UNIVERSIDADE ESTADUAL DE MARINGÁ
CENTRO DE TECNOLOGIA
DEPARTAMENTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

ALISON RAFAEL POLPETA FREITAS

Análise automática de coerência usando o modelo grade de entidades para o português

Maringá
2013

ALISON RAFAEL POLPETA FREITAS

Análise automática de coerência usando o modelo grade de entidades para o português

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Departamento de Informática, Centro de Tecnologia da Universidade Estadual de Maringá, como requisito parcial para obtenção do título de Mestre em Ciência da Computação.

Orientadora: Profa. Dra. Valéria Delisandra Feltrim

Maringá
2013

Dados Internacionais de Catalogação-na-Publicação (CIP)
(Biblioteca Central - UEM, Maringá – PR., Brasil)

F866a

Freitas, Alison Rafael Polpeta
Análise automática de coerência usando o modelo grade de entidades para o português / Alison Rafael Polpeta Freitas. -- Maringá, 2013.
85 f. : il., color., figs., tabs.

Orientador: Profª. Drª. Valéria Delisandra Feltrim.
Dissertação (mestrado) - Universidade Estadual de Maringá, Centro de Tecnologia, Departamento de Informática, Programa de Pós-Graduação em Ciência da Computação, 2013.

1. Coerência semântica. 2. Análise automática de coerência. 3. Modelo grade de entidades. I. Feltrim, Valéria Delisandra, orient. II. Universidade Estadual de Maringá. Centro de Tecnologia. Departamento de Informática. Programa de Pós-Graduação em Ciência da Computação. III. Título.

CDD 21.ed. 006.35

AHS-001524

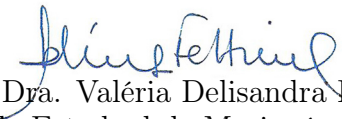
FOLHA DE APROVAÇÃO

ALISON RAFAEL POLPETA FREITAS

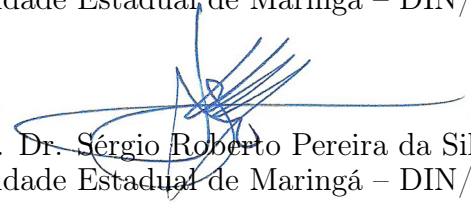
Análise automática de coerência usando o modelo grade de entidades para o português

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Departamento de Informática, Centro de Tecnologia da Universidade Estadual de Maringá, como requisito parcial para obtenção do título de Mestre em Ciência da Computação pela Banca Examinadora composta pelos membros:

BANCA EXAMINADORA



Profa. Dra. Valéria Delisandra Feltrim
Universidade Estadual de Maringá – DIN/UEM



Prof. Dr. Sérgio Roberto Pereira da Silva
Universidade Estadual de Maringá – DIN/UEM

Profa. Dra. Sandra Maria Aluísio
Universidade de São Paulo – ICMC/USP

Aprovada em: 01 de março de 2013.

Local da defesa: Sala 101, Bloco C56, *campus* da Universidade Estadual de Maringá.

DEDICATÓRIA(S)

À minha família, pelo apoio, incentivo e motivação sem os quais nada seria possível.

AGRADECIMENTO(S)

Agradeço à professora Valéria por todo auxílio, confiança e amizade que obtive sendo seu orientando e pelo conhecimento compartilhado, não apenas na área de pesquisa como também na docência, sendo para mim uma referência no que diz repeito ao significado de “aprender” e “ensinar”, sempre muito paciente com minhas limitações. Aos amigos e professores que me envolvi, em especial Murilo, Danilo e Vinícius, pelo incentivo e por ajudarem no desenvolvimento deste trabalho. À CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) pelo apoio financeiro e à Universidade Estadual de Maringá e ao Departamento de Informática que permitiram a realização desta pesquisa.

Análise automática de coerência usando o modelo grade de entidades para o português

RESUMO

Um texto bem escrito deve ser coerente de forma que cada novo trecho de informação possa ser interpretado de acordo com o contexto precedente, um fator fundamental para a inteligibilidade e interpretabilidade do texto. A coerência é normalmente refletida pela estrutura do discurso e suas relações, as quais podem ser analisadas por meios de teorias como a RST (*Rethorical Structure Theory*), *Centering* e o modelo LSA (*Latent Semantic Analysis*), sendo essas as principais fontes de estudos para o desenvolvimento de ferramentas que possam distinguir textos coerentes e incoerentes, seja na geração, sumarização ou avaliação automática de textos. Para a língua portuguesa, o SciPo (*Scientific Portuguese*) é um exemplo de ferramenta de auxílio à escrita que possui um módulo de análise de coerência para a detecção de potenciais problemas de coerência semântica em resumos. Baseado em LSA, esse módulo analisa os relacionamentos semânticos entre as sentenças de um resumo, de acordo com um conjunto pré-estabelecido de dimensões. Para uma das dimensões propostas para o módulo de análise de coerência, chamada Quebra de Linearidade, os resultados obtidos por meio de LSA foram pouco satisfatórios, sugerindo a utilização de outros modelos de coerência. Nesse contexto, este trabalho teve por objetivo investigar a aplicabilidade do modelo grade de entidades (do inglês *Entity-Grid*), um modelo estatístico baseado na teoria de *Centering*, na avaliação de coerência de resumos científicos escritos em português. Mais especificamente, se buscou avaliar se tal modelo poderia ser empregado na implementação de um classificador capaz de detectar problemas locais de coerência, semelhantes aos descritos na dimensão Quebra de Linearidade, visando a futura inclusão de tal classificador no módulo de análise de coerência. Os resultados obtidos nos experimentos se mostraram próximos aos resultados do modelo grade de entidades original para a língua inglesa e semelhantes aos resultados relatados por trabalhos relacionados para outras línguas. Os resultados também foram próximos ao obtido em experimento com juízes humanos, mostrando que o modelo grade de entidades tem potencial para ser usado no contexto do sistema SciPo.

Palavras-chave: coerência semântica, análise automática de coerência, modelo grade de entidades.

Automatic coherence analysis using the entity-grid model in portuguese

ABSTRACT

A well-written text should be coherent, it means that each new piece of information can be interpreted as the preceding context, a key factor for the readability and interpretability of the text. Coherence is usually reflected by the structure of discourse and its relations, which can be analyzed by theories such as RST (Rhetorical Structure Theory), Centering, and models like LSA (Latent Semantic Analysis), which are the main sources of studies for developing tools that can distinguish coherent texts of incoherent in the generation, automatic summarization and evaluation of texts. For the Portuguese language, SciPo (Scientific Portuguese) is an example of a writing tool which has a coherence analysis module that detects potential problems of semantic coherence in abstracts. Based on LSA, this module analyzes different semantic relationships among sentences, according to a pre-established set of dimensions. For one of these dimensions, named Linearity-break, evaluation results were not satisfactory, suggesting the use of other coherence models. In this context, our purpose was to investigate the applicability of the Entity-Grid model, a statistical model based on the Centering theory, in the evaluation of coherence in scientific abstracts written in Portuguese. More specifically, we aimed at assessing whether such model could be employed in the implementation of a classifier capable of detecting local coherence problems, similar to those assessed by the dimension Linearity-break, aiming at future inclusion of such classifier as part of coherence analysis module. Our experimental results are close to those of the original entity-grid model for English language and very similar to the results reported by related works for other languages. Results are also close to those obtained by human judges in an annotation experiment, showing that the entity-grid model can be applied in the context of SciPo system.

Keywords: semantic coherence, automatic analysis of coherence, entity-grid model.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AES	Automated Essay Scoring
AM	Aprendizagem de Máquina
AMADEUS	Amiable Article Development for User Support
API	Application Programming Interface
ARFF	Attribute-Relation File Format
AZ	Argumentative Zoning
AZPort	Argumentative Zoning for Portuguese
DC-UEL	Departamento de Computação da Universidade Estadual de Londrina
DIN-UEM	Departamento de Informática da Universidade Estadual de Maringá
HTML	Hypertext Markup Language
IDF	Inverse Document Frequency
Inf-UFPel	Departamento de Informática do Instituto de Física e Matemática da Universidade Federal de Pelotas
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
LSI	Latent Semantic Indexing
MAC	Módulo de Análise de Coerência
MR	Meta-regra
PDTB	The Penn Discourse Treebank
PLN	Processamento de Linguagem Natural
RST	Rhetorical Structure Theory
SciPo	Scientific Portuguese
SGML	Standard Generalized Markup Language
SMO	Sequential Minimal Optimization
SMOTE	Synthetic Minority Over-sampling Technique
SN	Sintagma Nominal
SVD	Singular Value Decomposition
SVM	Support Vector Machine
TOEFL	Test Of English as a Foreign Language
TF	Term Frequency
XML	Extensible Markup Language
W3C	World Wide Consortium
WEKA	Waikato Environment for Knowledge Analysis
WSJ	Wall Street Journal

SUMÁRIO

Capítulo

1

Introdução

Para uma grande variedade de aplicações na área de Processamento de Linguagem Natural (PLN), 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. Enquanto algumas dessas ferramentas focam em avaliar o texto já pronto e em lhe atribuir um nota (chamadas de *Scoring systems*), outras oferecem um suporte mais abrangente, apoiando o escritor em várias fases do processo de composição do texto, desde a geração e organização de ideias até a realização linguística. Independente do tipo de suporte oferecido, em geral, essas ferramentas realizam algum tipo de avaliação de coerência.

Exemplos de sistemas que consideram aspectos de coerência na avaliação de um texto são as ferramentas *Criterion* (Burstein et al., 2003; Higgins et al., 2004), *Intelligent Essay Assessor* (Landauer et al., 2003) e *Intellimetric* (Elliot, 2003). Essas ferramentas buscam avaliar a qualidade de redações (*essays*) escritas em inglês e são classificadas como *scoring systems*, já que retornam uma nota (*score*) considerando aspectos indicativos da qualidade do texto, tais como foco e coerência, organização, desenvolvimento, gramática, ortografia e estilo. Vale ressaltar que esses aspectos podem variar de uma ferramenta para outra. Além da nota dada ao texto, em geral, essas ferramentas também disponibilizam críticas e sugestões que podem ser utilizadas pelo escritor para melhorar o texto avaliado.

Outro exemplo de ferramenta que avalia aspectos de coerência em textos escritos em inglês é a Coh-Metrix (Graesser et al., 2004; McNamara et al., 2010, 2006). A Coh-Metrix é uma ferramenta *online* que analisa um texto e retorna um conjunto de

62 índices referentes à coesão, coerência e dificuldade de compreensão. É importante destacar que a ferramenta completa possui mais de 500 índices, porém apenas 62 são disponibilizadas publicamente. Uma versão desses índices para a língua portuguesa, chamada Coh-Metrix-Port foi proposta por Scarton e Aluísio (2010); Scarton et al. (2009). Nessa versão, 41 índices estão disponíveis publicamente, sendo que um subconjunto deles se refere à aspectos de coerência.

Outro exemplo para a língua portuguesa, é a ferramenta de auxílio à escrita é o SciPo (*Scientific Portuguese*)(Feltrim et al., 2006; Souza e Feltrim, 2011), um sistema desenvolvido para ajudar especificamente escritores iniciantes na escrita científica, em especial estudantes na área da Computação, fornecendo suporte para a escrita de resumos e introduções de teses e dissertações em português. Diferentemente das ferramentas para a língua inglesa, o SciPo não tem por objetivo dar uma nota ao texto. Em vez disso, busca apoiar o escritor na composição do texto desde o primeiro rascunho, auxiliando na estruturação esquemática do texto[1] e fornecendo um *feedback* indicativo dos pontos que podem ser melhorados. Além da análise e crítica da estrutura esquemática, o SciPo também possui um módulo de análise de coerência (MAC) que detecta potenciais problemas de coerência semântica em resumos. Basicamente, esse módulo analisa o nível de similaridade semântica existente entre diferentes componentes esquemáticos identificados no texto. O nível de similaridade semântica entre dois ou mais componentes esquemáticos é medido por meio de LSA (*Latent Semantic Analysis*)(Landauer et al., 1998). Especificamente, três tipos de relacionamentos semânticos, chamados por Souza e Feltrim (2012) de 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 (2012) propõem ainda uma quarta dimensão, chamada Quebra de Linearidade, em que se verifica a existência de uma quebra de linearidade entre sentenças adjacentes do resumo. Essa quebra se caracteriza pela dificuldade em se estabelecer um sentido lógico da sentença atual com a sentença anterior e/ou próxima e pode ser interpretada como uma barreira na leitura (por exemplo, falta de continuidade anafórica), demandando maior esforço cognitivo para a interpretação do texto. Assim como foi feito na implementação das dimensões 1, 2 e 3, Souza e Feltrim (2012) utilizaram atributos

[1]Neste trabalho, entende-se por estrutura esquemática ou retórica do texto, o conjunto de elementos que descrevem a organização (funcional) do texto, relacionando suas partes de forma a atribuir-lhe sentido.

extraídos com o uso da LSA para induzir um classificador capaz de detectar quebras de linearidades entre sentenças adjacentes de um resumo. No entanto, os resultados obtidos com tal classificador 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 por Souza e Feltrim (2012). Conforme sugerido pelos próprios autores, um modelo de coerência que fosse capaz de mapear o fluxo textual de forma mais refinada, como o modelo grade de entidades proposto por Barzilay e Lapata (2008), poderia obter melhores resultados para essa dimensão.

