Usando Grades de Entidades na Análise Automática de Coerência Local em Textos Científicos

Using Entity Grids to Automatically Evaluate Local Coherence in Scientific Texts

Alison Rafael Polpeta Freitas
Universidade Estadual de Maringá
arpfreitas@gmail.com

Valéria Delisandra Feltrim
Universidade Estadual de Maringá
vfeltrim@din.uem.br

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 local 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. Nos experimentos com resumos científicos, os resultados foram próximos ao obtido por juízes humanos, mostrando que o modelo grade de entidades tem potencial para ser aplicado no contexto investigado.

Palavras chave

coerência local, modelo grade de entidades, texto científico

Abstract

In this paper we investigate the applicability of Barzilay and Lapata's (2008) entity-grid model in the evaluation of local 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 text 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. In experiments with scientific abstracts, results are close to those obtained by human judges, showing that the entity-grid model can be used in the investigated context.

Keywords

local coherence, entity-grid model, scientific writing

1 Introdução

Para uma grande variedade de aplicações na área de Processamento de Linguagem Natural, 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. Um exemplo desse tipo de aplicação é a sumarização automática.

Outra 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, Chodorow e Leacock, 2003), *Intelligent Essay Assessor* (Landauer, Laham e Foltz, 2003) e *Intellimetric* (Elliot, 2003). Essas ferramentas buscam avaliar a qualidade de redações (*essays*) escritas em inglês e para isso analisam um conjunto de aspectos relativos a qualidade do texto que inclui algum tipo de avaliação de coerência.

Para a língua portuguesa, um exemplo é o sistema SciPo (Feltrim et al., 2006), desenvolvido 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 que detecta potenciais problemas de coerência semântica em resumos científicos (Souza e Feltrim, 2013). Atualmente, esse módulo é baseado na classificação de componentes retóricos e na similaridade semântica entre componentes medida por meio de *Latent Semantic Analysis* – LSA – (Landauer, Foltz e Laham, 1998). Especificamente, três tipos de relacionamentos semânticos ou dimensões são examinados: (1) Dimensão Título: verifica o relacionamento semântico entre o título do resumo e o componente Propósito; (2) Dimensão Propósito: verifica o relacionamento

semântico entre o componente Propósito e os componentes Metodologia, Resultado e Conclusão; e (3) Dimensão Lacuna-Contexto: verifica o relacionamento semântico entre o componente Lacuna e o componente Contexto.

Souza e Feltrim (2011) propõem ainda uma quarta dimensão, chamada Quebra de Linearidade, em que se verifica a existência de uma "quebra" entre sentenças adjacentes do resumo, que se caracteriza pela dificuldade em se estabelecer uma ligação clara da sentença atual com a sentença anterior, demandando maior esforço cognitivo para a interpretação do texto. No entanto, os resultados obtidos para essa dimensão foram pouco satisfatórios, mostrando que a LSA não foi capaz de capturar as quebras de linearidade por vezes sutis observadas no córpus de resumos científicos utilizado pelos autores. Conforme sugerido por Souza e Feltrim (2013), um modelo de coerência que fosse capaz de mapear o fluxo textual de forma mais refinada, como a grade de entidades proposta por Barzilay e Lapata (2008), poderia obter melhores resultados para essa dimensão.

Nesse contexto, este trabalho investigou a aplicabilidade do modelo grade de entidades (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, de modo semelhante ao proposto por Souza e Feltrim (2013), visando a futura inclusão de tal classificador no módulo de análise de coerência do sistema SciPo.

Dois tipos de experimentos foram realizados. Primeiramente foram feitos experimentos com um córpus jornalístico em português, visando a comparação, ainda que indireta, com outros trabalhos que utilizam o modelo grade de entidades. Para avaliar o desempenho do modelo com textos científicos foram feitos experimentos com um córpus de resumos científicos escritos em português.

