Análise Automática de Coerência Usando o Modelo Grade de Entidades para o Português

Alison R. P. Freitas, Valéria D. Feltrim

[1]Departamento de Informática – Universidade Estadual de Maringá (UEM)
Av. Colombo, 5790 – 87020-900 – Maringá – PR – Brasil

alisonrafael@ymail.com, valeria.feltrim@din.uem.br

Abstract. *In this paper we investigate the applicability of Barzilay and Lapata's (2008) entity-grid model in the evaluation of coherence in scientific abstracts written in Portuguese. More specifically, we focused on assessing whether such model could be employed in the implementation of a classifier capable of detecting linearity breaks that affect coherence. Our experimental results are close to those of the original entity-grid model for English and very similar to the results reported by related works for other languages. Results are also close to those obtained by human judges, showing that the entity-grid model can be applied in the investigated context.*

Resumo. *Este artigo apresenta os resultados de uma investigação acerca da aplicabilidade do modelo grade de entidades proposto por Barzilay e Lapata (2008) na avaliação de coerência em resumos científicos escritos em português. Mais especificamente, se buscou avaliar se tal modelo poderia ser empregado na implementação de um classificador capaz de detectar quebras de linearidade que afetam a coerência dos resumos. Os resultados experimentais se mostraram próximos aos do modelo original para a língua inglesa e semelhantes aos relatados por trabalhos relacionados para outras línguas. Os resultados também foram próximos ao obtido por juízes humanos, mostrando que o modelo grade de entidades tem potencial para ser aplicado no contexto investigado.*

1. Introdução

Para uma grande variedade de aplicações, a avaliação da coerência textual tem sido uma parte importante do processo. De modo geral, qualquer aplicação que envolva geração automática de texto em algum nível de processamento pode se beneficiar de métodos que possibilitem avaliar a coerência do texto gerado.

Uma categoria de aplicação que tem utilizado métodos de avaliação de coerência é a das ferramentas de auxílio à escrita, em especial aquelas com propósito educacional. Para a língua inglesa, são exemplos as ferramentas *Criterion* [Higgins et al. 2004, Burstein et al. 2003], *Intelligent Essay Assessor* [Landauer et al. 2003] e *Intellimetric* [Elliot 2003]. Essas ferramentas buscam avaliar a qualidade de redações escritas em inglês e para isso analisam um conjunto de aspectos relativos à qualidade do texto que inclui algum tipo de avaliação de coerência. Para a língua portuguesa, um exemplo é a ferramenta SciPo [Feltrim et al. 2006, Souza and Feltrim 2013], desenvolvida para ajudar escritores iniciantes na escrita científica, em especial na área da Ciência da Computação. Entre os vários recursos disponíveis, o SciPo possui um módulo de análise de coerência

Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology, pages 69–78,
Fortaleza, CE, Brazil, October 21–23, 2013.

que detecta potenciais problemas de coerência em resumos científicos. Atualmente, esse módulo é baseado na classificação de componentes retóricos e na similaridade semântica entre componentes medida por meio de LSA [Landauer et al. 1998]. Conforme sugerido pelos autores [Souza and Feltrim 2013], um modelo de coerência que fosse capaz de mapear o fluxo textual de forma mais refinada poderia melhorar os resultados do módulo.

Nesse contexto, este artigo apresenta os resultados da investigação acerca da aplicabilidade do modelo grade de entidades [Barzilay and Lapata 2008] na avaliação de coerência em resumos científicos escritos em português. Mais especificamente, se buscou avaliar se tal modelo poderia ser empregado na implementação de um classificador capaz de detectar quebras de linearidade que afetam a coerência dos resumos, de modo semelhante ao proposto por [Souza and Feltrim 2013], visando a futura inclusão de tal classificador no módulo de análise de coerência do sistema SciPo.

O modelo grade de entidades é apresentado na Seção 2, assim como outros trabalhos relacionados. Na Seção 3 é descrita a implementação do modelo grade de entidades para a língua portuguesa e os resultados dos experimentos de avaliação são apresentados na Seção 4. Por fim, na Seção 5 são apresentadas as conclusões deste trabalho, bem como as sugestões de trabalhos futuros.