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 semelhantes às definidas pela Teoria de *Centering* (Grosz et al., 1995). A teoria de *centering* preconiza que em um texto coerente o foco de atenção (uma entidade) tende a ser o mesmo 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 seria capaz de apreender padrões de transições característicos de textos coerentes/incoerentes.

Nesse contexto, este trabalho teve por objetivo investigar a aplicabilidade do modelo grade de entidades na avaliação de coerência de resumos científicos escritos em português. Mais especificamente, se buscou avaliar se tal modelo poderia ser empregado na implementação de um classificador capaz de detectar problemas locais de coerência, semelhantes aos descritos na dimensão Quebra de Linearidade de Souza e Feltrim (2012), visando a futura inclusão de tal classificador no MAC do sistema SciPo.

Para a implementação do modelo, um sistema de pré-processamento capaz de extrair entidades a partir de textos em português foi construído utilizando o *parser* PALAVRAS (Bick, 2002) como ferramenta principal para a identificação dos sintagmas nominais (SNs). Processamento adicional foi necessário 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. Vale destacar que, no modelo original, Barzilay e Lapata (2008) utilizaram um sistema de resolução automática de correferência, transformando cada grupo de entidades correferentes em uma única entidade. Embora existam na literatura vários trabalhos referentes à resolução de correferência em português (Chaves e Rino, 2008; Cuevas e Paraboni, 2008; Rossi et al., 2001; Silva e Rosa, 2010; Souza et al., 2008), até a finalização deste trabalho não foi encontrado um sistema pronto e disponível para uso. Desse modo, neste traballho, as entidades extraídas do texto foram lematizadas e agrupadas por lemas e todos os lemas extraídos são incluídos na grade de entidades. Além dos atributos extraídos da grade de entidades previstos no modelo

original, neste trabalho também foram incluídos atributos relacionados à variabilidade de vocabulário, mais especificamente, foi utilizado o modelo *Type/Token* para extrair atributos adicionais de forma similar ao relatado por Burstein et al. (2010).

Para o treinamento e teste do modelo foi utilizado como córpus um conjunto de 99 resumos científicos extraídos do córpus de resumos de Trabalhos de Conclusão de Curso (TCC) em Ciência da Computação compilado por Souza e Feltrim (2012), além de 40 resumos coletados diretamente com os alunos formandos do curso de graduação em Ciência da Computação da Universidade Estadual de Maringá, totalizando um conjunto de 139 resumos científicos. Além do córpus de resumos científicos, um corpus jornalístico composto de 286 textos extraídos dos corpora CSTNews (Cardoso et al., 2011), Summ-it (Collovini et al., 2007) e Temário (Rino e Pardo, 2007) foi utilizado para experimentos preliminares semelhantes aos realizados por Barzilay e Lapata (2008).

Para induzir classificadores capazes de discriminar entre textos "sem problemas de coerência"e textos "com problemas de coerência"foram utilizados algoritmos de aprendizagem de máquina. Os experimentos com o córpus jornalístico foram realizados utilizando o sistema SVM^{rank} (Joachims, 2006), que implementa o algoritmo SVM (*Support Vector Machine*) para problemas de ranqueamento (*ranking learning problem*). Os experimentos com o córpus de resumos científicos foram realizados usando as implementações fornecidas pelo ambiente WEKA (Witten e Frank, 2005) de três algoritmos de aprendizegem conhecidos: SVM, C4.5 e Naïve Bayes. Nos experimentos com o córpus jornalístico, os textos originais foram rotulados como “sem problemas” de coerência, enquanto versões criadas artificialmente por meio da permutação aleatória da ordem das sentenças dos textos originais foram rotuladas como “com problemas” de coerência. Nos experimentos com o córpus de resumos científicos, o julgamento humano foi utilizado para rotular cada texto como “sem problemas”/“com problemas”.

Os resultados obtidos nos experimentos com o córpus jornalístico se mostraram próximos aos resultados do modelo grade de entidades original para a língua inglesa (Barzilay e Lapata, 2008) e semelhantes aos resultados relatados por trabalhos relacionados para outras línguas, como o alemão (Filippova e Strube, 2007) e o japonês (Yokono e Okumura, 2010). Já os resultados obtidos com o córpus de resumos científicos não puderam ser comparados diretamente com os resultados do modelo original, uma vez que nos experimentos com esse córpus a fase de aprendizagem foi modelada como um problema de classificação e não como um problema de ranqueamento. Tal diferença de modelagem se deve ao fato deste trabalho ter por motivação a construção de um classificador para a dimensão Quebra de Linearidade a ser incorporado no módulo de análise de coerência do sistema SciPo. Dessa forma, esses resultados foram analisados apenas no contexto de sua aplicação e mostraram que o uso do modelo de grade de entidades no contexto do sistema SciPo é viável, alcançando resultados próximos ao obtido em experimento com juízes humanos.

Esta dissertação está organizada da seguinte maneira: no Capítulo 2 é apresentada a revisão bibliográfica que serviu de suporte ao desenvolvimento deste trabalho, abordando aspectos relacionados à coerência textual, assim como teorias linguísticas e modelos computacionais relacionados à coerência, incluindo o modelo grade de entidades e outros trabalhos relacionados. No Capítulo 4 são descritos os corpora utilizados para a avaliação do modelo grade de entidades implementado para a língua portuguesa, bem como os detalhes relativos à anotação do corpus de resumos científicos usado neste trabalho. No Capítulo 5, a implementação do modelo grade de entidades para a língua portuguesa é detalhada e os resultados dos experimentos realizados são apresentados e discutidos. Por fim, no Capítulo 6 são apresentadas as conclusões deste trabalho.

Capítulo

2

Coerência Textual: Definições, Teorias e Modelos relacionados

A Linguística Textual teve seu estudo iniciado na década de 60, momento em que os linguistas passam a argumentar que um texto não é simplesmente uma sequência de frases isoladas, mas uma unidade linguística com propriedades estruturais específicas. Mas foi apenas próximo à década de 80 que vários estudos sobre o texto apareceram, com teorias sobre a composição textual, surgindo a distinção entre fenômenos ligados à coesão e/ou coerência (Koch, 1994). Beaugrande e Dressler (1981) definem sete fatores responsáveis pela textualidade: (i) coesão, (ii) coerência, (iii) informatividade, (iv) situcionalidade, (v) intertextualidade, (vi) intencionalidade e (vii) aceitabilidade, sendo o fator (ii) coerência o foco deste trabalho.

Este capítulo apresenta a revisão bibliográfica que serviu de suporte ao desenvolvimento deste trabalho, abordando aspectos relacionados à coerência textual, assim como teorias linguísticas e modelos computacionais relacionados à coerência. A Seção 2.1 traz uma breve discussão acerca da definição de coerência textual. Na Seção 2.2 são apresentadas teorias e métodos empregados na análise da coerência textual, tanto de base linguística quanto estatística, incluindo, na Subseção 2.2.4, o modelo Grade de Entidades de Barzilay e Lapata (2008), que serviu de base para o desenvolvimento deste trabalho. Outros trabalhos relacionados ao modelo Grade de Entidades também são discutidos na Subseção 2.2.5.

2.1 Coerência Textual

Não existe um consenso na literatura quanto a uma definição de coerência textual, sendo que diversas definições foram formuladas ao longo dos anos por teóricos como van Dijk e Kintsch (1983), Charolles (1978) e Koch e Travaglia (2003). Embora os autores apresentem pontos de vista distintos, todos concordam que a coerência é um dos fatores principais responsáveis para se alcançar a textualidade, fazendo com que o texto não seja um amontoado de palavras ou frases sem sentido entre si (Koch e Travaglia, 2003).

Para Koch e Travaglia (2003), a coerência está diretamente ligada à possibilidade de estabelecer um sentido para o texto, sendo um princípio de interpretabilidade, ligada à inteligibilidade do texto em uma situação de comunicação e à capacidade que o receptor tem para calcular o sentido desse texto. É construída a partir da interação entre a situação de comunicação, os conceitos ativados pelo texto e os conhecimentos armazenados na memória dos participantes. Dessa maneira, a coerência está vinculada ao texto, mas não depende somente dele.

Diferentemente de Koch e Travaglia (2003), para os quais a coerência é uma propriedade inteiramente global ligada à composição do texto como um todo, Charolles (1978) estabelece dois níveis de coerência que se complementam, denominados coerência (i) microestrutural e (ii) macroestrutural. A **coerência microestrutural** corresponde a um nível local de organização do texto formada por relações estabelecidas entre palavras ou frases sucessivamente ordenadas, enquanto a **coerência macroestrutural** corresponde a um nível global de organização formada por sequências maiores do texto.

Charolles (1978) também propôs **quatro meta-regras** para a formação de um texto coerente. São elas:

1. **Meta-regra de repetição (MR-I):** um texto coerente deve ter elementos repetidos.

   Esta meta-regra está relacionada à recuperação de termos de frases anteriores por meio de pronomes, elipses, elementos lexicais ou substitutivos, o que constitui um processo de repetição ou recorrência. Essa meta-regra pode ser interpretada como um mecanismo de coesão textual definido por autores como Beaugrande e Dressler (1981), de modo que ela constitui uma condição necessária, mas não suficiente para que um texto seja coerente.

2. **Meta-regra de progressão (MR-II):** um texto coerente deve apresentar renovação do suporte semântico.

   A segunda meta-regra complementa a primeira, posto que estipula que um texto coerente não pode simplesmente repetir indefinidamente seu próprio assunto, evitando a circularidade do discurso e apresentando informações novas à medida que vai progredindo.

3. **Meta-regra de não-contradição (MR-III):** cada parte de um texto coerente deve fazer sentido com o que foi escrito anteriormente.

   Com base na lógica, Charolles (1978) recorda que é inadmissível que uma mesma proposição seja conjuntamente verdadeira e não verdadeira, ou falsa e não falsa.

4. **Meta-regra de relação (MR-IV):** para que uma sequência ou um texto seja coerente, é preciso que os fatos que se denotam no mundo representado estejam diretamente relacionados.

   Charolles (1978) ressalta que essa regra traz em si intensas características pragmáticas. As relações entre os enunciados podem ocorrer tanto com a presença explícita de conectivos, como sem a presença explícita dos mesmos. Nesse último caso, a conexão entre os fatos ocorre somente mediante a fatores pragmáticos, como conhecimento de mundo e conhecimento compartilhado.

É importante destacar que as quatro meta-regras de Charolles (1978) não satisfazem sozinhas todas as condições necessárias para um texto ser avaliado como coerente, fato reconhecido pelo próprio autor. Também não foram pensadas como tendo caráter normativo, de modo a estabelecer regras que constituem a coerência textual, mas sim como diretrizes para a composição e avaliação de textos coerentes por meio de orientações referentes à continuidade temática, progressão semântica, possíveis contradições entre ideias e pertinência do conteúdo abordado.

Assim como Charolles (1978), van Dijk (1983) e van Dijk e Kintsch (1983) também estabelecem dois níveis básicos de coerência, denominando-os de coerência (i) local e (ii) global. De modo similar a coerência microestrutural/macroestrutural de Charolles (1978), a **coerência local** de van Dijk e Kintsch (1983) diz respeito às partes do texto como sentenças ou proposições e suas conexões lineares, enquanto a **coerência global** se refere à totalidade do texto ou a fragmentos maiores, em que se estabelece uma conexão com o tema, ideia, essência ou finalidade.

Embora seja feita essa distinção de níveis, tanto Charolles (1978) quanto van Dijk e Kintsch (1983) afirmam que existe uma interdependência entre os mesmos, pois um conjunto de sentenças linearmente coerentes (coerência local) contribui para se alcançar a coerência global e, da mesma forma, a coerência local pode ser construída a partir da coerência global.

van Dijk e Kintsch (1983) afimam ainda que a coerência textual pode ser analisada sob **quatro perspectivas**. São elas:

1. **Coerência semântica:** refere-se à relação entre os significados dos elementos que constituem as sentenças, que devem ser relacionados ou complementares. A coerência semântica também se refere à relação entre sentenças sucessivas do discurso, que devem ter seus conteúdos interligados. A incoerência semântica ocorre

quando o significado dos elementos das sentenças ou os conteúdos de sentenças sucessivas são semânticamente distantes. A coerência semântica pode se manifestar tanto em nível local quanto global.