Os resultados experimentais com o córpus jornalístico se mostraram próximos aos do modelo original para a língua inglesa e semelhantes aos relatados por trabalhos relacionados para outras línguas. Nos experimentos com resumos científicos, os resultados obtidos com o modelo foram próximos ao obtido por dois juízes humanos, mostrando que o modelo tem potencial para ser aplicado no contexto do sistema SciPo.

O restante deste artigo está organizado da seguinte forma: 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

Como explicitado pelo próprio nome, o modelo grade de entidades é baseado em uma grade (ou matriz) de entidades e busca aprender propriedades relativas à coerência local semelhantes às definidas pela Teoria de *Centering* (Grosz, Weinstein e Joshi, 1995). 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 uma probabilidade para 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 seria capaz de apreender padrões de transições característicos de textos coerentes/incoerentes.

Cada texto é representado por uma grade em que as linhas correspondem às sentenças e as colunas às entidades. Por entidade se entende uma classe de sintagmas nominais correferentes. 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 uma certa probabili-

dade 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.

1. *[The Justice Department]S is conducting an [anti-trust trial]O against [Microsoft Corp.]X with [evidence]X that [the company]S is increasingle attempting to crush [competitors]O.*
2. *[Microsoft]O is accused of trying to forcefully buy into [markets]X where [its own products]S are not competitive enough to unseat [established brands]O.*

(a)

	Department	Trial	Microsoft	Evidence	Competitors	Markets	Products	Brands
1	S	O	S	X	O	-	-	-
2	-	-	O	-	-	X	S	O

(b)

Figura 1: (a) Exemplo de sentenças com anotações sintáticas e (b) grade de entidades correspondente – adaptado de Barzilay e 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, uma grade apenas com entidades salientes pode ser construída e as probabilidades das transições podem ser calculadas separadamente para cada classe de saliência.

Vários trabalhos buscaram estender o modelo grade de entidades. Filippova e 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 e 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, Tetreault e Andreyev (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 (*essays*) 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 e Charniak (2011) estenderam o modelo por meio da adição de atributos entidade-específicos que buscam distinguir entre entidades importantes e entidades menos importantes. 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, Ng e Kan (2011) combinaram o modelo grade de entidades com relações discursivas semelhantes as da RST (Mann e 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 e 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. Diferentemente do modelo original, não foi utilizado um resolvedor automático de correferência. Neste trabalho, a identificação de entidades seguiu uma abordagem similar a de Elsner e Charniak (2011), em que apenas sintagmas nominais que possuem o mesmo núcleo são considerados correferentes. Adicionalmente, para diminuir a duplicação de entidades, os SNs foram lematizados e agrupados por lemas antes de serem incluídos na grade. Uma visão geral das etapas de processamento para a construção da grade de entidades e extração do vetor de característica é mostrada na Figura 3.

	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 e Lapata, 2008).

