científicos: Avaliando Quebras de Linearidade

Leandro Lago da Silva[1], Valéria Delisandra Feltrim[1]

[1]Departamento de Informática – Universidade Estadual de Maringá (UEM)
CEP 87020-900 – Maringá – PR – Brazil

leandro@datacampo.com.br, vfeltrim@din.uem.br

Abstract. *This paper presents an extension of the coherence analysis module that is part of the writing tool called SciPo, allowing it to automate the analysis of the coherence dimension called Linearity Break. The proposed implementation is based on a combination of the entity grid model and information from the rhetorical structure of scientific abstracts, allowing it to generate messages that indicate possible linearity breaks in specific regions of the abstract. Experiments have shown that the combination of the entity grid model and information from the rhetorical structure is feasible and can be used as part of SciPo.*

Resumo. *Este artigo apresenta uma extensão do módulo de análise de coerência que é parte da ferramenta SciPo, visando à análise automática da dimensão chamada Quebra de Linearidade. A implementação proposta é baseada na combinação do modelo grade de entidades com informações provenientes da estrutura retórica do resumo, permitindo que o módulo gere mensagens que indiquem possíveis quebras de linearidade em regiões específicas do resumo. Experimentos mostraram que a combinação do modelo grade de entidades com a estrutura retórica é viável e pode vir a ser utilizada como parte da ferramenta SciPo.*

1. Introdução

A ferramenta SciPo [Feltrim et al. 2006] foi desenvolvida para auxiliar escritores iniciantes na escrita científica, em especial na escrita de resumos e introduções na área da Ciência da Computação. A ferramenta é voltada para a língua portuguesa e possui um módulo de análise de coerência (MAC), que detecta potenciais problemas de coerência textual em resumos.

O MAC é baseado na classificação de componentes retóricos e em Análise de Semântica Latente (*LSA*) [Landauer et al. 1998]. Atualmente, três tipos de relacionamentos semânticos, chamados de dimensões, são examinados pelo MAC [Souza and Feltrim 2013]. Uma quarta dimensão, chamada Quebra de Linearidade, foi proposta para o MAC, mas não chegou a ser automatizada. Essa dimensão busca identificar problemas de coerência locais que se caracterizam pela dificuldade em se estabelecer uma ligação clara da sentença atual com as sentenças adjacentes. Segundo os autores, os resultados obtidos com LSA para essa dimensão foram insatisfatórios,

sugerindo o uso de outros modelos de coerência, como a de grade de entidades proposta por [Barzilay and Lapata, 2008].

Visando a automatização da dimensão Quebra de Linearidade, este trabalho propõe utilizar informações provenientes da estrutura retórica em conjunto com a grade de entidades para gerar mensagens que indiquem possíveis problemas de coerência local em regiões específicas do resumo, indicando, por exemplo, que uma possível quebra de linearidade foi detectada em certo componente retórico. Os resultados experimentais mostram que a proposta é viável de ser incluída do MAC da ferramenta SciPo.

A Seção 2 apresenta a proposta. A Seção 3 apresenta a metodologia e os resultados das avaliações são mostrados nas seções 4 e 5. Por fim, a Seção 6 traz as conclusões do trabalho.

2. Análise Automática de Quebra de Linearidade

Vários trabalhos têm usado a grade de entidades para automatizar em algum nível a análise de coerência [Barzilay and Lapata 2008; Burstein et al. 2010; Elsner and Charniak 2011; Castro Jorge et al. 2014; Dias et al. 2014; Freitas and Feltrim 2014]. Uma característica comum a esses trabalhos é a análise do texto completo, o que é útil em vários contextos de aplicação.

Freitas e Feltrim (2014) mostraram que o uso da grade de entidades possibilita a identificação de resumos com quebras de linearidade, no entanto, a análise do texto como um todo não permite a identificar a localização das quebras. Informar que o texto possui quebras de linearidade sem dar indicar a região em que as quebras ocorrem é de pouca utilidade para uma ferramenta de auxílio à escrita como o SciPo. Assim, é preciso que as sugestões geradas pela ferramenta sejam mais específicas, informando, ainda que de forma aproximada, em qual trecho do texto a quebra foi detectada.

A solução proposta foi usar a grade de entidades na análise de trechos menores constituídos por um ou mais componentes retóricos. Essa análise por trechos permite a geração de mensagens que indiquem quebras de linearidade em um componente ou grupo de componentes retóricos específicos, permitindo a geração de mensagens mais específicas por parte da ferramenta.