2. **Coerência sintática:** refere-se aos meios sintáticos usados para expressar a coerência semântica, como o uso de conectivos, pronomes, artigos, advérbios, entre outros meios. Equivale aos recursos coesivos da língua que auxiliam na construção da coerência.

3. **Coerência estilística:** refere-se ao estilo de escrita conforme o gênero textual, sem a mistura de registros linguísticos, como o uso de linguagem informal em um texto formal. A coerência estilística envolve o léxico e as estruturas frasais adequadas ao texto produzido.

4. **Coerência pragmática:** refere-se ao texto visto como uma sequencia de atos de fala realizados de forma apropriada e com informações claras. Por exemplo, a resposta de uma pergunta pode se manifestar por meio de uma afirmação, de outra pergunta, de uma promessa, de uma negação e não de algo totalmente desconectado ao tema da pergunta. Nesse sentido, os fatores pragmáticos equivalem ao fato do produtor e o interlocutor reconhecerem o esquema textual empregado na situação comunicativa, considerando-se o contexto extralinguístico e o significado pretendido.

É importante ressaltar que uma incoerência pode ser caracterizada sob mais de uma perspectiva. Por exemplo, problemas estilísticos podem equivaler também a falhas lexicais, caracterizando tanto incoerência estilística quanto sintática.

2.2 Teorias e Métodos Relacionados à Coerência Textual

Conforme discutido na seção anterior, a coerência é um aspecto textual multifacetado e carregado de subjetividade. Dessa forma, estabelecer um modelo capaz de avaliar ou "mensurar" a coerência textual não é uma tarefa trivial, especialmente quando se pretende que tal modelo seja passível de uso por sistemas computacionais.

Nesta seção são apresentadas duas teorias – a Teoria da Estrutura Retórica (RST) (Mann e Thompson, 1988) e a teoria de *Centering* (Grosz et al., 1995) – e dois modelos computacionais – a Análise de Semântica Latente (LSA)(Landauer et al., 1998) e o modelo Grade de Entidades (Barzilay e Lapata, 2008) – relacionados à coerência textual. Vale ressaltar que o modelo Grade de Entidades é o foco principal deste trabalho, de modo que o mesmo é apresentado mais detalhadamente na Subseção 2.2.4.

2.2.1 Teoria da Estrutura Retórica

A Teoria da Estrutura Retórica (*Rhetorical Structure Theory*: RST) (Mann e Thompson, 1988) é uma teoria discursiva cujo objetivo é descrever a organização do texto por meio de relações entre suas partes funcionais. Essas partes são porções de texto chamadas de *text spans*, constituídas de intervalos lineares e ininterruptos de texto, e o conjunto de relações definidas entre essas partes forma a estrutura retórica do texto.

Segundo Mann e Thompson (1988), a RST fornece meios de se obter o conteúdo proposicional implícito de um texto por meio de relações proposicionais entre suas partes. Ainda segundo os autores, essas proposições são essenciais para a coerência do texto, de tal forma que, se um texto for coerente, será sempre possível extrair a sua estrutura retórica. Por esse motivo, as relações retóricas também são chamadas *relações de coerência*.

Na RST original foram definidas 26 relações retóricas para relacionar as proposições expressas em um texto. Essas relações se estabelecem entre duas ou mais proposições de segmentos adjacentes do texto, sendo uma nuclear (N), que indica a informação principal, e outra informação adicional complementar, chamada de satélite (S). Quando ambas as informações são igualmente importantes, tem-se uma relação multinuclear, em que se tem mais de um núcleo e nenhum satélite. O conjunto dessas relações e suas nuclearidades são apresentadas na Tabela 2.1.

Para o estabelecimento de uma relação entre duas proposições, Mann e Thompson (1988) definem que quatro campos devem ser observados e que o analista deve ter seu julgamento baseado no contexto e nas intenções do escritor:

1. Restrições sobre o núcleo (N);
2. Restrições sobre o satélite (S);
3. Restrições sobre a combinação do núcleo e do satélite (N+S); e
4. Efeito que a relação em questão causa no leitor.

Vamos considerar, por exemplo, a relação *Purpose*:

1. Restrições sobre (N): apresenta uma ação;
2. Restrições sobre (S): apresenta uma situação não realizada;
3. Restrições sobre (N+S): S apresenta uma situação que pode realizar N; e
4. Efeito: o leitor reconhece que a atividade em N pode ser iniciada por meio de S.

Dessa maneira, existe um conjunto de restrições para cada uma das 26 relações estabelecidas na RST que pode ser consultado na obra de Mann e Thompson (1988). O mesmo conjunto traduzido para o português pode ser consultado em Pardo (2005).

Tabela 2.1: Relações e nuclearidade da RST (Mann e Thompson, 1988).

Relação	**Multinuclear**
Antithesis	Não
Background	Não
Circumstance	Não
Concession	Não
Condition	Não
Elaboration	Não
Enablement	Não
Evaluation	Não
Evidence	Não
Interpretation	Não
Justify	Não
Means	Não
Motivation	Não
Non-volitional Cause	Não
Non-volitional Result	Não
Otherwise	Não
Purpose	Não
Restatement	Não
Solutionhood	Não
Summary	Não
Volitional Cause	Não
Volitional Result	Não
Contrast	Sim
Joint	Sim
List	Sim
Sequence	Sim

Apresentamos os exemplos descritos em Pardo (2005) para exemplificar o uso de algumas das relações. No trecho de texto *"Embora você não goste, trabalhar é importante."* a proposição expressa pela primeira oração é o satélite e a proposição expressa pela segunda oração é o núcleo da relação retórica de oposição *Concession*. A estrutura retórica construída pela RST também pode ser representada por árvores, como pode ser observado na Figura 2.1, que ilustra o exemplo apresentado anteriormente.

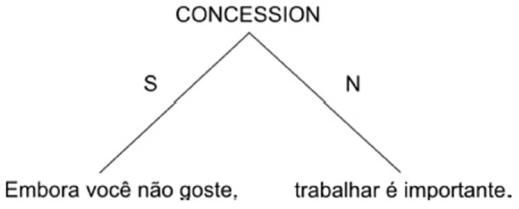


Figura 2.1: Exemplo de estrutura retórica RST (Pardo, 2005).

Ainda no exemplo apresentado na Figura 2.1, se adicionarmos a oração "O trabalho enobrece o homem." às sentenças já existentes, temos uma nova relação, *Justify*, em que

a nova oração é a satélite e as proposições que formam a relação *Concession* são o núcleo da relação *Justify*, formando uma hierarquia de relações. Esta hierarquia formada pelas relações *Concession* e *Justify* é apresentada na Figura 2.2.

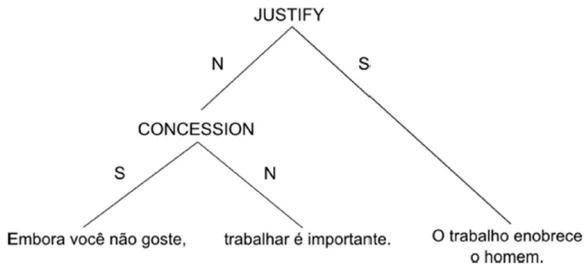


Figura 2.2: Exemplo de estrutura retórica com as relações Concession e Justify (Pardo, 2005).

Já no trecho de texto "O garoto chegou da escola e fez sua lição de casa. Depois, foi brincar com os amigos.", há uma relação *Sequence* (indicando uma "sequência" de eventos) entre as proposições expressas pelas orações "O garoto chegou da escola", "e fez sua lição de casa." e "Depois, foi brincar com os amigos.", sendo que todas são consideradas núcleos da relação, caracterizando uma relação multinuclear, diferentemente dos exemplos anteriores em que as relações são estabelecidas entre uma oração nuclear e outra satélite. A representação desse exemplo é apresentada na Figura 2.3.

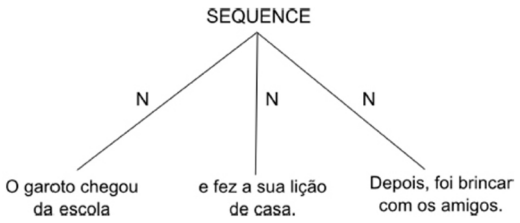


Figura 2.3: Exemplo de estrutura retórica com a relação Sequence (Pardo, 2005).

No trabalho de Taboada e Mann (2006), os autores abordam diversas áreas de aplicação da RST, sendo utilizada na linguística computacional em um grande número de tarefas, tais como análise estrutural, sumarização, tradução e avaliação de textos e geração de linguagem natural.

2.2.2 Análise de Semântica Latente

A Análise de Semântica Latente (*Latent Semantic Analysis*: LSA) (Landauer et al., 1998) é um método estatístico para a extração e representação de conhecimento aplicado em córpus, o qual foi originalmente desenvolvido na *Bell Laboratories e University of Colorado*[1] no contexto de recuperação de informações, tendo sido utilizado em um grande número de aplicações.

[1] `http://lsa.colorado.edu/`

O método simula o processo de aquisição de conhecimento de forma associativa, assim como uma pessoa associa ações/percepções/palavras ao seus respectivos contextos, formando uma grande base de relações semânticas (Kintsch, 2002).

A ideia básica da LSA é construir um espaço semântico em que a semelhança entre os termos se dá pela ocorrência em contextos comuns. Por exemplo, dadas ocorrências como "O médico curou o paciente." e "O cirurgião operou o paciente.", as palavras "médico" e "cirurgião" podem ser consideradas similares, já que ocorrem no mesmo contexto com a palavra "paciente" (Mani, 2001). Para isso, a LSA extrai, a partir da análise de um córpus, uma matriz formada por termos e suas respectivas quantidades de ocorrências nos documentos (contextos) do córpus, como pode ser visto no exemplo da Tabela 2.2.

Tabela 2.2: Exemplo de matriz de co-ocorrência de termos.

	Documento 1	Documento 2	Documento 3	Documento N
Termo 1	0	2	1	...
Termo 2	4	3	5	...
Termo 3	3	4	2	...
Termo K	...	...	...	...

Após a construção da matriz, o valor de cada célula é submetido a uma função de transformação que atribui um peso a cada entrada de acordo com sua importância em relação às outras entradas. Essa função é também conhecida como normalização e é obtida por meio do modelo TF–IDF (*term frequency – inverse document frequency*). O modelo é constituído basicamente de duas etapas. Na primeira etapa (TF) atribui-se um peso às entradas de acordo com a frequência de termos em um único documento. Esse cálculo é realizado por meio de um processo simples que envolve a divisão da frequência de cada termo pelo total de termos do documento. Por exemplo, considere um documento contendo 100 palavras em que determinado termo ocorre 3 vezes. O peso TF desse termo será (3/100) = 0,03. A segunda etapa é a indexação pela frequência inversa (IDF) dos documentos por meio do emprego da fórmula:

$$IDF = \log(\frac{N}{n_k})$$

em que N é o número de documentos do córpus e n_k é o número de documentos em que o termo k ocorre. Por exemplo, considere um córpus com 10 milhões de documentos e que determinado termo ocorre em 1.000 desses documentos, logo o peso IDF é calculado como $\log(\frac{10.000.000}{1.000}) = 4$. Com a combinação das duas etapas obtém-se o valor TF–IDF utilizado para calcular o peso de cada termo nos documentos:

$$\textit{TF-IDF} = \textit{TF} \times \textit{IDF}$$

Após normalizar as entradas da matriz, uma técnica matemática derivada da álgebra linear e denominada *Singular Value Decomposition* – SVD (Golub e Reinsch, 1970) é

utilizada para reduzir a dimensionalidade da matriz e encontrar os principais padrões associativos nos dados, ignorando as influências menos importantes.

Descrevemos a seguir todo processo realizado pela LSA utilizando a SVD de acordo com Deerwester et al. (1990).

Considere uma matriz de termos como a apresentada na Tabela 2.2 e com seus valores normalizados pelo modelo TF–IDF, em que:

- t = número de termos, ou linhas
- d = número de documentos, ou colunas
- X = matriz $t \times d$

A matriz X é decomposta em um produto de outras três matrizes, $X = TSD$, sendo que:

- m = número de dimensões, $m \leqslant min(t, d)$
- T = matriz de vetores singulares à esquerda, $t \times m$
- S = matriz diagonal $m \times m$ de valores singulares em ordem decrescente,
- D = matriz de vetores singulares à direita, $m \times d$

A Figura 2.4 apresenta esquematicamente o processo realizado pela SVD de decomposição da matriz termos × documentos em outras três matrizes (T, S e D).