2. Modelo Grade de Entidades

O modelo grade de entidades é baseado em uma grade (ou matriz) de entidades. Tal representação do discurso permite que propriedades relativas à coerência local, semelhantes às definidas pela Teoria de *Centering* [Grosz et al. 1995], sejam aprendidas. A teoria de *centering* preconiza que em um texto coerente o foco de atenção (uma entidade) tende a se manter em sentenças adjacentes e que certos tipos de transições entre focos de atenção são preferíveis a outros. O modelo grade de entidades generaliza essa teoria, modelando na grade todas as transições de todas as entidades de um texto e, posteriormente, calculando a probabilidade de cada tipo de transição. Como na teoria de *centering*, o modelo grade de entidades assume que as entidades mais relevantes do discurso aparecerão em funções sintáticas importantes, como sujeito e objeto. Desse modo, o modelo permite que padrões de transições característicos de textos coerentes/incoerentes sejam aprendidos.

Cada texto é representado por uma grade em que as linhas correspondem às sentenças e as colunas às entidades. Os sintagmas nominais constituem as entidades e sintagmas nominais correferentes representam uma única entidade. Para cada entidade, as células correspondentes da grade contêm informações sobre sua presença/ausência na sequência de sentenças, bem como informações sobre as suas funções sintáticas. Dessa forma, cada célula da grade é preenchida com uma letra representando se a entidade em questão aparece na função de sujeito (S), objeto (O) ou nenhuma das anteriores (X). A ausência de uma entidade na sentença é sinalizada pelo símbolo ($-$). A Figura 1(b) mostra a grade de entidades gerada para o texto de duas sentenças mostrado em (a).

Uma transição é uma sequência $\{S, O, X, -\}_n$ que representa as ocorrências de uma entidade e suas funções sintáticas em n sentenças adjacentes. As transições podem ser obtidas a partir da grade de entidades como subsequências contínuas de cada coluna e possuem certa probabilidade de ocorrência na grade. Dessa forma, cada texto pode ser representado por um conjunto fixo de transições e suas probabilidades, usando a notação padrão de vetor de características. A Figura 2 exemplifica o vetor de características para

dois documentos d_1 e d_2 considerando-se as transições de tamanho dois.

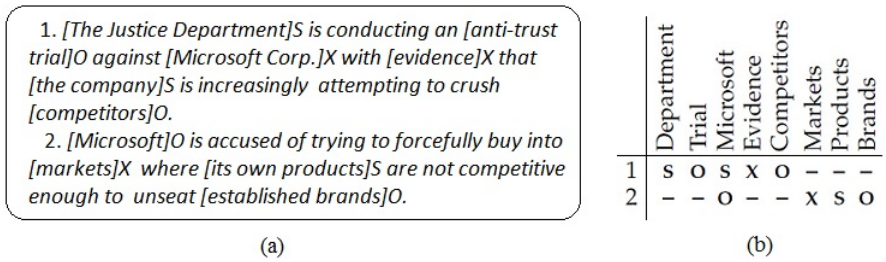

Figura 1. (a) Exemplo de sentenças com anotações sintáticas e (b) grade de entidades correspondente (adaptado de [Barzilay and Lapata 2008]).

	ss	so	sx	s-	os	oo	ox	o-	xs	xo	xx	x-	-s	-o	-x	- -
d_1	.01	.01	0	.08	.01	0	0	.09	0	0	0	.03	.05	.07	.03	.59
d_2	.02	.01	.01	.02	0	.07	0	.02	.14	.14	.06	.04	.03	.07	0.1	.36

Figura 2. Vetores de características para transições de tamanho dois dadas as categorias sintáticas {S, O, X, -} [Barzilay and Lapata 2008].

Outro aspecto incorporado ao modelo é a saliência. A saliência de uma entidade é definida com base na frequência de ocorrência da entidade no texto. Por exemplo, entidades mencionadas duas ou mais vezes são consideradas salientes. A partir dessa informação, as probabilidades das transições podem ser calculadas separadamente para cada classe de saliência.