A partir da identificação dos componentes retóricos, feita por meio de um classificador retórico, a análise da dimensão Quebra de Linearidade pode ser iniciada. Em uma primeira etapa da análise, grades de entidades individuais são construídas para todos os componentes retóricos compostos de pelo menos duas sentenças. A partir de cada grade é extraído um vetor de características que então é testado por um classificador que atribui uma de duas categorias possíveis: Com Quebra ou Sem Quebra. Sempre que um trecho é classificado como Com Quebra, uma sugestão é gerada ao usuário indicando que aquele componente retórico específico possui uma possível quebra de linearidade. O usuário, por sua vez, pode acatar a sugestão, retornar ao texto para modificá-lo e reenviá-lo para uma nova análise, ou pode ignorar a sugestão dada, o que faz com que o processo de análise prossiga.

Em uma segunda etapa, novas grades de entidades são construídas para todos os pares de componentes adjacentes. O processo de classificação se repete como na primeira etapa e caso a análise continue, uma nova etapa é iniciada. A cada nova etapa, grupos maiores de componentes retóricos, gerados por meio da adição de um

componente adjacente, são usados para gerar as grades de entidades e realizar a classificação. A análise continua enquanto não forem detectadas quebras de linearidade e termina quando houver um único grupo de componentes retóricos que corresponde ao resumo completo.

3. Metodologia

Para a identificação dos componentes retóricos foi utilizado o classificador AZPort [Feltrim et al. 2006], que classifica cada sentença de um resumo em uma de seis categorias retóricas: Contexto, Lacuna, Propósito, Metodologia, Resultado e Conclusão.

Para a construção das grades de entidades foi utilizado o sistema de Freitas (2013), que implementa o modelo de grade de entidades conforme proposto por Barzilay e Lapata (2008) para o português. Dois tipos de conhecimento linguístico foram considerados na construção das grades: (i) a função sintática das entidades (se sujeito (S), objeto (O), nenhum dos anteriores (X) ou ausente na sentença (-)) e (ii) a saliência, definida com base nas frequências das entidades observadas no discurso. Entidades que ocorrem pelo menos duas vezes no texto foram consideradas salientes.

A partir da grade de entidades foram extraídas as probabilidades de todas as possíveis transições de tamanho dois. Uma transição é uma sequência $\{S; O; X; -\}_n$ que representa as ocorrências da entidade em *n* sentenças adjacentes. As transições podem ser obtidas como sequências contínuas de cada coluna com certa probabilidade de ocorrência na grade. Dessa maneira, cada texto é representado por um conjunto fixo de transições e suas probabilidades, usando a notação padrão de vetor de características. Considerando a presença (+) ou a ausência (-) das informações sintáticas e de saliência, quatro configurações diferentes do modelo foram obtidas por meio das combinações de função sintática (+/-) e saliência (+/-).

Foram criados dois classificadores para a dimensão Quebra de Linearidade: um para classificar componentes retóricos isolados e o outro para classificar resumos completos. Os classificadores foram induzidos com o algoritmo J48 disponível no ambiente Weka [Witten and Frank 2005] e os resultados foram obtidos por meio de validação cruzada estratificada com 10 partições. O treinamento e teste dos classificadores foram feitos com o CorpusTCC [Souza and Feltrim 2013], um *corpus* composto por 408 resumos extraídos de monografias de conclusão de curso de graduação em Computação.

O classificador de componentes foi treinado com pares de componentes retóricos extraídos a partir dos resumos. Ao todo foram utilizados 1.160 pares de compostos por no mínimo três sentenças, sendo 580 pares originais e 580 pares gerados pela inversão das sentenças na fronteira dos componentes. O classificador de resumos completos foi treinado com 816 resumos, sendo 408 resumos originais e 408 resumos gerados pela inversão da ordem das sentenças. Em ambos os casos (pares e resumos), as versões geradas artificialmente foram consideradas Com Quebra enquanto os textos originais foram considerados Sem Quebra. A opção pela geração de versões artificiais para o treinamento dos classificadores se deu devido ao pequeno número de resumos originais anotados como tendo quebra de linearidade, o que deixa o *corpus* altamente desbalanceado.

4. Avaliação dos Classificadores

O classificador de componentes obteve taxa de acerto de 95,17% com a grade de entidades na configuração Sintático+ Saliência+. Dada a quantidade de pares usados no treinamento era esperado que essa configuração obtivesse melhor resultado, uma vez que ela incorpora mais conhecimento sobre as entidades.

O classificador de textos completos também obteve sua melhor taxa de acerto (85,05%) com a grade de entidades na configuração Sintático+ Saliência+. Essa taxa de acerto é menor do que a obtida com o classificador de componentes, provavelmente devido à diferença na quantidade de exemplos de treinamento.

Os resultados obtidos mostram que a grade de entidades é capaz de detectar quebras de linearidade mesmo em trechos pequenos, compostos de poucas sentenças. Assim, os dois classificadores que obtiveram os melhores resultados foram utilizados na dimensão Quebra de Linearidade.