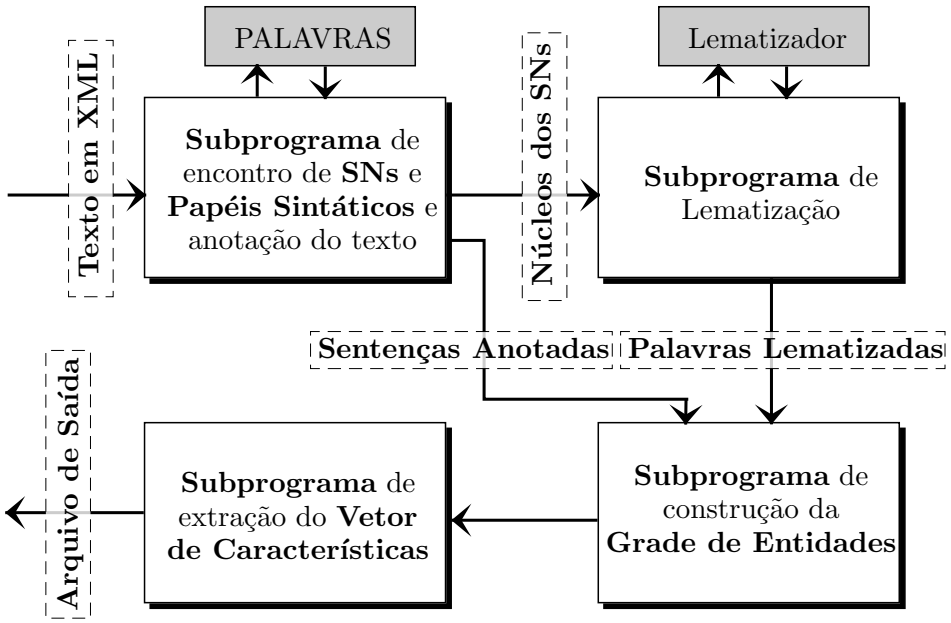

Figura 3: Etapas de processamento para a construção da grade de entidades e extração do vetor de características.

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 empregada resolução de correferência, as configurações só variam nos aspectos sintático e saliência, sendo portanto representadas por `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 e 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+`, a grade de entidades é dividida em duas – 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) semelhantes aos utilizados por Burstein, Tetreault e Andreyev (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 o 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, mais especificamente para o inglês (Barzilay e Lapata, 2008; Elsner e Charniak, 2011), para o alemão (Filippova e Strube, 2007) e para o japonês (Yokono e Okumura, 2010). O objetivo, nesse caso, foi validar a implementação feita neste trabalho, bem como avaliar se o comportamento do modelo aplicado à 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 local em resumos científicos, uma vez que a motivação para este estudo está na melhoria do módulo de análise de coerência da ferramenta SciPo e na implementação da Dimensão Quebra de Linearidade. Os experimentos (1) e (2) e os respectivos resultados são apresentados a seguir, nas Seções 4.1 e 4.2, respectivamente.

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 textos –, Summ-it (Collovini et al., 2007) – 50 textos – e Temário (Rino e Pardo, 2007) – 100 textos. A preparação do córpus seguiu o mesmo procedimento descrito em Barzilay e 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 × 20 versões permutadas), que foi separado aleatoriamente em conjuntos de treinamento (2/3 dos pares) e teste (1/3 dos pares). A descrição completa desse córpus está disponível em Freitas (2013).

Como em Barzilay e Lapata (2008), a ordenação de sentenças foi tratada como um problema de ranqueamento, em que o modelo é usado para ranquear diferentes versões de um mesmo texto, esperando que as versões mais coerentes fiquem no topo do *ranking*. 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 e Lapata (2008). Esse modelo estima um valor de coerência V para um texto T por meio da média dos valores de similaridade para todos os pares de sentenças de T. Para o cálculo da LSA foi utilizada a implementação de Souza e Feltrim (2013) e os dois córpus compilados para este trabalho foram utilizados na criação do espaço semântico.

A métrica de avaliação seguiu a de Barzilay e 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* (LSA) e do modelo grade de entidades com os atributos *Type/Token*, representado na tabela por suas quatro configurações possíveis (`Sintático[+/-] Saliência[+/-]`). Como os textos jornalísticos são provenientes de córpus 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"). Os melhores resultados obtidos por cada modelo estão destacados em negrito.

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 outros dois córpus (média de sentenças por texto: CSTNews e Summit ≈ 16; Temário ≈ 29) e isso pode ter influenciado o resultado da *baseline*, que é baseada na média de similaridade entre pares de sentenças. Embora não seja possível uma comparação direta, no

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 %

Tabela 1: Percentual de acertos da *baseline* e do modelo grade de entidades para o experimento 1.

Trabalho	Língua	Córpus		Melhor resultado
Barzilay e Lapata (2008)	Inglês	100 textos originais	(T)	83,0 %
		100 textos originais	(A)	89,9 %
Elsner e Charniak (2011)	Inglês	1004 textos originais		84,0 %
Filippova e Strube (2007)	Alemão	100 textos originais		69,0 %
Yokono e Okumura (2010)	Japonês	100 textos originais		59,4 %
		300 textos originais		77,3 %

Tabela 2: Resumo dos resultados obtidos por trabalhos relacionados.