Vários trabalhos buscaram estender o modelo grade de entidades. [Filippova and Strube 2007] modificaram o processo de seleção de entidades correferentes usando medidas de similaridade semântica em vez de resolução de correferência. Os experimentos foram realizados com textos jornalísticos escritos em alemão. [Yokono and Okumura 2010] estenderam o modelo original visando sua aplicação para a língua japonesa por meio da adição de atributos baseados em mecanismos coesivos. A representação das entidades na grade por meio de funções sintáticas também foi refinada pela adição de marcadores de tópico específicos da língua japonesa. Os experimentos foram realizados com textos jornalísticos escritos em japonês. [Burstein et al. 2010] combinaram o modelo grade de entidades com atributos relacionados à qualidade de escrita, como erros gramaticais, uso de vocabulário e estilo, visando aplicar o modelo em redações escritas em inglês por estudantes de perfis variados. Também foram utilizados atributos do tipo *Type/Token* para medir a variedade léxica das entidades que ocorrem em cada função sintática. [Elsner and Charniak 2011] estenderam o modelo por meio da adição de atributos que buscam distinguir entre entidades mais ou menos importantes, mapeando características relacionadas à proeminência no discurso, tipo de entidade nomeada e correferência. Também modificaram o processo de identificação de entidades, reconhecendo todo substantivo ou nome próprio como uma entidade em vez de usar apenas os núcleos dos sintagmas nominais. Os experimentos foram realizados com textos jornalísticos escritos em inglês. [Lin et al. 2011] combinaram o modelo grade de entidades com relações discursivas semelhantes as da RST [Mann and Thompson 1988]. Desse modo, em vez de serem representadas na grade por suas funções sintáticas, as entidades são representadas pela relação retórica em que aparecem. Os experimentos foram realizados com textos jornalísticos escritos em inglês.

3. Modelo Grade de Entidades para o Português

O modelo grade de entidades para o português foi implementado segundo a proposta original de [Barzilay and Lapata 2008]. Para extrair as entidades foi construído um sistema de pré-processamento que utiliza o *parser* PALAVRAS [Bick 2002] como ferramenta principal para a identificação dos sintagmas nominais (SNs). Processamento adicional foi realizado para desmembrar os SNs complexos identificados pelo *parser* em SNs simples, a partir dos quais as entidades puderam ser extraídas para a construção da grade de entidades. Diferente do modelo original, não foi utilizado um resolvedor automático de correferência. Neste trabalho, a identificação de entidades seguiu uma abordagem similar a de [Elsner and Charniak 2011], em que apenas sintagmas nominais que possuem o mesmo núcleo foram considerados correferentes. Para minimizar os efeitos da falta de resolução de correferência, os SNs foram lematizados e agrupados por lemas antes de serem incluídos na grade.

A partir da grade de entidades, o vetor de características é extraído de acordo com a configuração escolhida para o modelo. No modelo original, as configurações possíveis são definidas por *Correferência[+/-] Sintático[+/-] Saliência[+/-]*, representando a consideração (+) ou não (-) de tal conhecimento na construção do vetor. No caso do modelo implementado neste trabalho, como não foi utilizada resolução de correferência, as configurações só variam nos aspectos sintático e saliência, sendo portanto representadas por *Correferência- Sintático[+/-] Saliência[+/-]*.

Na configuração *Sintático+*, o vetor de características contém as probabilidades de todas as transições possíveis considerando-se as funções sintáticas S, O, X, $-$. No modelo original, o tamanho da transição é um parâmetro que pode ser ajustado conforme necessário. Neste trabalho foram consideradas apenas as transições de tamanho dois, uma vez que esse é o tamanho de transição comumente utilizado por outros trabalhos.