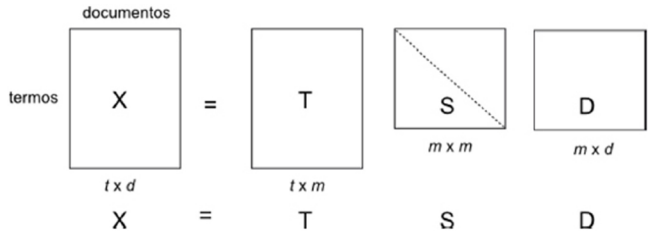


Figura 2.4: Matriz original X como um produto de três matrizes (Deerwester et al., 1990).

O próximo passo é ilustrado na Figura 2.5, em que a dimensão das matrizes é reduzida, eliminando as linhas e colunas correspondentes aos menores valores da matriz S, assim como as colunas da matriz T e as linhas da matriz D. Para a redução das dimensões, reduz-se o número m de dimensões para um valor $k < m$ e, dessa maneira, a matriz reduzida S' afeta diretamente nas dimensões das matriz T e D.

Com o produto de $TS'D$ obtém-se os elementos mais significativos da matriz e com um custo computacional reduzido para seu processamento, já que matrizes com um número grande de dimensões podem necessitar de horas ou até dias para seu processamento (Miller, 2003).

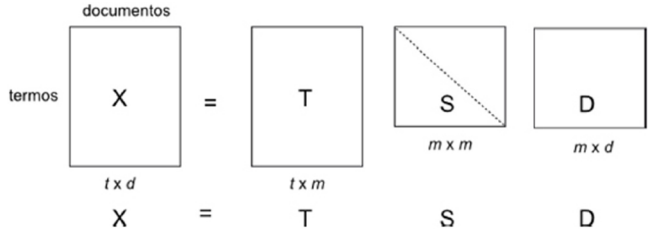


Figura 2.5: Processo de redução de dimensões da matriz criada pela SVD (Deerwester et al., 1990).

Encontrar o número correto de dimensões k para ser utilizado na redução de dimensionalidade do vetor S é uma tarefa crucial e difícil para a SVD. Se o número de dimensões for considerado pequeno, pode-se perder informações importantes e, se considerado grande, pode-se adicionar informações sem relevância. Trabalhos empíricos envolvendo córpus volumosos apresentam valores entre 200 e 400 dimensões (Landauer et al., 1998; Wade-Stein e Kintsch, 2004).

A partir da representação em forma de vetores permitida pela LSA, pode-se medir a similaridade de conceitos relacionados entre duas palavras ou sentenças, calculando-se o produto co-seno (ou produto cartesiano) entre os vetores que representam tais palavras, sentenças ou textos. Ljungstrand e Johansson (1998) definem essa medida de similaridade por meio da fórmula:

$$sim(X, Y) = \frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}}$$

em que $X = (x1, x2, ..., xn)$ e $Y = (y1, y2, ..., yn)$ são dois vetores com n dimensões que representam os textos a serem comparados utilizando o modelo *bag of words*. O resultado apresenta valores limitados entre [-1, 1], sendo -1 o menor valor possível de similaridade e 1 o mais alto grau de similaridade.

A Figura 2.6 ilustra a comparação entre três vetores que possuem respectivamente as palavras “casa”, “sacada” e “corrida”. Observa-se que as palavras “casa” e “sacada” possuem 0,65 de similaridade, enquanto as palavras “casa” e “corrida”, assim como “sacada” e “corrida” possuem 0,01 de similaridade. Esses valores se devem ao fato das palavras “casa” e “sacada” estarem presentes em contextos semelhantes, ao contrário dos outros pares de palavras.

A similaridade entre duas cadeias pode ser alta mesmo que existam poucas palavras em comum entre elas. Considere o exemplo em que são analisadas três sentenças, duas com conceitos relacionados, mas com palavras distintas, e uma terceira sentença com conceito não relacionado às outras duas sentenças, mas com uma palavra semelhante a uma delas. Na Tabela 2.3 são apresentados os valores finais[2] (co-seno) da comparação entre os vetores de termos.

[2]Valores da análise em inglês, obtidos no endereço: `http://lsa.colorado.edu/cgi-bin/LSA-matrix.html`. Os valores obtidos variam de acordo com o córpus.

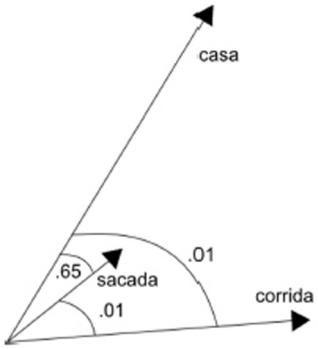


Figura 2.6: Comparação LSA entre três palavras (Wade-Stein e Kintsch, 2004).

Tabela 2.3: Valores atribuídos pela LSA na comparação entre três sentenças.

	"O raio da esfera"	"O diâmetro do círculo"	"A música das esferas"
"O raio da esfera"	1	0,66	0,08
"O diâmetro do círculo"	0,66	1	0,15
"A música das esferas"	0,08	0,15	1

Na Tabela 2.3, a diagonal da matriz apresenta os valores da comparação entre as mesmas sentenças e por isso o nível máximo de similaridade é alcançado com o valor 1. Também pode-se observar que a comparação entre as sentenças "O raio da esfera" e "O diâmetro do círculo" apresentam alto valor de similaridade (0,66), já que são sentenças de um mesmo contexto. Por outro lado, ao comparar a primeira sentença com a terceira, "A música das esferas", o valor de similaridade é de apenas 0,08, mostrando que, embora as sentenças apresentem a mesma palavra "esfera", o seu uso é utilizado em contextos diferentes.

Como já dito, inicialmente a LSA foi desenvolvida no contexto de Recuperação de Informação com a *Latent Semantic Indexing* (LSI), mas recentemente tem sido aplicada nas mais diversas aplicações, como em ferramentas de *Automated Essay Scoring* (Franzke et al., 2005; Higgins et al., 2004; Landauer et al., 2003), em que, por exemplo, se compara o texto avaliado com um texto base escrito por um profissional e a nota emitida leva em consideração a similaridade semântica entre os textos.

A LSA possui algumas limitações, como o fato de não considerar a ordem das palavras, relações sintáticas ou lógicas, nem morfológicas. Assim, frases do tipo "João ama Maria" e "Maria ama João" são interpretadas pela LSA como sendo iguais e o valor do co-seno atribuído a elas é 1, o que não é verdade. Mesmo com tais limitações, a LSA em geral apresenta bons resultados nos contextos em que é aplicada.

2.2.3 Teoria de Centering

Centering (Grosz et al., 1995) é uma teoria discursiva voltada para a avaliação da coerência local do discurso, partindo do princípio da compatibilidade entre os centros

(ou focos) de atenção dos enunciados e na escolha de expressões de referenciamento. Essa teoria tem sido utilizada em várias aplicações de PLN, tais como geração automática de texto (Karamanis, 2004), sumarização automática (Orăsan, 2003) e na avaliação de textos (Barzilay e Lapata, 2008; Hasler, 2004, 2008; Miltsakaki e Kukich, 2004), tratando aspectos relacionados à coerência local, como a resolução de anáforas (Abraços e Lopes, 1994; Palomar e Martínez-Barco, 2001; Tetreault, 2001).

Os autores da teoria propõem que o discurso é composto por segmentos textuais que, por sua vez, são compostos por uma sequência de enunciados e que tais enunciados possuem *centros de atenção*, que são as entidades mais salientes das frases e que servem de ligação entre um enunciado e outro no segmento do discurso. A saliência de uma entidade normalmente reflete suas funções gramaticais e formas linguísticas, sendo as entidades mais salientes mais prováveis de aparecerem em posições sintáticas proeminentes, como sujeitos e objetos, conforme afirma Grosz et al. (1995).

Assumindo que a coerência global está relacionada aos diversos segmentos do texto, enquanto a coerência local está relacionada aos enunciados adjacentes de um mesmo segmento, a teoria de *Centering* apresenta um modelo para o tratamento da coerência local, com um sistema de restrições e regras que governam as relações entre os centros de atenção e as formas escolhidas para construção dos enunciados.

Adaptamos o exemplo de Grosz et al. (1995), no qual temos o segmento da Figura 2.7 com quatro enunciados (*utterances* ou U_i), sendo que o último enunciado do segmento pode ser formado por (U4a) ou (U4b):

(U1) João foi à sua loja de música favorita para comprar um piano.
(U2) Ele frequentava a loja há vários anos.
(U3) Estava animado por finalmente poder comprar um piano.
(U4a) Ele chegou justo quando a loja fechava.
(U4b) Ela estava fechando assim que João chegou.

Figura 2.7: Exemplo de enunciados adaptados de Grosz et al. (1995).

Segundo os autores, o segmento S é intuitivamente mais coerente se finalizado com o enunciado (U_{4a}) em vez de (U_{4b}), embora expressem exatamente a mesma informação. Isso se deve ao enunciado (U_{4a}) formar um segmento com um único foco de atenção, "João", enquanto o uso do enunciado (U_{4b}) faria o foco de atenção oscilar entre "João" e "loja de música". A teoria de *Centering* foi desenvolvida para capturar essas diferenças de continuidade.

Os enunciados que compõem um segmento textual são representados pela sequência U_i – U_n. Para cada enunciado U_i é associado um conjunto ordenado de entidades de contexto do discurso, denominados *Forward-Looking Centers* e representados por $C_f(U_i)$. Esse conjunto representa os potenciais centros de atenção dos próximos enunciados. Por

exemplo, no enunciado (U_1) do segmento S apresentado anteriormente, identificamos o seguinte conjunto de entidades $C_f(U_1)$ = {João, loja de música, piano}. A ordenação do conjunto de entidades é de vital importância para a teoria. Um critério de ordenação usual é a função gramatical subcategorizada pelo verbo principal: sujeito > objeto direto > objeto indireto > outras subcategorizações > adjuntos.

O primeiro elemento do conjunto $C_f(U_i)$ é o mais saliente e é denominado *Preferred Center*, sendo representado por $C_p(U_i)$. De acordo com a teoria, em um segmento com maior nível de coerência, o elemento C_p do enunciado (U_i) deve ser o centro de atenção do enunciado (U_{i+1}). No exemplo do segmento S, $C_p(U_1)$ = {João}.

Outro elemento pertencente ao conjunto $C_f(U_i)$ é o *Backward-Looking Center*, representado por $C_b(U_i)$. O $C_b(U_i)$ estabelece uma relação coerente com o enunciado imediatamente anterior, desde que o enunciado corrente não seja o primeiro do segmento e, portanto, possui $C_b(U_i)$ = *{vazio}*. Podemos dizer que $C_b(U_i)$ é uma referência a um elemento de $C_f(U_{i-1})$, desde que $i > 1$, sendo de fato a entidade central do enunciado anterior. No exemplo do segmento S, $C_b(U_2)$ = {João}, já que no enunciado (U_2) o pronome "Ele" faz uma referência à "João", que se encontra no conjunto $C_f(U_{i-1})$.

Segundo Miltsakaki e Kukich (2004), espera-se que $C_b(U_i)$ e $C_p(U_{i-1})$ sejam idênticos em uma sequência de enunciados, não sendo uma condição obrigatória, porém satisfatória e que contribui para o grau de coerência local do discurso. Ainda segundo os autores, a relação entre esses conjuntos em enunciados subsequentes caracteriza sempre uma de quatro transições definidas na teoria: *Continue, Retain, Smooth-Shift* e *Rough-Shift*. Essas transições se dão pelas relações de equivalência entre os conjuntos C_b e C_p, como pode ser observado na Tabela 2.4.

Tabela 2.4: Relações de equivalência entre os conjuntos C_b e C_p e as transições da teoria de *Centering* (Walker et al., 1998).

	$C_b(U_i) = C_b(U_{i-1})$ ou $C_b(U_{i-1}) =$ *{vazio}*	$C_b(U_i) \neq C_b(U_{i-1})$
$C_b(U_i) = C_p(U_i)$	*Continue*	*Smooth-Shift*
$C_b(U_i) \neq C_p(U_i)$	*Retain*	*Rough-Shift*

A teoria considera que essas transições definem o grau de coerência do discurso, sendo que a transição *Continue* é preferível à transição *Retain*, que por sua vez é preferível à *Smooth-Shift*, que é preferível à transição *Rough-Shift*. No entanto, cabe ressaltar que embora um segmento formado somente por transições *Continue* seja considerado coerente pela teoria, sua leitura não é agradável, já que não há progressão.