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-`.

Ainda na Tabela 1, vale observar que a variação na configuração `Sintático[+/-] Saliência[+/-]` não resultou em um padrão de resultados que permitisse a julgar sobre a melhor configuração, variando conforme o córpus utilizado. Esse mesmo comportamento pode ser observado nos trabalhos relacionados e no modelo original.

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 e 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, Tetreault e Andreyev (2010). Para a anotação manual, os anotadores foram instruídos a marcar o resumo como "com problemas", caso fossem encontradas barreiras na leitura (por exemplo, dificuldade de se estabelecer uma ligação semântica entre sentenças), 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 estatística *Kappa* foi de 0,70%, valor próximo ao obtido por Burstein, Tetreault e Andreyev (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 córpus manualmente anotados e que o mesmo nível de desbalanceamento foi observado nos córpus de redações (*essays*) utilizados por Burstein, Tetreault e Andreyev (2010).

Nesse experimento, a tarefa foi modelada como um problema de classificação binária. O treinamento e o teste do modelo foram realizados no ambiente WEKA (Witten e Frank, 2005) utilizando-se os seguintes algoritmos de aprendizado de máquina: SMO – *Sequential Minimal Optimization* – (Platt, 1998), uma implementação do algoritmo SVM para classificação; J48, uma implementação em Java e de código aberto do algoritmo C4.5 (Quinlan, 1993) que gera árvores de decisão; e *Naïve Bayes*, um algoritmo probabilístico baseado na regra da probabilidade condicional de *Bayes*. A escolha pelo 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, Tetreault e Andreyev, 2010); e o *Naïve Bayes* foi escolhido por ser um algoritmo de aprendizado 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* (F_1) 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 apenas o modelo grade de entidades.

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, Tetreault e Andreyev (2010) ($K = 0,61$) em experimento semelhante. O algoritmo SMO apresentou os piores resultados. Também é possível observar que enquanto os valores de *F-measure* são relativamente altos (acima de 0,8 para o algoritmo 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 induzidos para a classe majoritária ("sem problema"), elevando por consequência os valores da medida *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. Os resultados completos, detalhados por algoritmo de aprendizado e por classe, expressos nas cinco medidas de avaliação utilizadas, estão disponíveis em Freitas (2013).

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 e 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 vetor de características teve pouca influência nos resultados, sendo que, em alguns casos, os valores com TT+ permaneceram iguais aos valores com TT−. Uma discreta contribuição dos atributos TT pode ser notada nos resultados calculados com o algoritmo *Naïve Bayes*.

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.

Para avaliar o efeito do desbalanceamento do córpus nos resultados, os experimentos com os três algoritmos de aprendizado foram refeitos utilizando-se a técnica de balanceamento SMOTE (Chawla et al., 2002) – *Synthetic Minority Oversampling Technique* – (Chawla et al., 2002), também disponível no WEKA (Witten e Frank, 2005), a qual realiza *oversampling*, isto é, adiciona ao conjunto novos casos da classe minoritária gerados sinteticamente a partir de casos já existentes. Esses novos casos são gerados na vizinhança de cada caso da classe minoritária. Segundo Chawla et al. (2002), esse método produz resultados melhores do que a simples replicação de casos existentes, uma vez que essa prática pode levar a modelos muito específicos, prejudicando o poder de generalização do modelo (*overfitting*).

A Tabela 5 apresenta os resultados após a classe minoritária ter sido aumentada em 400%, valor que deixa o córpus com um balanceamento próximo ao perfeito. Assim como na Tabela 3, os resultados são mostrados em termos das medidas *F-measure* e *Kappa* para cada algoritmo de aprendizado e configuração do modelo grade de entidades.