[Barzilay and Lapata 2008] explicam que várias classes de saliência podem ser consideradas na configuração *Saliência+*. No entanto, como o tamanho do vetor de características aumenta conforme aumenta o número de classes de saliência, é comum usar apenas duas classes: entidades salientes e não salientes. Como no modelo original, neste trabalho foram consideradas salientes as entidades mencionadas duas ou mais vezes. Assim, na configuração *Saliência+*, são construídas duas grades – uma para entidades salientes e outra para entidades não salientes. As probabilidades das transições são computadas separadamente para cada grade e depois incluídas no vetor de características.

Além dos atributos previstos no modelo original, neste trabalho também foram extraídos atributos do tipo *Type/Token* (TT) no mesmo formato utilizado por [Burstein et al. 2010], que buscam medir a variedade léxica das entidades que ocorrem em cada função sintática. Quando a configuração é *Sintático+*, quatro atributos TT são calculados: um para cada função sintática (S_TT, O_TT, X_TT), mais um para a combinação das três funções (SOX_TT). O atributo S_TT representa a proporção de entidades que aparecem como sujeito (S) em relação ao número total de sujeitos observados na grade de entidades. O mesmo tipo de proporção é calculada para as outras funções sintáticas e para a combinação de todas as funções. Quando a configuração é *Sintático-*, apenas um atributo TT é calculado, representando o número de entidades diferentes na grade dividido pelo número de ocorrências das entidades nas sentenças.

4. Experimentos e Resultados

Para avaliar o modelo grade de entidades para português foram realizados dois tipos de experimentos: (1) um experimento de ordenação de sentenças usando um córpus jornalístico e (2) um experimento de classificação baseado no julgamento de juízes humanos usando um córpus de resumos científicos. O experimento (1) buscou replicar os experimentos realizados para outras línguas. O objetivo, nesse caso, foi validar a implementação feita neste trabalho, bem como avaliar se o comportamento do modelo aplicado a língua portuguesa é semelhante ao observado para outras línguas. O experimento (2) buscou avaliar o desempenho do modelo no contexto de um classificador capaz de detectar problemas de coerência em resumos científicos, uma vez que a motivação para este estudo está na melhoria do módulo de análise de coerência da ferramenta SciPo.

4.1. Experimento 1: Ordenação de Sentenças

Nesse experimento foi utilizado um córpus de 286 textos jornalísticos extraídos dos córpora CSTNews [Cardoso et al. 2011] (136), Summ-it [Collovini et al. 2007] (50) e Temário [Rino and Pardo 2007] (100). A preparação do córpus seguiu o mesmo procedimento descrito em [Barzilay and Lapata 2008]. Para cada texto foram geradas aproximadamente 20 versões sintéticas em que a ordem original das sentenças foi permutada aleatoriamente e assumiu-se que o texto com as sentenças na ordem original deve ser mais coerente que a maioria dos textos com as sentenças permutadas. Desse modo, os textos originais foram marcados como "sem problemas" de coerência e as versões permutadas como "com problemas". Como resultado foi obtido um conjunto de 5.720 pares $\{$*texto_original, versão_permutada*$\}$ (286 textos $\times$ 20 versões permutadas), que foi separado aleatoriamente em conjuntos de treinamento e teste, sendo 2/3 para treinamento e 1/3 para teste.

Como em [Barzilay and Lapata 2008], a ordenação de sentenças foi tratada como um problema de ranqueamento, em que o modelo é usado para ranquear versões de um mesmo texto, esperando que a versão mais coerente fique no topo do *ranquing*. Desse modo, para treinar e testar o modelo foi utilizado o sistema SVM^{rank} [Joachims 2006], que implementa o algoritmo SVM (*Support Vector Machine*) para problemas de ranqueamento. Como *baseline* foi utilizado um modelo baseado em LSA semelhante ao implementado por [Barzilay and Lapata 2008]. Para o cálculo da LSA foi utilizada a implementação de [Souza and Feltrim 2013] e todos os córpora empregados neste trabalho foram utilizados na criação do espaço semântico.