5. Avaliação da Dimensão Quebra de Linearidade

A avaliação da dimensão Quebra de Linearidade foi avaliada com um conjunto de 28 resumos originais, sendo 14 resumos Com Quebra e 14 resumos Sem Quebra. Os resumos Com Quebra foram selecionamos manualmente do CorpusTCC por dois anotadores humanos. Os anotadores também identificaram, nesses resumos, os pares de sentenças que caracterizavam as quebras. Foram identificados 18 pares de sentenças com quebra de linearidade. Os 14 resumos Sem Quebra de linearidade foram selecionados aleatoriamente a partir do restante do CorpusTCC.

O primeiro experimento buscou verificar a acurácia da dimensão na identificação das quebras de linearidade. A taxa de acerto observada foi de 67,86%. Ao todo, 15 resumos foram avaliados como Com Quebra, sendo que 10 dos 14 resumos Com Quebra foram corretamente identificados.

Outro experimento, realizado apenas com os 14 resumos Com Quebra, buscou verificar a acurácia da dimensão em relação à identificação dos pares de sentenças anotados com quebra de linearidade. No total foram avaliados 73 pares de sentenças, sendo 18 pares Com Quebra e 55 pares Sem Quebra. Ao todo, 15 pares de sentenças foram avaliados como tendo quebra, sendo que nove dos 18 pares Com Quebra foram corretamente identificados. A cobertura para a classe Com Quebra foi mais baixa nesse segundo experimento, o que era esperado. De fato, identificar o par de sentenças que caracteriza a quebra de linearidade é uma tarefa difícil mesmo para anotadores humanos.

6. Conclusões

Este artigo apresentou uma proposta para a automatização da dimensão Quebra de Linearidade de modo que ela pudesse ser incluída no MAC da ferramenta SciPo. A proposta utiliza a grade de entidades como modelo para a avaliação de coerência de resumos científicos e o seu diferencial está na forma como o modelo é aplicado no contexto do MAC. O uso da grade de entidades para a análise de trechos menores de textos, juntamente com as informações provenientes da estrutura retórica do resumo, permite a geração de críticas e sugestões mais específicas, tornando-as mais úteis para os usuários da ferramenta SciPo.

Os resultados experimentais mostraram que a proposta de analisar trechos menores de texto usando a grade de entidades como modelo de coerência é viável, embora o desempenho dependa do tamanho do *corpus* de treinamento. Para que se tivesse um número maior de exemplos de treinamento, os textos com quebra de linearidade foram gerados artificialmente. Embora a geração das versões artificiais tenha buscado simular quebras de linearidade, os experimentos com textos originais mostraram que as quebras existentes nesses textos são sutis, causando uma queda no desempenho do MAC em relação aos resultados obtidos para os classificadores com validação cruzada.

Agradecimentos

A CNPq pelo apoio financeiro.

Referências

Barzilay, R. and Lapata, M. (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics*, v. 34, p. 1–34.

Burstein, J., Tetreault, J. and Andreyev, S. (2010) Using entity-based features to model coherence in student essays. In: Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, California, p. 681–684.

Castro Jorge, M.L.R., Dias, M.S. and Pardo, T.A.S. (2014). Building a Language Model for Local Coherence in Multi-document Summaries using a Discourse-enriched Entity-based Model. In: *Proceedings of the Brazilian Conference on Intelligent Systems*, São Carlos, SP, p. 44 - 49.

Elsner, M. and Charniak, E. (2011) Extending the entity grid with entity-specific features. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: shortpapers*, Portland, Oregon, p. 125–129.

Feltrim, V. D., Teufel, S., Nunes, M. G. V. and Aluísio, S. M. (2006) Argumentative zoning applied to criquing novices scientific abstracts. In: Shanahan, J. G.; Qu, Y.; Wiebe, J., eds. *Computing Attitude and Affect in Text: Theory and Applications*, Dordrecht, The Netherlands, p. 233–246.

Freitas, A. R. P. (2013). Análise automática de coerência usando o modelo grade de entidades para o português. *Dissertação de mestrado*, Universidade Estadual de Maringá, 85p.

Freitas, A.R.P. and Feltrim, V.D. (2014) Usando Grades de Entidades na Análise Automática de Coerência Local em Textos Científicos. *Linguamática*, v.6, n.1, p 29-40.

Landauer, T., Foltz, P. and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, v. 25, p. 259–284.

Souza, V. M. A. and Feltrim (2013). A coherence analysis module for SciPo: providing suggestions for scientific abstracts written in Portuguese. *Journal of the Brazilian Computer Society*, 19(1), p 59-73.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann – Elsevier.