Conforme pode ser observado na Tabela 5, os resultados usando *oversampling* foram melhores para os três algoritmos testados, sendo que o J48 continuou apresentando o melhor resultado, especialmente em termos da

	Naïve Bayes		SMO		J48	
TT−	F_1	**Kappa**	F_1	**Kappa**	F_1	**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**

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

	Sem Problema (117)			Com problema (22)			Média ponderada (139)		
	Precision	**Recall**	F_1	**Precision**	**Recall**	F_1	**Precision**	**Recall**	F_1
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

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

medida *Kappa*. A configuração `Sintático- Saliência+` continuou sendo a melhor e a contribuição dos atributos *_TT ficou mais evidente, elevando drasticamente o valor *Kappa* da configuração `Sintático- Saliência+` para os algoritmos *Naïve Bayes* e SMO. Vale destacar que enquanto o valor da *F-measure* ponderada teve pouca variação em relação aos valores sem *oversampling* (Tabela 3), o valor da medida *Kappa* melhorou significativamente, especialmente para os algoritmos *Naïve Bayes* e SMO. Isso se deve ao fato da medida *Kappa* refletir de forma mais apropriada o desempenho nas duas classes consideradas. A variação do valor da medida *Kappa* de acordo com o percentual de *oversampling* da classe minoritária é mostrada na Figura 4.

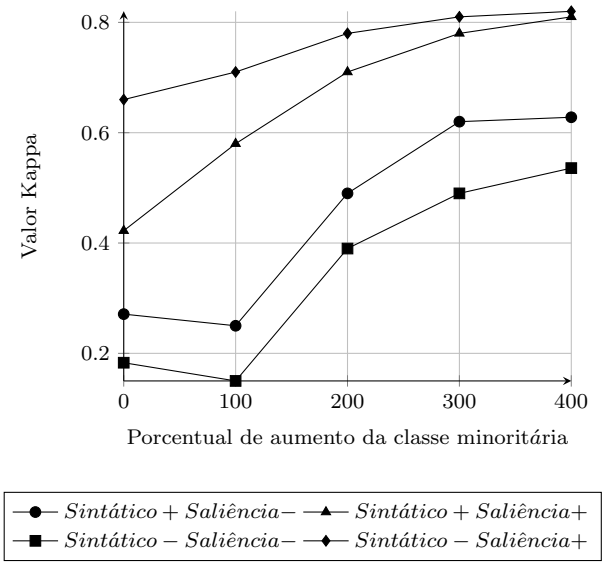

Figura 4: Variação dos valores da medida *Kappa* de acordo com o percentual de *oversampling* (SMOTE).

Quando se considera as medidas *Precision*, *Recall* e *F-measure* para cada classe, é possível notar que os resultados com *oversampling* ficaram mais uniformes, uma vez que os valores para a classe minoritária melhoraram, alcançando valores semelhantes aos obtidos para a classe majoritária. A Tabela 6 mostra os resultados em termos de *Precision*, *Recall* e *F-measure* para cada classe. Os valores sem *oversampling* são mostrados na primeira metade da tabela e os valores com *oversampling* são mostrados na segunda parte.

Com o objetivo de observar a influência do tamanho do córpus nos resultados foi realizado um experimento em que se aumentou artificialmente e gradativamente o tamanho do córpus. Para isso, foi utilizado novamente o algoritmo SMOTE sobre o córpus já balanceado pelo *oversampling*. Nesse caso, como o córpus já estava balanceado, a cada execução o SMOTE selecionava aleatoriamente uma das classes para a criação de novos exemplos. Dessa forma, para cada vez que o tamanho foi aumentado, o algoritmo foi aplicado duas vezes para que o balanceamento fosse mantido.