A métrica de avaliação seguiu a de [Barzilay and Lapata 2008], em que dadas todas as comparações entre pares, a acurácia é medida como a quantidade de predições corretas feitas pelo modelo dividida pelo número de pares existentes no conjunto de teste. Na Tabela 1 é mostrado o percentual de acertos da *baseline* e do modelo grade de entidades com os atributos *Type/Token*, representado na tabela por suas quatro configurações possíveis. Como os textos jornalísticos são provenientes de córpora diferentes, os resultados são mostrados considerando-se o córpus de origem dos textos originais, além dos resultados calculados para o córpus jornalístico como um todo (coluna "Todos juntos").

Conforme pode ser observado na Tabela 1, o modelo grade de entidades superou a *baseline* em todos os casos, com exceção do córpus Temário, em que a *baseline* foi superior em 4%. De fato, os textos do córpus Temário são maiores do que os textos dos

Tabela 1. Percentual de acertos da *baseline* (LSA) e do modelo grade de entidades para o Experimento 1.

Modelo	Cstnews	Summit	Temário	Todos juntos
LSA	61,429%	56,000%	**79,000%**	67,000%
Sintático+ Saliência-	64,000%	48,235%	60,455%	62,105%
Sintático+ Saliência+	**74,444%**	50,294%	59,242%	58,105%
Sintático- Saliência-	69,444%	63,824%	**74,848%**	**68,579%**
Sintático- Saliência+	70,889%	**72,059%**	65,455%	67,368%

outros dois córpora (média de sentenças por texto: CSTNews e Summit $\approx$ 16; Temário $\approx$ 29) e isso pode ter influenciado o resultado da *baseline*, que é baseada na média de similaridade entre pares de sentenças. Com relação a variação na configuração *Sintático[+/-] Saliência[+/-]*, observa-se que não existe um padrão de resultados que permita julgamento sobre a melhor configuração. Esse mesmo comportamento pode ser observado nos trabalhos relacionados e no modelo original. No geral, os resultados obtidos pelo modelo grade de entidades para o português são semelhantes aos relatados pelos trabalhos desenvolvidos para outras línguas, ficando abaixo apenas dos resultados obtidos pelo modelo original. A Tabela 2 apresenta um resumo dos melhores resultados relatados pelos trabalhos relacionados considerando-se sempre a configuração *Correferência-*.

Tabela 2. Resumo dos resultados obtidos por trabalhos relacionados.

Trabalho	Córpus de origem	Tamanho	% de acerto
[Barzilay and Lapata 2008]	*Associated Press articles from the North American News Corpus on the topic of earthquakes*	100 textos originais	83%
	Narratives from the National Transportation Safety Boards aviation accident database	100 textos originais	89,9%
[Elsner and Charniak 2011]	*WSJ articles from the Penn Treebank*	1004 textos originais	84%
[Filippova and Strube 2007]	*German corpus of newpaper articles*	100 textos originais	69%
[Yokono and Okumura 2010]	*Articles in 2003 from Asahi newspaper corpus*	100 textos originais	59,4%
		300 textos originais	77,3%

4.2. Experimento 2: Classificação Baseada no Julgamento Humano

Nesse experimento foi utilizado um córpus de 139 resumos científicos: 99 resumos extraídos do córpus de resumos de Trabalhos de Conclusão de Curso em Ciência da Computação compilado por [Souza and Feltrim 2013], e 40 resumos experimentais coletados diretamente com os autores – alunos formandos do curso de graduação em Ciência da Computação da Universidade Estadual de Maringá. O córpus foi preparado para experimentos utilizando o julgamento humano acerca do nível de coerência dos resumos nos mesmos moldes do trabalho de [Burstein et al. 2010]. Para a anotação manual, os anotadores foram instruídos a marcar o resumo como "com problemas", caso fossem encontradas barreiras na leitura (por exemplo, sentenças "soltas", mudanças bruscas no foco de atenção), caracterizando a quebra de linearidade; caso contrário, os anotadores foram instruídos a marcar o resumo como "sem problemas".