Além das transições, a teoria de *Centering* apresenta um conjunto de restrições e regras quanto a forma como os centros de atenção podem ser utilizados para a composição de um texto coerente, sendo que algumas dessas informações já foram apresentadas anteriormente:

Restrições:

1. Cada enunciado (U_i) possui somente um C_b;
2. Todos os elementos de $C_f(U_i)$ são realizados em (U_i); e
3. O $C_b(U_i)$ é o elemento mais proeminente em $C_f(U_{i-1})$ e que é realizado no enunciado (U_i).

Regras:

1. Se um elemento de $C_f(U_{i-1})$ é realizado por um pronome em (U_i), então $C_b(U_i)$ também deve ser realizado por um pronome; e
2. As transições possuem uma ordem de preferência:

$$\textit{Continue} > \textit{Retain} > \textit{Smooth-Shift} > \textit{Rough-Shift}.$$

Dadas as definições da teoria, podemos justificar a afirmação feita no início desta seção em relação ao segmento S. Para isso, apresentamos as entidades identificadas em cada enunciado do segmento e suas respectivas transições nas Tabelas 2.5, 2.6, 2.7, 2.8 e 2.9. Vale lembrar que as transições ocorrem entre pares de enunciados adjacentes ao se verificar a relação entre os conjuntos C_b e C_p dos mesmos, de acordo com a Tabela 2.4.

Tabela 2.5: Entidades e transição identificada no enunciado (U_1).

João foi à sua loja de música favorita para comprar um piano.		
$C_f(U_1)$ = {João, loja de música, piano}	$C_p(U_1)$ = {João}	$C_b(U_1) = \varnothing$
Transição = $\varnothing$		

Tabela 2.6: Entidades e transição identificada no enunciado (U_2).

Ele frequentava a loja há vários anos. (Ele = João)		
$C_f(U_2)$ = {João, loja de música}	$C_p(U_2)$ = {João}	$C_b(U_2)$ = {João}
Transição(U_1 - U_2) = *Continue*		

Tabela 2.7: Entidades e transição identificada no enunciado (U_3).

Estava animado por finalmente poder comprar um piano. (Estava = João)		
$C_f(U_3)$ = {João, piano}	$C_p(U_3)$ = {João}	$C_b(U_3)$ = {João}
Transição(U_2 - U_3) = *Continue*		

Tabela 2.8: Entidades e transição identificada no enunciado (U_{4a}).

Ele chegou justo quando a loja fechava. (Ele = João)		
$C_f(U_{4a})$ = {João, loja de música}	$C_p(U_{4a})$ = {João}	$C_b(U_{4a})$ = {João}
Transição(U_3 - U_{4a}) = *Continue*		

Tabela 2.9: Entidades e transição identificada no enunciado (U_{4b}).

Ela estava fechando assim que João chegou. (Ela = loja de música)		
$C_f(U_{4b})$ = {loja de música, João}	$C_p(U_{4b})$ = {loja de música}	$C_b(U_{4b})$ = {João}
Transição(U_3 - U_{4b}) = *Retain*		

Desse modo, a afirmação anteriormente apresentada de que o segmento S é intuitivamente mais coerente se finalizado com o enunciado (U_{4a}) em vez de (U_{4b}) se comprova na teoria de *Centering* pelo fato da transição entre U_3 - U_{4a} ser do tipo *Continue*, como visto na Tabela 2.8, enquanto a transição entre U_3 - U_{4b} é do tipo *Retain*, como visto na Tabela 2.9, o qual o assunto é mantido, no entanto, demonstra-se a intenção de introduzir no discurso um novo assunto.

2.2.4 Modelo Grade de Entidades

O modelo Grade de Entidades proposto por Barzilay e Lapata (2008) também tem seu foco na coerência local do texto, uma vez que analisa o relacionamento das transições entre sentenças adjacentes. Embora trate de coerência local, os autores afirmam que esse nível é incontestavelmente necessário para se conseguir a coerência global.

A premissa fundamental do modelo é que textos com coerência local possuem certas regularidades na distribuição de substantivos e pronomes, chamados de entidades. Tal premissa é baseada na teoria de *Centering* e em outras teorias de discurso baseada em entidades (e.g., Givon (1987), Prince (1981) apud Barzilay e Lapata (2008)).

No modelo, cada texto é representado por uma matriz ou grade de entidades (denominada *grid*) que captura a distribuição das entidades do discurso nas sentenças do texto. Nessa grade, as linhas correspondem às sentenças do texto e as colunas às entidades identificadas. Por entidade do discurso se entende uma classe de sintagmas nominais correferentes. Caso não seja possível utilizar resolução de correferência na identificação das entidades, cada sintagma nominal identificado corresponderá a uma entidade na grade.

Para cada ocorrência de uma entidade no texto, a célula correspondente da grade contêm informações sobre sua presença ou ausência na sequência de sentenças, bem como informações sobre o papel gramatical do sintagma nominal que a representa na sentença. Dessa forma, cada célula da grade é preenchida por uma letra que representa se a entidade

em questão aparece no papel de sujeito (S), objeto (O) ou nenhum (X). Entidades ausentes na sentença são sinalizadas por lacunas ($-$).

Considere o exemplo da Figura 2.8 extraído de Barzilay e Lapata (2008), que apresenta um texto de seis sentenças com as respectivas anotações sintáticas:

(1) [The Justice Department]$_S$ is conducting an [anti-trust trial]$_O$ against [Microsoft Corp.]$_X$ with [evidence]$_X$ that [the company]$_S$ is increasingly 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]$_O$.
(3) [The case]$_S$ revolves around [evidence]$_O$ of [Microsoft]$_S$ aggressively pressuring [Netscape]$_O$ into merging [browser software]$_O$.
(4) [Microsoft]$_S$ claims [its tactics]$_S$ are commonplace and good economically.
(5) [The government]$_S$ may file [a civil suit]$_O$ ruling that [conspiracy]$_S$ to curb [competition]$_O$ through [collusion]$_X$ is [a violation of the Sherman Act]$_O$.
(6) [Microsoft]$_S$ continues to show [increased earnings]$_O$ despite [the trial]$_X$.

Figura 2.8: Exemplo de sentenças com anotações sintáticas, extraído de Barzilay e Lapata (2008).

A Figura 2.9 representa um fragmento da grade de entidades extraída a partir das seis sentenças da Figura 2.8.

	Department	Trial	Microsoft	Evidence	Competitors	Markets	Products	Brands	Case	Netscape	Software	Tactics	Government	Suit	Earnings	
1	S	O	S	X	O	–	–	–	–	–	–	–	–	–	–	1
2	–	–	O	–	–	X	S	O	–	–	–	–	–	–	–	2
3	–	–	S	O	–	–	–	–	S	O	O	–	–	–	–	3
4	–	–	S	–	–	–	–	–	–	–	–	S	–	–	–	4
5	–	–	–	–	–	–	–	–	–	–	–	–	S	O	–	5
6	–	X	S	–	–	–	–	–	–	–	–	–	–	–	O	6

Figura 2.9: Fragmento da grade de entidades extraída a partir das sentenças da Figura 2.8, extraído de Barzilay e Lapata (2008).

Como o exemplo da Figura 2.8 possui seis sentenças, a grade de entidades representada na Figura 2.9 também possui seis linhas. Considere, por exemplo, a coluna da grade referente a entidade *Trial*, $[O - - - - X]$. Este registo representa que a entidade *Trial* está presente nas sentenças 1 e 6 (como O e X, respectivamente), mas ausente nas outras sentenças. É importante destacar que neste exemplo a grade de entidades foi construída utilizando-se resolução de correferência na fase de identificação das entidades. Nesse caso, mesmo que uma entidade ocorra de diferentes formas linguísticas no texto, por exemplo,

Microsoft Corp., *Microsoft* e *the company*, ela é mapeada como uma entrada única na grade, como é o caso da coluna *Microsoft* $[SOSS - S]$.

Uma particularidade do processo de construção da grade ocorre quando uma mesma entidade é referenciada mais de uma vez na mesma sentença e com diferentes funções gramaticais. Seguindo a ideia da ordenação dos centros de atenção da teoria de *Centering*, escolhe-se para ser uma entrada da grade a entidade que possuir a função gramatical mais proeminente, sendo considerada a seguinte ordem de proeminência: sujeito > objeto > outras funções gramaticais. Por exemplo, a entidade *Microsoft* é mencionada duas vezes na Sentença 1 com funções gramaticais diferentes, $[\textit{Microsoft Corp.}]_X$ e $[\textit{the company}]_S$, mas é representada somente por S na célula correspondente da grade, já que sujeito (S) é a função gramatical mais proeminente.

O modelo considera que grades de entidades extraídas a partir de textos coerentes são suscetíveis a possuírem algumas colunas densas, ou seja, em que a maior parte de seus elementos são diferentes de lacunas ($-$), como é o caso da coluna da entidade *Microsoft* na Figura 2.9, e muitas colunas esparsas constituídas principalmente de lacunas, como é o caso das entidades *Markets* e *Earnings*. Também espera-se que as funções sintáticas sujeito (S) e objeto (O) apareçam em maior número nas colunas densas, pois essas características são menos presentes em textos com baixa coerência.

Assim como na teoria de *Centering*, o modelo se baseia nos padrões das transições de entidades locais. Uma transição de entidade local é uma sequência $\{S, O, X, -\}_n$ que representa as ocorrências da entidade e suas funções sintáticas em n sentenças adjacentes. As transições locais podem ser facilmente obtidas a partir da grade de entidades como subsequências contínuas de cada coluna e possuem uma certa probabilidade de ocorrência na grade. Por exemplo, a probabilidade de ocorrer a transição $[S-]$ na grade da Figura 2.9 é 0,08, já que a transição ocorre seis vezes em um total de 75 possíveis transições de tamanho 2, logo $6 \div 75 = 0{,}08$.

Dessa maneira, cada texto pode ser representado por um conjunto fixo de transições locais e suas probabilidades, usando a notação padrão de vetor de características. O vetor de características extraído de uma grade de entidades x_{ij} que representada um documento d_i é dado por:

$$\Phi(x_{ij}) = (p_1(x_{ij}), p_2(x_{ij}), ..., p_m(x_{ij}))$$

em que m é o número de todas as transições locais pré-definidas e $p_t(x_{ij})$ é a probabilidade da transição t na grade x_{ij}.

A representação por meio de vetor de características é bastante útil, pois permite a utilização dos dados por algoritmos de aprendizagem de máquina, que por sua vez podem descobrir novos padrões relevantes de distribuição de entidades a serem utilizados na avaliação de coerência.

Como se pode notar, existe um grande número de transições passíveis de serem incluídas no vetor de características (todas as de tamanho dois, três, quatro e assim por diante). Uma forma de limitar o tamanho do vetor de características é considerar apenas transições de um dado tamanho, normalmente tamanho dois ou três, ou verificar apenas as transições mais frequentes nos documentos. Isso impede que haja uma explosão do número de características, o que causaria um aumento excessívo da quantidade necessária de dados de treinamento.

Na Tabela 2.10 é apresentado um exemplo com espaço de transições de tamanho dois, em que a primeira linha (d_1) representa as probabilidades calculadas a partir da grade de entidades da Figura 2.9, enquanto o restante das linhas ilustram os valores de outros documentos (d_2 e d_3).

Tabela 2.10: Vetores de características com todas as transições de tamanho dois dadas as categorias sintáticas {S, O, X, -} para os documentos d_1, d_2 e d_3, adaptado de Barzilay e 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
d_3	0.2	0	0	.03	.09	0	.09	.06	0	0	0	.05	.03	.07	.17	.39

Um outro aspecto que pode ser incorporado à grade de entidades é a saliência das entidades. Barzilay e Lapata (2008) definem a saliência de uma entidade com base na frequência de ocorrência da entidade no discurso. 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 (probabilidades das transições para entidades salientes vs. probabilidades das transições para entidades não salientes).

Em resumo, três tipos de conhecimento linguístico podem ser considerados na construção da grade de entidade: (i) informações sobre correferência, (ii) informações sobre a função sintática e (iii) informações sobre saliência. Em relação a (i) correferência, essa informação pode ser considerada ou não no momento da identificação das entidades. Em relação (ii) função sintática, essa informação pode ser considerada ou não no momento da representação da ocorrência da entidade em uma sentença (linha da grade). Se a função sintática for considerado, a grade será como exemplificado na Figura 2.9; caso contrário, a grade marcará apenas a presença (P) ou ausência (-) das entidades nas sentenças representadas. Em relação a (iii) saliência, a grade pode ser composta por todas as entidades do discurso, caso em a informação de saliência não está sendo considerada, ou a grade pode ser dividida em duas – uma composta apenas de entidades salientes e uma composta de entidades não salientes. Nesse caso, as probabilidades das transições são computadas separadamente para cada grade, mas todas são incluídas no vetor de

características que presenta o texto. Dessa forma, considerando a presença (+) ou a ausência (-) desses três tipos de informação, oito configurações diferentes da grade de entidades podem ser obtidas fazendo-se as combinações de correferência(+/-), função sintática (+/-) e saliência (+/-). Cabe ressaltar que a presença ou não de tais informações na grade de entidades afetará diretamente a configuração do vetor de características e probabilidades das transições.