Os resultados desse experimento foram calculados para dois cenários usados nos experimentos anteriores: (1) o modelo na configuração `Sintático- Saliência+ *_TT+` treinado e testado com o J48, por ser o cenário que apresentou os melhores resultados, e (2) o modelo na configuração `Sintático- Saliência+ *_TT−` treinado e testado com o SMO, por ser o cenário que

	Naïve Bayes		**SMO**		**J48**	
***_TT−**	F_1	**Kappa**	F_1	**Kappa**	F_1	**Kappa**
Sintático+ Saliência−	0,718	0,460	0,725	0,478	0,806	0,612
Sintático+ Saliência+	0,762	0,525	0,830	0,663	0,904	0,808
Sintático− Saliência−	0,631	0,294	0,670	0,383	0,769	0,544
Sintático− Saliência+	0,615	0,290	0,587	0,253	0,912	0,824
***_TT+**						
Sintático+ Saliência−	0,772	0,554	0,832	0,667	0,806	0,612
Sintático+ Saliência+	0,797	0,596	0,802	0,144	0,904	0,808
Sintático− Saliência−	0,706	0,433	0,729	0,466	0,764	0,536
Sintático− Saliência+	0,890	0,780	0,868	0,735	0,916	0,833

Tabela 5: Resultados do modelo grade de entidades para o córpus de resumos científicos balanceado com SMOTE em termos de *F-measure* e *Kappa*.

Sem *oversampling*

	Sem Problema (117)			**Com problema (22)**		
*_TT−	**Precision**	**Recall**	F_1	**Precision**	**Recall**	F_1
Sintático+ Saliência−	0,877	0,915	0,895	0,412	0,318	0,359
Sintático+ Saliência+	0,906	0,975	0,939	0,769	0,455	0,571
Sintático− Saliência−	0,862	0,957	0,907	0,444	0,182	0,258
Sintático− Saliência+	0,934	0,966	0,950	0,778	0,636	0,700
*_TT+						
Sintático+ Saliência−	0,882	0,897	0,890	0,400	0,364	0,381
Sintático+ Saliência+	0,906	0,966	0,935	0,714	0,455	0,556
Sintático− Saliência−	0,862	0,957	0,907	0,444	0,182	0,258
Sintático− Saliência+	0,934	0,966	0,950	0,778	0,636	0,700

Com *oversampling* (SMOTE)

	Sem Problema (117)			**Com problema (110)**		
*_TT−	**Precision**	**Recall**	F_1	**Precision**	**Recall**	F_1
Sintático+ Saliência−	0,812	0,812	0,812	0,800	0,800	0,800
Sintático+ Saliência+	0,915	0,899	0,907	0,893	0,909	0,901
Sintático− Saliência−	0,849	0,675	0,752	0,716	0,873	0,787
Sintático− Saliência+	0,929	0,897	0,913	0,895	0,927	0,911
*_TT+						
Sintático+ Saliência−	0,807	0,821	0,814	0,806	0,791	0,798
Sintático+ Saliência+	0,915	0,899	0,907	0,893	0,909	0,901
Sintático− Saliência−	0,848	0,667	0,746	0,711	0,873	0,784
Sintático− Saliência+	0,922	0,915	0,918	0,910	0,918	0,914

Tabela 6: Resultados usando o J48 sem e com *oversampling* (SMOTE).

apresentou os piores resultados. Como medidas de avaliação foram utilizadas a medida Kappa e a acurácia (percentagem de acerto), uma vez que o aumento gradativo do córpus buscou mantê-lo balanceado. O gráfico mostrando os resultados para os dois cenários é apresentado na Figura 5.

Como esperado, o aumento do tamanho do córpus influenciou positivamente os valores de acurácia e *Kappa* nos dois cenários avaliados, porém de maneiras diferentes. Conforme pode ser observado na Figura 5, os resultados calculados para o cenário (1) – melhor cenário – permaneceram os mesmos após o balanceamento (227), aumentaram significativamente no primeiro aumento de tamanho (454) e se estabilizaram a partir desse ponto. Já os resultados para o cenário (2) – pior cenário – aumentaram de forma acentuada e contínua desde o balanceamento e pelos consecutivos aumentos de tamanho, começando a estabilizar a partir do terceiro aumento no tamanho (908). Ainda sim, os resultados para o cenário (2) permaneceram abaixo dos resultados para o cenário (1) em todas as avaliações. Esse experimento confirma o que já havia sido observado por Barzilay e Lapata (2008), que a partir de um certo número de exemplos – e considerando-se o balanceamento entre as