A concordância entre anotadores foi avaliada por meio de um experimento em que dois anotadores treinados anotaram separadamente os 40 resumos experimentais. A concordância medida por meio da medida *Kappa* foi de 0,70%, valor próximo ao obtido por [Burstein et al. 2010] (K = 0,68%) em experimento semelhante. O restante do córpus foi anotado por apenas um dos anotadores treinados no experimento. No total, a anotação manual resultou em 117 (84%) resumos marcados como "sem problemas" e 22 (16%)

como "com problemas". Vale ressaltar que esse desbalanceamento é característico de córpora manualmente anotados e que o mesmo nível de desbalanceamento foi observado nos córpora de redações utilizados por [Burstein et al. 2010].

Nesse experimento a tarefa foi modelada como um problema de classificação binária. O treinamento e teste do modelo foi realizado no ambiente WEKA [Witten and Frank 2005] utilizando-se os algoritmos SMO, J48 e *Naïve Bayes*. A escolha do algoritmo SMO se deu por ele ser uma implementação de SVM, que é o algoritmo utilizado no Experimento 1; o J48 foi escolhido por ser uma implementação do C4.5, que é o algoritmo de aprendizado utilizado por [Burstein et al. 2010]; e o *Naïve Bayes* foi escolhido por ser um algoritmo simples, rápido e de larga utilização em tarefas que envolvem classificação textual. Os resultados foram calculados aplicando-se 10-*fold cross-validation* ao córpus de 139 resumos. Os resultados em termos das medidas *F-measure* e *Kappa* são mostrados para cada algoritmo de aprendizado e configuração do modelo na Tabela 3. Os valores de *F-measure* representam a média das *F-measures* calculadas para as duas classes, ponderada pelo número de exemplos de cada classe. Os resultados listados como TT+ foram calculados adicionando-se ao modelo os atributos do tipo *Type/Token*. Os resultados listados como TT- foram calculados utilizando o modelo grade de entidades original.

Tabela 3. Resultados do modelo grade de entidades para o experimento 2.

	Naïve Bayes		**SMO**		**J48**	
TT-	**F-meas.**	**Kappa**	**F-meas.**	**Kappa**	**F-meas.**	**Kappa**
Sintático+ Saliência-	0,663	0,211	0,769	0,000	0,810	0,256
Sintático+ Saliência+	0,741	0,053	0,802	0,144	0,882	0,515
Sintático- Saliência-	0,707	0,211	0,769	0,000	0,804	0,183
Sintático- Saliência+	0,799	0,168	0,766	-0,014	**0,910**	**0,650**
TT+						
Sintático+ Saliência-	0,731	0,262	0,766	-0,014	0,809	0,271
Sintático+ Saliência+	0,770	0,114	0,802	0,144	0,876	0,494
Sintático- Saliência-	0,740	0,223	0,769	0,000	0,804	0,183
Sintático- Saliência+	0,799	0,168	0,797	0,127	**0,910**	**0,650**

Conforme pode ser observado na Tabela 3, os melhores resultados foram obtidos com o algoritmo J48, sendo que o melhor resultado ($K = 0,65$) se aproxima do valor obtido pelos juízes humanos ($K = 0,70$) e ultrapassa o melhor sistema de [Burstein et al. 2010] ($K = 0,61$). Também é possível observar que enquanto os valores de *F-measure* são relativamente altos (acima de 0,8 para o J48), os valores da medida *Kappa* são mais baixos e apresentam maior variação entre os diferentes modelos. Isso pode ser atribuído ao forte desbalanceamento do córpus (84%/16%), que eleva o desempenho dos classificadores para a classe majoritária, elevando por consequência os valores de *F-measure*. A medida *Kappa*, por sua vez, prioriza os acertos para a classe minoritária, em que a probabilidade de acerto "ao acaso" é menor, fornecendo assim uma medida mais realista do desempenho do classificador nesse contexto de desbalanceamento.