Avaliação do Modelo

Barzilay e Lapata (2008) demonstraram o uso do modelo Grade de Entidades em três diferentes experimentos: (1) ordenação de sentenças, (2) avaliação de coerência em sumários gerados automaticamente e (3) avaliação de inteligibilidade. Os três experimentos utilizaram os três tipos de conhecimento linguístico: o sintático, a correferência e a saliência.

O experimento (1) abordou a tarefa de ordenação de sentenças, que é essencial para uma série de problemas de PLN, tais como geração e sumarização automática. Desse modo, o modelo grade de entidades foi usado para ranquear diferentes sequências de sentenças de um mesmo texto, esperando que as sequências mais coerentes ficassem no topo do *ranquing*. Para esse experimento, os dados usados para treinamento e teste do modelo foram gerados sinteticamente a partir de 200 documentos, sendo 100 extraídos de um corpus jornalístico (especificamente textos sobre terremotos) e 100 extraídos de um corpus de narrativas técnicas sobre acidentes aéreos. Para cada documento original foram geradas aproximadamente 20 versões sintéticas contendo permutações aleatórias das sentenças. Com isso assumiu-se que o texto com as sentenças na ordem original deve ser mais coerente que a maioria dos textos com as sentenças permutadas. Assim, o conjunto final de textos usados no treinamento foi de aproximadamente 4.000 textos. Cada texto teve em média 10,4 e 11,5 sentenças, respectivamente. Oito modelos foram construídos, um para cada configuração possível, e todos foram treinados usando o pacote SVM^{light} (Joachims, 1999) que implementa o algoritmo SVM (*Support Vector Machine*), sendo a aprendizagem formulada como um problema de ranqueamento (*ranking learning problem*). Para se determinar a acurácia do modelo, foi analisado o *ranquing* dado para cada par texto original vs. permutação. Em ambos os domínios (Terremotos e Acidentes), o modelo completo `Correferência+ Sintático+ Saliência+` obteve o melhor desempenho, alcançando uma acurácia de 87,2% para o córpus Terremoto e 90,4% para o córpus Acidentes. Os resultados completos desse experimento mais o resultado obtido por meio de uma *baseline* que utiliza LSA são apresentados na Tabela 2.11.

O experimento (2) abordou a tarefa de avaliação de coerência em sumários multidocumentos gerados automaticamente. Os dados usados nesse experimento foram sumários automaticamente gerados por seis sistemas diferentes de sumarização automática multidocumentos, todos provenientes da conferência DUC 2003. Esses sumários foram gerados a

Tabela 2.11: Experimento (1): Acurácia medida como a porcentagem de ranqueamentos corretos entre pares de texto no conjunto de teste (adaptado de Barzilay e Lapata (2008)).

Modelo	Terremotos	Acidentes
Correferência+ Sintático+ Saliência+	87,2	90,4
Correferência+ Sintático+ Saliência-	88,3	90,1
Correferência+ Sintático- Saliência+	86,6	88,4
Correferência- Sintático+ Saliência+	83,0	89,9
Correferência+ Sintático- Saliência-	86,1	89,2
Correferência- Sintático+ Saliência-	82,3	88,6
Correferência- Sintático- Saliência+	83,0	86,5
Correferência- Sintático- Saliência-	81,4	86,0
Latent Semantic Analysis	81,0	87,3

partir de 16 agrupamentos de documentos aleatoriamente selecionados. Além dos sumários automaticamente gerados, também foram usados os sumários gerados por humanos. Todos os sumários foram julgados quanto a coerência por juízes humanos, tendo sido atribuída uma nota entre 1 e 7 para cada sumário. A partir desses sumários avaliados por humanos, foi construído um conjunto de 144 pares de sumários usados para o treinamento, mais 80 pares usados para o teste. Mais uma vez foram construídos oito modelos, um para cada configuração possível, e todos foram treinados usando o SVM^{light} (Joachims, 1999) com a aprendizagem modelada como um problema de ranqueamento. O modelo `Correferência- Sintático+ Saliência+` obteve o melhor desempenho, alcançando uma acurácia de 83,8%. Os resultados completos desse experimento mais o resultado obtido por meio de uma baseline que utiliza LSA são apresentados na Tabela 2.12.

Tabela 2.12: Experimento (2): Acurácia medida como a porcentagem de ranqueamentos corretos entre pares de texto no conjunto de teste (adaptado de Barzilay e Lapata (2008)).

Modelo	Acurácia
Correferência+ Sintático+ Saliência+	80,0
Correferência+ Sintático+ Saliência-	75,0
Correferência+ Sintático- Saliência+	78,8
Correferência- Sintático+ Saliência+	83,8
Correferência+ Sintático- Saliência-	71,3
Correferência- Sintático+ Saliência-	78,8
Correferência- Sintático- Saliência+	77,5
Correferência- Sintático- Saliência-	73,8
Latent Semantic Analysis	52,5

O experimento (3) abordou a tarefa de avaliação de inteligibilidade. Mais especificamente, buscou-se avaliar se o modelo grade de entidades seria útil para um sistema que avaliasse a inteligibilidade de textos (segundo Barzilay e Lapata (2008), inteligibilidade indica a facilidade com que um documento pode ser lido e compreendido). A premissa é que informações sobre a coerência seriam um bom indicativo do nível de inteligibilidade. Diferente dos experimentos (1) e (2) em que a tarefa foi modelada como um problema de ranqueamento, nesse experimento a tarefa foi modelada como um problema de classificação. A unidade de classificação foi um texto completo e a tarefa foi predizer se o texto era fácil ou difícil de ler. Como dados para esse experimento foram utilizados textos coletados da *Encyclopedia Britannica* e da *Britannica Elementary*, sendo a última uma versão da primeira voltada para crianças. Foram coletados 107 artigos da *Encyclopedia Britannica* e suas versões correspondentes na *Britannica Elementary*, totalizando 214 artigos. Os textos da *Encyclopedia Britannica* foram anotados como "difícil de ler" enquanto as versões da *Britannica Elementary* foram anotados como "fácil de ler". Nesse experimento, além das diferentes configurações do modelo grade de entidades, também foram testados atributos retirados do trabalho de Schwarm e Ostendorf (2005), que foram desenvolvidos especificamente para avaliação de inteligibilidade. Todos os modelos foram treinados usando o SVM^{light} (Joachims, 1999) com o caso formulado como um problema de classificação binário. Os testes foram realizados usando-se 5-*fold cross-validation*. O melhor desempenho foi obtido pelo modelo `Correferência- Sintático+ Saliência+` combinado com o modelo de Schwarm e Ostendorf (2005), alcançando acurácia de 88,79%. Vale ressaltar que o modelo de Schwarm e Ostendorf (2005) sozinho obteve acurácia de 78,56%. Quando testado sozinho, modelo grade de entidades obteve o melhor desempenho na configuração `Correferência+ Sintático+ Saliência+`, alcançando acurácia de 50,9%.

2.2.5 Trabalhos Relacionados ao Modelo Grade de Entidades

Filippova e Strube (2007)

Barzilay e Lapata (2008) sugeriram que a integração de conhecimento semântico para o agrupamento de entidades (em oposição a resolução de correferência) poderia melhorar os resultados do modelo grade de entidades. Seguindo essa linha, Filippova e Strube (2007) realizaram um estudo usando a API *WikiRelate!* (Strube e Ponzetto, 2006) para calcular relações semânticas entre entidades. A escolha pela *WikiRelate!* como fonte de conhecimento semântico se deu pelo fato do córpus utilizado ser de textos jornalísticos, que em geral contêm entidades que são nomes próprios (pessoas, locais, empresas) e a Wikipédia seria então a melhor escolha. Desse modo, o trabalho de Filippova e Strube (2007) teve dois objetivos: verificar se a integração de conhecimento semântico de fato melhora os resultados atingidos usando-se correferência; e verificar se o uso do

relacionamento semântico é confiável para agrupamento de entidades no caso de não haver disponibilidade de resolução de correferência.

Os testes foram realizados nos mesmos moldes do experimento (1) de Barzilay e Lapata (2008), com a diferença do corpus utilizado para gerar as versões permutadas – 100 textos jornalísticos em alemão com anotação manual de funções sintáticas e resolução manual de correferência. Utilizando o processo original para a identificação e agrupamento das entidades, o melhor desempenho foi alcançado com o modelo `Correferência+ Sintático- Saliência+`, com acurácia de 75%. Os resultados utilizando a *WikiRelate!* para agrupar entidades relacionadas ficaram em torno de 5% abaixo dos modelos que usam correferência, mostrando que o uso da *WikiRelate!* no processo de agrupamento de entidades é melhor do que não usar informação nenhuma (`Correferência-`), mas não é tão bom quanto o uso de informação de correferência.

Yokono e Okumura (2010)

Yokono e Okumura (2010) estenderam o modelo grade de entidades original visando sua aplicação para a língua japonesa por meio da adição de atributos baseados em mecanismos coesivos. Três mecanismos coesivos foram considerados: conjunções, referência por meio de pronomes demonstrativos e coesão léxica. A representação das entidades na grade por meio de funções sintáticas também foi refinada pela adição de marcadores de tópico específicos da língua japonesa.

O tratamento das conjunções foi feito se considerando a relação entre sentenças adjacentes e criando uma grade de entidades separada para cada grupo de relações de conjunção (três grupos foram utilizados). No tratamento da referência foram considerados apenas os pronomes demonstrativos, sendo que foi incorporado ao modelo a probabilidade de uma entidade correferente ocorrer na sentença anterior a atual. O valor desse atributo é calculado como a proporção do número de casos em que uma sentença anterior a atual, que contém um pronome demonstrativo, contém uma entidade relacionada a esse pronome. No tratamento da coesão léxica foram utilizadas correntes lexicais (*lexical chains*) identificadas pelo sistema de Mochizuki et al. (1999) apud Yokono e Okumura (2010) para o agrupamento de entidades. Nesse caso, a grade de entidades passa a ser uma grade de grupos de entidades.

Foram realizados dois experimentos com 400 textos jornalísticos em japonês nas áreas de política, saúde e educação, sendo os textos separados em dois corpora: um com 100 textos e um com 300 textos. O primeiro experimento se baseou na tarefa de discriminar entre o texto original e sua permutação, no mesmo formato do experimento (1) de Barzilay e Lapata (2008) (conforme descrito na Subseção 2.2.4). Para esse experimento, o modelo estendido teve desempenho superior ao do modelo original. Para o córpus de 100 textos, a acurácia do modelo original foi de 54,7%, enquanto a acurácia do modelo estendido com todos os atributos novos foi de 59,4%. Para o córpus de 300 textos, a acurácia do

modelo original foi de 58% e o melhor desempenho foi obtido pelo modelo estendido com os atributos novos relativos as conjunções e as correntes léxicas, com uma acurácia de 77,3%.

No segundo experimento foram utilizados os resultados da avaliação feita por juízes humanos em sumários criados automaticamente. A nota de cada sumário foi atribuída com base nas respostas dadas a um questionário que aponta pontos problemáticos nos sumários. Desse modo, para um par de sumários extraídos do mesmo texto, aquele com a menor nota é o mais coerente. Os modelos foram treinados com o córpus de 300 textos do experimento anterior e testado com os sumários avaliados manualmente. Para esse experimento, a acurácia do sistema original foi de 50,2% e o melhor desempenho foi obtido pelo modelo estendido com todos os atributos novos, exceto os relativos as conjunções, com acurácia de 58,9%.

Burstein, Tetreault e Andreyev (2010)

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 a textos de um domínio diferente: redações (*essays*) escritas por estudantes de perfis variados. Os novos atributos foram chamados de GUMS - *Grammar, usage, mechanics errors* e *style*) e são extraídos por ferramentas de AES (*Automated Essay Scoring*) previamente desenvolvidas pelo grupo de pesquisa do ETS (*Educational Testing Service*)[3].