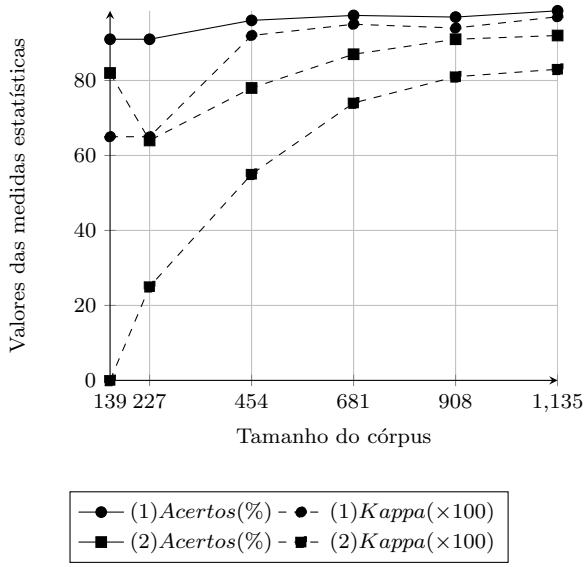

Figura 5: Variação dos valores de acurácia e medida *Kappa* de acordo com o aumento artificial do tamanho do córpus.

classes – o desempenho do modelo grade de entidades se estabiliza. Esses resultados também reforçam a configuração utilizada no cenário (1) como a melhor configuração para esse córpus, independentemente do seu tamanho.

5 Conclusões e Trabalhos Futuros

Este trabalho teve por objetivo avaliar o modelo grade de entidades proposto por Barzilay e Lapata (2008) na avaliação de coerência em textos 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 e 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 automático 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, Tetreault e Andreyev (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, ainda que indireta, dos resultados deste trabalho com os obtidos para outras línguas, mostrando que os resultados são próximos aos relatados para a língua inglesa e superiores ao relatado para a língua japonesa e alemã. 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 ($K = 0,70$) na classificação dos resumos usando duas classes: "sem problemas"/"com problemas".

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. Para isso é preciso encontrar formas de se mapear os resultados do modelo em um *feedback* que seja útil ao usuário do sistema. 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 com modelo grade de entidades por meio da combinação do modelo original com conhecimentos provenientes de outras fontes, como os índices calculados pela Coh-Metrix-Port (Scarton e Aluísio, 2010; Scarton, Almeida e Aluísio, 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.

Agradecimentos

Os autores agradecem a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro na realização deste trabalho.

Referências

Barzilay, Regina e Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34:1–34, March, 2008.

Bick, Eckhard. 2002. *The Parsing System Palavras - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Famework*. Tese de doutoramento, Department of Linguistics – Aarhus: Aarhus University Press – DK.

Burstein, Jill, Martin Chodorow, e Claudia Leacock. 2003. Criterion online essay evaluation: An application for automated evaluation of student essays. Em *Proceedings of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence*, pp. 3–10.

Burstein, Jill, Joel Tetreault, e Slava Andreyev. 2010. Using entity-based features to model coherence in student essays. Em *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pp. 681–684.

Cardoso, Paula Christina Figueira, Erick Galani Maziero, Maria Lucia del Rosario Castro Jorge, Eloize Rossi Marques Seno, Ariani Di Felippo, Lúcia Helena Machado Rino, Maria das Graças Volpe Nunes, e Thiago Alexandre Salgueiro Pardo. 2011. Cstnews - a discourse-annotated corpus for single and multi-document summarization of news texts in brazilian portuguese. Em *Proceedings of the 3rd RST Brazilian Meeting*, pp. 88–105, Cuiabá/MT, Brazil.

Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, e W. Philip Kegelmeyer. 2002. Smote: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Collovini, Sandra, Thiago Ianez Carbonel, Juliana Thiesen Fuchs, Jorge César Barbosa Coelho, Lúcia Helena Machado Rino, e Renata Vieira. 2007. Summ-it: um corpus anotado com informações discursivas visando sumarização automática. Em *Anais do XXVII Congresso da SBC: V Workshop em Tecnologia da Informação e da Linguagem Humana (TIL 2007)*.

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

Elsner, Micha e Eugene Charniak. 2011. Extending the entity grid with entity-specific features. Em *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pp. 125–129.

Feltrim, Valéria Delisandra, Simone Teufel, Maria das Graças Volpe Nunes, e Sandra Maria Aluísio. 2006. Argumentative zoning applied to criquing novices scientific abstracts. Em James G. Shanahan, Yan Qu, e Janyce Wiebe, editores, *Computing Attitude and Affect in Text: Theory and Applications*, pp. 233–246, Dordrecht, The Netherlands. Springer.

Filippova, Katja e Michael Strube. 2007. Extending the entity-grid coherence model to semantically related entities. Em *Proceedings of the Eleventh European Workshop on Natural Language Generation*, ENLG '07, pp. 139–142.

Freitas, Alison Rafael Polpeta. 2013. *Análise Automática de Coerência Usando o Modelo Grade de Entidades para o Português*. Dissertação de Mestrado, Departamento de Informática – Universidade Estadual de Maringá, Maringá/PR - Brasil.

Grosz, Barbara J., Scott Weinstein, e Aravind K. Joshi. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21:203–225.

Higgins, Derrick, Jill Burstein, Daniel Marcu, e Cláudia Gentile. 2004. Evaluating multiple aspects of coherence in student essays. Em *Proceedings of the 2004 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 185–192.

Joachims, Thorsten. 2006. Training linear svms in linear time. Em *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pp. 217–226, New York, NY, USA. ACM.

Landauer, Thomas K., Peter W. Foltz, e Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse processes*, 25:259–284.

Landauer, Thomas K., Darrell Laham, e Peter W. Foltz. 2003. *Automated essay scoring and annotation of essays with the intelligent essay assessor.* Lawrence Erlbaum Associates, Mahwah, NJ.

Lin, Ziheng, Hwee Tou Ng, e Min-Yen Kan. 2011. Automatically evaluating text coherence using discourse relations. Em *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pp. 997–1006.

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

Platt, John. 1998. Fast training of support vector machines using sequential minimal optimization. Em *B. Schoelkopf and C. Burges and A. Smola, editors, Advances in Kernel Methods - Support Vector Learning*.

Quinlan, Ross. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA.

Rino, Lúcia Helena Machado e Thiago Alexandre Salgueiro Pardo. 2007. *A coleção TeMário e a avaliação de sumarização automática*, volume 1. IST Press, Lisboa, Portugal.

Scarton, Carolina Evaristo, Daniel Machado de Almeida, e Sandra Maria Aluísio. 2009. Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: adaptando as métricas do coh-metrix para o português. Em *Proceedings of the 7th Brazilian Symposium in Information and Human Language Technology (STIL 2009)*. 1 CD-ROM v1.

Scarton, Carolina Evaristo e Sandra Maria Aluísio. 2010. Coh-metrix-port: a readability assessment tool for texts in brazilian portuguese. Em *Proceedings of the 9th International Conference on Computational Processing of the Portuguese Language, Extended Activities Proceedings*, PROPOR '10. 1 CD-ROM v1.

Souza, Vinícius Mourão Alves e Valéria Delisandra Feltrim. 2011. An analysis of textual coherence in academic abstracts written in portuguese. Em *Proceedings of the Sixth Corpus Linguistics Conference (CL 2011)*, pp. 1–13, Birmingham, UK.

Souza, Vinícius Mourão Alves e Valéria Delisandra Feltrim. 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, Ian H. e Eibe Frank. 2005. *Data Mining - Practical Machine Learning Tools and Techniques*. Morgan Kaufmann – Elsevier.

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