Analisando as diferentes configurações do modelo com base no algoritmo J48, fica evidente a contribuição do aspecto saliência. O melhor resultado ($K = 0,65$) foi obtido com a configuração *Sintático- Saliência+* e o segundo melhor ($K = 0,52$) com a configuração *Sintático+ Saliência+*. Curiosamente, neste caso, o modelo mais simples, que não considera a função sintática das entidades (*Sintático-*), se saiu melhor do que o modelo mais rico (*Sintático+*). De fato, esse comportamento também foi observado

por [Filippova and Strube 2007] e pode ser atribuído ao tamanho reduzido do córpus de treinamento. Uma vez que a configuração *Sintático+* gera um vetor de características quatro vezes maior que a configuração *Sintático-*, um número maior de exemplos de treinamento pode ser necessário para que o modelo possa se beneficiar das informações relativas ao aspecto sintático. Quanto aos atributos *Type/Token* (TT), observa-se que a sua inclusão no modelo teve pouca influência nos resultados, sendo que, em alguns casos, os valores com TT+ permaneceram iguais aos valores com TT-.

Na Tabela 4, os resultados obtidos com o melhor modelo (*Sintático- Saliência+* treinado com J48) são detalhados por classe e comparados com os obtidos por uma *baseline* simples que classifica todos os textos como "sem problemas". Como pode ser observado, o modelo é superior a *baseline* para as duas classes.

Tabela 4. Melhor modelo (*Sintático- Saliência+* treinado com J48) vs. *baseline*.

	Sem Problema (117)			Com problema (22)			Média ponderada (139)		
	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
Melhor modelo	0,934	0,966	0,950	0,778	0,636	0,700	0,909	0,914	0,910
Baseline	0,842	1,000	0,914	0,000	0,000	0,000	0,708	0,841	0,769

5. Conclusões e Trabalhos Futuros

Este trabalho teve por objetivo implementar e avaliar o modelo grade de entidades proposto por [Barzilay and Lapata 2008] para a língua portuguesa, visando sua aplicação na avaliação de coerência em resumos científicos. A motivação está em encontrar um modelo de coerência capaz de mapear o fluxo textual de forma mais refinada do que o modelo baseado em LSA proposto por [Souza and Feltrim 2013], visando melhorar os resultados obtidos no âmbito da detecção de quebras de linearidades entre sentenças adjacentes de um resumo. O modelo grade de entidades para o português foi implementado segundo a proposta original, com exceção do tratamento de correferências, que não foi realizado neste trabalho. Além dos atributos previstos no modelo original, neste trabalho também foram avaliados atributos do tipo *Type/Token* de forma similar a realizada por [Burstein et al. 2010].

A avaliação do modelo foi feita de dois modos. Primeiramente buscou-se reproduzir o mesmo cenário de testes empregado pelos trabalhos encontrados na literatura. Isso permitiu a comparação dos resultados deste trabalho com os obtidos para outras línguas, mostrando que os resultados são próximos aos relatados para outras línguas. Em um segundo momento buscou-se avaliar o modelo na tarefa de avaliação de coerência em resumos científicos. Os resultados mostraram que o uso do modelo grade de entidades é viável nesse contexto. O melhor resultado ($K = 0,65$), alcançado com o algoritmo J48 com o modelo na configuração *Sintático- Saliência+*, é próximo ao obtido por juízes humanos ($K = 0,70$) e superior ao relatado por [Burstein et al. 2010] para o seu melhor sistema ($K = 0,61$).

Um desdobramento natural deste trabalho é a aplicação efetiva do modelo grade de entidades no módulo de análise de coerência do sistema SciPo, possibilitando a avaliação extrínseca do modelo no contexto de uma ferramenta de auxílio à escrita científica. Também pretende-se avaliar o modelo no contexto de outras aplicações que possam se beneficiar de um modelo de coerência, como é o caso da sumarização automática. Outra linha de trabalhos futuros aborda a melhoria dos resultados obtidos

por meio da combinação do modelo original com conhecimentos provenientes de outras fontes, como os índices calculados pela Coh-Metrix-Port [Scarton and Aluísio 2010, Scarton et al. 2009]. A inclusão de um sistema de resolução automática de correferência no modelo atual também será explorada em trabalhos futuros, já que os melhores resultados da literatura foram obtidos utilizando-se esse tipo de conhecimento.