Além dos atributos GUMS, Burstein et al. (2010) também utilizaram atributos do tipo *Type/Token* (denominados de *_TT) para medir a variedade léxica das entidades que ocorrem em cada função sintática. Dessa forma, os atributos *_TT são expressos como $\{S, O, X, SOX\}$_TT para o modelo na configuração `Sintático+` e $\{P\}$_TT para o modelo na configuração `Sintático-`. Por exemplo, o atributo S_TT representa a proporção de entidades que aparecem como sujeito (S) em relação ao número total de sujeitos (Ss) observados na grade de entidades. O mesmo foi feito para as outras funções sintáticas. Também foram utilizados atributos do tipo *Type/Token* para representar a proporção do uso de *Shell nouns* (Aktas e Cortes, 2008; Hinkel, 2004) nas três funções sintáticas (S, O, X). *Shell nouns* são substantivos usados como mecanismos de coesão e que comumente fazem referência a algum ponto do texto, como é o caso da palavra "*fact*" na frase "*This fact can be explained...*". O uso dos atributos *_TT_Shellnouns se baseou no fato da coesão ser intimamente relacionada à coerência.

Para testar o desempenho da proposta, três corpora de *essays* diferentes foram utilizados, totalizando aproximadamente 800 textos: (1) Adultos não nativos na língua inglesa (TOEFL); (2) Adultos nativos e não nativos (GRE); e (3) Adolescentes (*high-school*)

[3]http://www.ets.org/

nativos e não nativos que utilizaram uma ferramenta de AES (*Criterion*). Todos os textos foram anotados por dois anotadores humanos treinados, que avaliaram a qualidade da coerência com base em quão fácil era a leitura do texto, isto é, sem "barreiras na coerência" (por ex., sentença(s) confusa(s)). A anotação final foi feita em uma escala binária (baixa coerência vs. alta coerência) e o valor da medida *Kappa* calculada em um subconjunto de 100 textos para dois anotadores foi de 0,677%.

Diferentemente de Barzilay e Lapata (2008), Burstein et al. (2010) modelaram seus experimentos como um problema de classificação. Os testes foram realizados com o algoritmo C5.0 (versão comercial do conhecido algoritmo C4.5) e foi usado *n-fold cross-validation* para testar diferentes configurações do modelo. A saída obtida por meio de um método de votação pela maioria (*majority vote*) também foi considerada. Uma vez que os corpora eram bastante desbalanceados, os resultados foram expressos em termos das medidas *Kappa (K), Precision, Recall e F-measure*. O modelo original foi testado na configuração `Correferência- Sintático+ Saliência+`. Em todas as configurações o modelo estendido teve desempenho melhor do que o modelo original (para os três corpora avaliados). O melhor resultado foi alcançado pelo sistema de votação aplicado a saída de três configurações diferentes do modelo estendido para o córpus (2), com $K = 0,61$ e *F-measure* = 0,91.

Elsner e Charniak (2011)

Elsner e Charniak (2011) estenderam o modelo grade de entidades por meio da adição de atributos entidade-específicos. Enquanto o modelo original trata todas as entidades de forma igualitária, o modelo estendido busca distinguir entre entidades importantes e entidades menos importantes, aplicando um conjunto de oito atributos que mapeiam características relacionadas à proeminência no discurso, tipo de entidade nomeada e correferência.

Elsner e Charniak (2011) também modificaram o processo de identificação de entidades de Barzilay e Lapata (2008), reconhecendo todo nome (substantivo ou nome próprio) como uma entidade em vez de usar apenas os núcleos dos sintagmas nominais (SNs). Na grade de entidades, esses nomes que não são núcleos receberam a função sintática *X* (outros). Também não foi utilizado um resolvedor automático de correferência. Em vez disso, apenas as menções que possuíam o mesmo núcleo (*head noun*) foram consideradas correferentes. Essa nova abordagem na identificação de entidades resultou em um aumento de 4% no desempenho quando comparado a um modelo original na configuração `Correferência-`.

Dois tipos de experimentos foram realizados utilizando um córpus de 1004 artigos do WSJ (*Wall Street Journal*). O primeiro experimento se baseou na tarefa de discriminar entre o texto original e sua permutação, no mesmo formato do experimento (1) de Barzilay e Lapata (2008). O segundo experimento se baseou na tarefa de inserção de sentença

(Elsner e Charniak, 2008 apud Elsner e Charniak (2011)), que consiste em remover cada sentença de um texto e testar se o modelo prefere reinserir a sentença na sua posição original. Em ambos os experimentos, o modelo estendido foi superior ao modelo original. No primeiro experimento, o modelo original obteve acurácia de 79,5%, enquanto o modelo estendido obteve acurácia de 84%. Já no segundo experimento, a média de acerto do modelo original foi de 21% e a do modelo estendido foi de 24%.

Lin, Tou Ng e Kan (2011)

Diferente dos trabalhos anteriores em que a proposta foi estender o modelo original com algum tipo de atributo adicional, Lin et al. (2011) propuseram um modelo novo, que combina o modelo grade de entidades com relações discursivas semelhantes as da RST (Mann e Thompson, 1988). Esse novo modelo também usa uma grade de entidades (sentenças x entidades), mas em vez das entidades serem representadas na grade por seus papeis sintáticos (S, O, X), são representadas pelo tipo de relação retórica em que aparecem (*Temporal*, *Contingency*, *Comparison* e *Expansion*). Desse modo, a grade de entidades é utilizada para se calcular as probabilidades de transições entre relações retóricas em vez de transições entre funções sintáticas. Para a identificação das relações retóricas, Lin et al. (2011) utilizaram a D-LTAG (*Discourse Lexicalized Tree Adjoining Grammar*) (Webber, 2004 apud Lin et al. (2011)), conforme a marcação utilizada no PDTB (*The Penn Discourse Treebank*) (Prasad et al., 2008 apud Lin et al. (2011)).

Os experimentos foram realizados no mesmo formato do experimento (1) de Barzilay e Lapata (2008), utilizando, inclusive, os mesmos corpora (Acidentes e Terremotos). Experimentos adicionais utilizando um córpus de 1040 artigos do WSJ também foram realizados. Os resultados do modelo original de Barzilay e Lapata (2008) na configuração `Sintático+ Saliência+` foram utilizados como *baseline*. Os resultados mostraram que o novo modelo superou o modelo original para os córpora WSJ e Terremotos, mas manteve o mesmo desempenho para o córpus Acidentes. Para o córpus WSJ, o novo modelo obteve acurácia de 88%, enquanto a *baseline* obteve acurácia de 85,7%. Para o córpus Terremotos, o novo modelo obteve acurácia de 86,5%, enquanto a *baseline* obteve acurácia de 83,6%. Para o córpus Acidentes, o novo modelo obteve acurácia de 89,9%, enquanto a *baseline* obteve acurácia de 89,4%.

Capítulo

3

Ambiente SciPo

Uma vez que o ambiente SciPo e o seu módulo de análise de coerência (MAC) serviram de motivação para este trabalho, este capítulo apresenta uma visão geral do ambiente, mostrando suas funcionalidades e sua arquitetura. O MAC, um dos elementos da arquitetura do SciPo, também é detalhado neste capítulo.

O SciPo é um sistema *web* cujo o objetivo principal é auxiliar escritores iniciantes na escrita de textos científicos em português. Seu foco principal são os resumos e introduções de textos científicos na área de ciencia da computação, ajudando escritores a estruturarem corretamente seu texto. O SciPo permite que o usuário escolha entre dois modos de trabalho:

1. O processo *top-down* o qual o escritor começa com um texto-exemplo com a estrutura retórica já definida. Esse modo foi herdado do projeto AMADEUS (Aluísio et al., 2001);

2. O processo *bottom-up* o qual o sistema detecta automaticamente e analisa a estrutura retórica do texto submetido à análise.

De fato, esses dois modos são diferentes formas de começar a escrita do texto, porém pertencem ao mesmo processo cíclico de refinamento, dado que a estrutura retórica detectada e avaliada em (2) pode ser melhorada usando os recursos disponíveis em (1). O sistema contém quatro base de conhecimento: (i) base de dados de resumo, (ii) base de dados de de introduções, (iii) medidas e regras de similaridade, e (iv) regras de crítica. A base de dados de resumos contém 52 exemplos de estrutura esquemática extraídas de resumos originais. Cada exemplo conta com a descrição dos componentes retóricos que compõem a estrutura, estratégias e padrões léxicos. Do mesmo modo, a base de

dados de introduções contém 48 exemplos de estrutura esquemática para introduções com descrições dos componentes retóricos, estratégias e padrões léxicos para cada exemplo. Nos dois casos, todas as informações foram anotadas manualmente conforme o modelo de estrutura retórica apropriada (Aluisio e Oliveira, 1996; Feltrim et al., 2003). O usuário pode visualizar livremente esse banco de dados e pesquisar por ocorrências de estruturas retóricas específicas.

Na Tabela 3.1 é mostrado o modelo de estrutura retórica usado em resumos com uma breve descrição da função de cada componente.

Tabela 3.1: Modelo de estrutura retórica para resumos utilizado pelo SciPo.

Componente	Função
Contexto	Apresenta conhecimento aceito pela comunidade científica e que é usado para contextualizar a pesquisa
Lacuna	Apresenta um problema (uma lacuna) em uma área de pesquisa específica que será abordado no estudo em questão (prepara o leitor para o propósito)
Propósito	Apresenta o propósito/objetivo da pesquisa
Metodologia	Apresenta brevemente os materiais e métodos utilizados
Resultado	Apresenta brevemente os principais resultados
Conclusão	Apresenta as principais conclusões/contribuições/limitações da pesquisa

A terceira base de conhecimento, medidas e regras de similaridade, refere-se às regras estabelecidas com base na similaridade entre as listas (encontro de padrões) e na busca pelo vizinho mais próximo(Kriegsman e Barletta, 1993). Essas regras são usadas para recuperar uma determinada estrutura retórica, conforme solicitado pelo usuário.

A quarta base de conhecimento se refere as regras de crítica. Essas regras foram formuladas com base nos modelos estruturais utilizados pelo ambiente SciPo e tentam corrigir padrões de estruturação considerados problemáticos. Foram consideradas sugestões prescritivas de boas estruturas de Resumos e Introduções encontradas na literatura e também os principais desvios observados no corpus, com relação aos modelos estruturais, que possam prejudicar de alguma forma o entendimento do texto. As regras consideram desvios de conteúdo (falta de componentes tidos como essenciais) e desvios de ordem (ordem de ocorrência dos componentes na estrutura), além de diferenciar desvios "graves" (apresentados ao usuário como "críticas") de desvios "leves" (apresentados ao usuário como "sugestões"). Dessa forma, são quatro as classes de regras: críticas referentes (1) ao conteúdo e (2) a ordem, e sugestões de melhoria referentes (3) ao conteúdo e (4) a ordem.

Um quinto elemento da arquitetura SciPo é um classificador textual, chamado AZPort, que detecta automaticamente a estrutura retórica de um resumo. O AZPort é um classificador Naive Bayes que implementa a abordagem de *Argumentative Zoning* proposta por Teufel e Moens (2002), adaptado ao contexto de resumos científicos escrito em

português. Seguindo os componentes estruturais do modelo de estrutura retórica proposto por Feltrim et al. (2003), o AZPort classifica cada sentença do texto em uma das seguintes categorias: Contexto, Lacuna, Propósito, Metodologia, Resultado e Conclusão. Mais detalhes sobre o AZPort podem ser encontrados em Feltrim (2004) e Feltrim et al. (2006). O uso do AZPort é que torna possível incorporar o processo *bottom-up* no ambiente SciPo. A Figura 3.1 apresenta uma visão geral da arquitetura do SciPo, mostrando como os diferentes processos se relacionam com as bases de conhecimento.

Como mostrado na Figura 3.1, uma vez que o usuário decidiu como será a estrutura retórica do texto, a qual pode ter sido tanto detectada automaticamente (processo *bottom-up*) ou construída de modo explícito (processo *top-down*), ele receberá um *feedback* do sistema acerca da estrutura. Esse processo é repetido quantas vezes forem necessárias até que uma estrutura aceitável (pelas regras do ambiente) tenha sido construída. O usuário pode utilizar os exemplos do córpus (bases de resumos e introduções) para reutilizar determinados padrões léxicos em seu texto.