References

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

Bick, E. (2002). *The Parsing System Palavras - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Famework*. PhD thesis, Department of Linguistics – Aarhus: Aarhus University Press – DK.

Burstein, J., Chodorow, M., and Leacock, C. (2003). Criterion online essay evaluation: An application for automated evaluation of student essays. In *Proceedings of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence*, pages 3–10.

Burstein, J., Tetreault, J., and Andreyev, S. (2010). Using entity-based features to model coherence in student essays. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 681–684.

Cardoso, P., Maziero, E., Jorge, M., Seno, E., Di Felippo, A., Rino, L., Nunes, M., and Pardo, T. (2011). Cstnews - a discourse-annotated corpus for single and multi-document summarization of news texts in brazilian portuguese. In *Proceedings of the 3rd RST Brazilian Meeting*, pages 88–105, Cuiabá/MT, Brazil.

Collovini, S., Carbonel, T. I., Fuchs, J. T., Coelho, J. C., Rino, L., and Vieira, R. (2007). Summ-it: um corpus anotado com informações discursivas visando sumarização automática. In *Anais do XXVII Congresso da SBC: V Workshop em Tecnologia da Informação e da Linguagem Humana (TIL 2007)*.

Elliot, S. (2003). Intellimetric: From here to validity. In *Shermis, M.; Burstein, J., eds. Automatic Essay Scoring: A Cross-Disciplinary Perspective.*, pages 71–86, Hillsdale, NJ. Lawrence Erlbaum Associates.

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: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 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., and Wiebe, J., editors, *Computing Attitude and Affect in Text: Theory and Applications*, pages 233–246, Dordrecht, The Netherlands. Springer.

Filippova, K. and Strube, M. (2007). Extending the entity-grid coherence model to semantically related entities. In *Proceedings of the Eleventh European Workshop on Natural Language Generation*, ENLG '07, pages 139–142.

Grosz, B. J., Weinstein, S., and Joshi, A. K. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21:203–225.

Higgins, D., Burstein, J., Marcu, D., and Gentile, C. (2004). Evaluating multiple aspects of coherence in student essays. In *Proceedings of the 2004 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 185–192.

Joachims, T. (2006). Training linear svms in linear time. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 217–226, New York, NY, USA. ACM.

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

Landauer, T. K., Laham, D., and Foltz, P. W. (2003). *Automated essay scoring and annotation of essays with the intelligent essay assessor.* Lawrence Erlbaum Associates, Mahwah, NJ.

Lin, Z., Ng, H. T., and Kan, M.-Y. (2011). Automatically evaluating text coherence using discourse relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 997–1006.

Mann, W. C. and Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Rino, L. H. M. and Pardo, T. A. S. (2007). *A coleção TeMário e a avaliação de sumarização automática*, volume 1. IST Press, Lisboa, Portugal.

Scarton, C. and Aluísio, S. M. (2010). Coh-metrix-port: a readability assessment tool for texts in brazilian portuguese. In *Proceedings of the 9th International Conference on Computational Processing of the Portuguese Language, Extended Activities Proceedings*, PROPOR '10. 1 CD-ROM v1.

Scarton, C. E., Almeida, D. M., and Aluísio, S. M. (2009). Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: adaptando as métricas do coh-metrix para o português. In *Proceedings of the 7th Brazilian Symposium in Information and Human Language Technology (STIL 2009)*. 1 CD-ROM v1.

Souza, V. M. A. and Feltrim, V. D. (2013). A coherence analysis module for scipo: providing suggestions for scientific abstracts written in portuguese. *Journal of the Brazilian Computer Society*, 19:59–73.

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

Yokono, H. and Okumura, M. (2010). Incorporating cohesive devices into entity grid model in evaluating local coherence of japanese text. In Gelbukh, A., editor, *Proceedings of the 11th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2010)*, number 6008 in Lecture Notes in Computer Science, pages 303–314. Springer Berlin/Heidelberg.