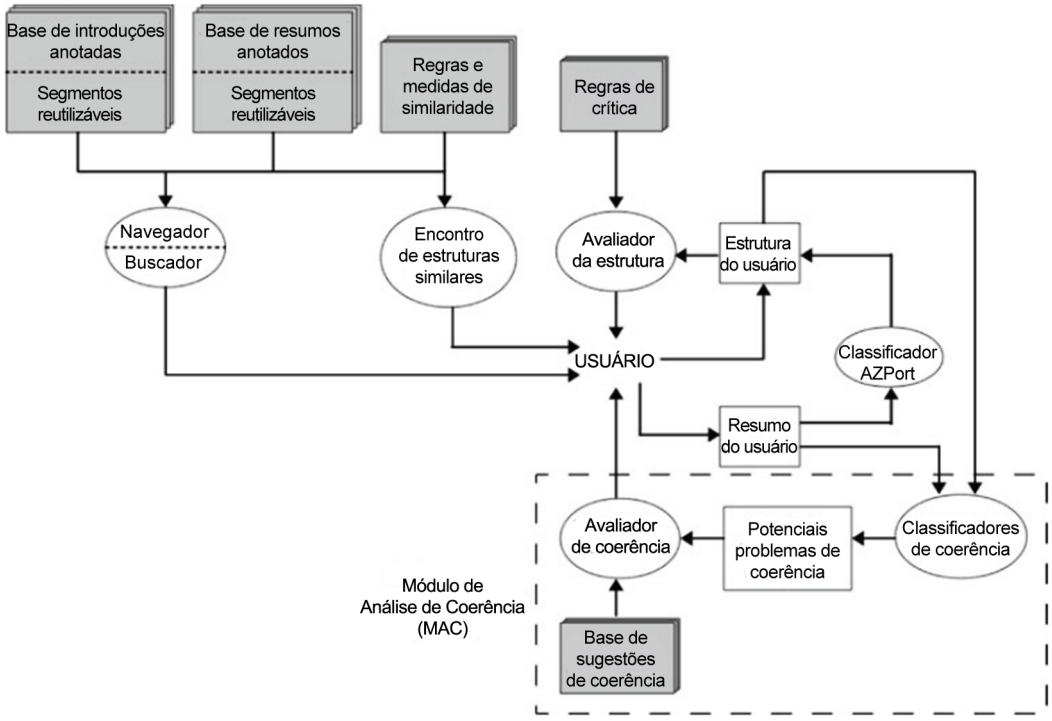


Figura 3.1: Arquitetura do SciPo com o módulo de análise de coerência (MAC) (adaptado de Souza e Feltrim (2012)).

Por último, mas não menos importante, o sexto elemento da arquitetura SciPo é o módulo de análise de coerência (MAC), responsável em por gerar *feedback* a respeito da coerência do texto. Com base no texto do resumo e em sua estrutura retórica, o MAC identifica potenciais problemas relacionados a três tipos de relacionamentos semânticos predefinidos e seleciona, a partir de uma base de sugestões de coerência, o *feedback* apropriado. Conforme comentado na introdução (Cap. 1), basicamente, esse módulo analisa o nível de similaridade semântica existente entre os diferentes

componentes esquemáticos identificados no texto. O nível de similaridade semântica entre dois ou mais componentes esquemáticos é medido por meio de LSA (*Latent Semantic Analysis*)(Landauer et al., 1998). Os três tipos de relacionamentos semânticos ou dimensões examinados, conforme proposto por Souza e Feltrim (2012), são:

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.

Na Figura 3.1 é possível observar a parte referente ao MAC na arquitetura do ambiente SciPo na área tracejada. O MAC é formado por uma base de sugestões de coerência, um conjunto de cinco classificadores de coerência (um para a dimensão Título, três para a dimensão Propósito e um para a dimensão Lacuna-Contexto), e um analisador de coerência, o qual seleciona a(s) sugestão(ões) apropriada(s) com base nos potenciais problemas de coerência encontrados.

O processo de análise de coerência inicia chamando o classificador AZPort para detectar automaticamente a estrutura retórica do resumo. Se algum problema na estrutura esquemática é encontrado, o usuário recebe as críticas e/ou sugestões do SciPo referentes a estrutura, para então fazer a correção. Caso contrário, a estrutura retórica e o texto são enviados para o MAC para o processamento da LSA e a extração de atributos. Baseado nos valores dos atributos extraídos, os cinco classificadores analisam cada sentença do resumo e os resultados são enviados para o analisador de coerência. Caso algum potencial problema de coerência tenha sido detectado, a(s) sugestão(ões) serão selecionadas pelo analisador e apresentadas ao usuário. As sugestões referem-se às três dimensões descritas anteriormente. A ciclo de refinamento continua até que o sistema não tenha mais nenhuma sugestão a oferecer ou até que o usuário decida parar o processo.

É importante salientar que o usuário é livre para rejeitar as sugestões realizadas pelo sistema. De fato, aspectos semânticos são ambíguos por natureza, portanto, não se pode excluir a possibilidade de que o usuário e o MAC discordem a respeito de algum problema de coerência identificado. Seguindo essa linha de raciocínio, decidiu-se deixar o usuário livre para aceitar ou rejeitar as sugestões oferecidas pelo MAC. Essa liberdade de escolha dada ao usuário já era implementada na primeira versão do ambiente SciPo e, desse modo, decidiu-se por mantê-la no MAC.

Além dos experimentos de avaliação intrínseca, foi realizado um experimento com usuários reais, em que buscou-se medir a eficácia das sugestões na escrita de resumos acadêmicos. Primeiramente, os usuários – 8 estudantes de mestrado em Ciência da

Computação da Universidade Estadual de Maringá, os quais não tinham familiaridade com os conceitos dos componentes e das estruturas retóricas – realizaram as correções da estrutura retórica do resumo conforme as sugestões do SciPo. Depois dessa etapa, foram apresentados ao processo de refinamento com as sugestões de coerência. Após utilizarem o MAC, todos os usuários responderam um questionário com questões sobre suas impressões sobre o MAC, tais como facilidade de uso, relevância das sugestões apresentadas, alterações realizadas, entre outras. A maioria dos usuários concordaram que as sugestões do MAC são relevantes e ajudam a melhorar a coerência do texto, sendo que as informações apresentadas junto às sugestões resultaram em pouca ou nenhuma dúvida quanto em como melhorar seus resumos. Todos os resultados referentes a avaliação do MAC podem ser encontrados em Souza e Feltrim (2012).

Capítulo

4

Coleta e Anotação de Córpus

Para a investigação dos problemas relacionados à coerência textual fez-se necessária a compilação de um córpus específico para esse propósito. Como este trabalho buscou também a realização de experimentos semelhantes ao realizados por trabalhos relacionados, foi necessário coletar um córpus jornalístico em português. Para os experimentos com o córpus de resumos científicos, foram coletados resumos em português escritos por alunos de graduação como parte de seus trabalhos de conclusão de curso (TCCs). A descrição dos textos coletados é apresentada na Seção 4.1. Nas Seções 4.2 e 4.3 são descritas as anotações feitas nos dois corpora.

4.1 Coleta

4.1.1 Textos Jornalísticos

Com o objetivo de replicar o experimento (1) de ordenação de sentenças realizado por Barzilay e Lapata (2008) e por vários dos trabalhos relacionados, foi coletado um conjunto de textos jornalísticos. Os textos foram coletados de três corpora que abordam a tarefa de sumarização automática: (1) *CSTNews* (Cardoso et al., 2011), (2) *Summ-it* (Collovini et al., 2007) e (3) Temário (Rino e Pardo, 2007). Apesar de serem trabalhos que visam a automatização do sumário, é de interesse deste trabalho os textos originais que foram utilizados para a extração dos sumários, pois, intuitivamente, esses textos se caracterizam como coerentes, tornando-os aptos a serem utilizados para avaliar a implementação realizada neste trabalho. Um total de 286 textos jornalísticos foram coletados, sendo:

1. *CSTNews* (Cardoso et al., 2011): 136 textos provenientes dos seguintes jornais: a Folha de São Paulo, Estadão, O Globo, Gazeta do Povo e Jornal do Brasil[1].

2. *Summ-it* (Collovini et al., 2007): 50 textos do caderno de Ciências da Folha de São Paulo retirados do córpus PLN-BR (Recursos e Ferramentas para a Recuperação de Informação em Bases Textuais em Português do Brasil)[2].

3. Temário (Rino e Pardo, 2007): 100 textos da Folha de São Paulo (Seções: Especial, Mundo e Opinião) e do Jornal do Brasil (Seções: Internacional e Política)[3].

A Tabela 4.1 apresenta informações sobre o tamanho em sentenças dos textos coletados de cada córpus, sendo que os textos do Temário são, na média, os maiores.

Tabela 4.1: Variação do tamanho dos textos jornalísticos em número de sentenças.

Córpus	**Textos**	**Nº. mínimo**	**Nº. máximo**	**Média**
CSTNews	136	3	48	16,01
Summ-it	50	4	17	16,22
Temário	100	5	69	29,12

4.1.2 Textos Científicos

A coleta do córpus de resumos científicos foi feita a partir do córpus compilado para o trabalho de Souza e Feltrim (2012). Os resumos desse córpus foram extraídos de Trabalhos de Conclusão de Curso (TCCs) na área da Ciência da Computação e são provenientes de três fontes: Departamento de Informática da Universidade Estadual de Maringá (DIN-UEM), Departamento de Computação da Universidade Estadual de Londrina (DC-UEL) e Departamento de Informática do Instituto de Física e Matemática da Universidade Federal de Pelótas (Inf-UFPel). Os resumos provenientes do DIN-UEM foram coletados diretamente com seus autores ou respectivos orientadores, enquanto os outros resumos foram encontrados no Sistema de Arquivamento e Indexação de Documentos do DC-UEL[4] e no Acervo Digital da Biblioteca de Ciência e Tecnologia da UFPel[5]. No total, são 385 resumos que datam de 1999 a 2009, sendo 205 resumos provenientes do DIN-UEM, 98 do DC-UEL e 82 do Inf-UFPel.

A Tabela 4.2 apresenta a distribuição dos resumos entre as áreas de pesquisa identificadas, bem como os totais de palavras distribuídos pelas áreas e a quantidade média de palavras de cada resumo. Também é mostrado o total de palavras de todo o córpus (64.104 palavras) e a média geral da quantidade de palavras.

[1]Disponível em `http://www2.icmc.usp.br/~taspardo/`
[2]`http://www.nilc.icmc.usp.br/plnbr/index.htm`
[3]Disponível em `http://www.linguateca.pt/Repositorio/TeMario/`
[4]`http://www2.dc.uel.br/nourau/`
[5]`http://www.ufpel.tche.br/prg/sisbi/bibct/acervodigital.html`

Tabela 4.2: Distribuição e totais de palavras do córpus de resumos científicos por área de conhecimento (Souza e Feltrim, 2011).

Área de Pesquisa	**Nº. Resumos**	**Nº. Palavras**	**Média**
Banco de Dados	51	8.720	170,98
Inteligência Computacional	70	11.149	159,27
Engenharia de Software	92	15.482	168,28
Hipermídia	32	5.755	179,84
Sistemas Digitais	36	6.454	179,27
Sist. Distribuídos e Ling. Programação	19	4.192	220,63
Programação Gráfica e Proc. de Imagens	41	6.565	160,12
Redes de Computadores	44	5.787	131,52
Total	385	64.104	171,23

Todos os resumos desse córpus se encontram anotados de acordo com as quatro dimensões propostas por Souza e Feltrim (2012). São elas:

1. Dimensão Título: verifica o relacionamento semântico do componente Propósito com o Título do resumo;

2. Dimensão Propósito: verifica o relacionamento semântico dos componentes Metodologia, Resultado e Conclusão com o componente Propósito;

3. Dimensão Lacuna-Contexto: verifica o relacionamento semântico entre os componentes Lacuna e Contexto;

4. Dimensão Quebra de Linearidade: verifica se existe a quebra de sentido lógico entre sentenças adjacentes do resumo.

Como o objetivo deste trabalho está relacionado a dimensão Quebra de Linearidade, foram coletados todos os textos marcados com dimensão Quebra de Linearidade igual a "sim", reduzindo assim o córpus de resumos científicos para 101 resumos. Essa escolha se deu por se julgar que os textos marcados como "sim" seriam os mais prováveis de conter problemas de coerência.

Visando aumentar o número de textos coletados, também foram adicionados 40 resumos experimentais da turma do último ano do curso de Ciência da Computação do DIN-UEM. Como os resumos foram elaborados no primeiro semestre do respectivo ano (2012), de modo espontâneo e sem correções, também estariam propensos a problemas de coerência semelhantes aos detectados pela dimensão Quebra de Linearidade.

Dessa forma, um conjunto de 141 resumos foi coletado. Como havia dois resumos com número de sentenças igual a 1, eles foram removidos do córpus, uma vez que não seriam aptos à análise pelo modelo grade de entidades. Assim, o córpus de resumos científicos compilado para este trabalho ficou com 139 resumos. A Tabela 4.2 apresenta informações sobre o tamanho em sentenças dos 139 resumos do córpus.

Tabela 4.3: Variação do tamanho dos textos científicos em número de sentenças.

Córpus	Textos	Número mínimo	Número máximo	Média
Córpus científico	139	2	18	5,96

4.2 Anotações Preliminares

Para auxiliar a manipulação e compreensão das informações contidas no córpus, tanto por humanos quanto por ferramentas computacionais, é necessário que o córpus esteja eletronicamente anotado com um conjunto de etiquetas apropriadas. Para essa anotação, utilizou-se a linguagem XML (*Extensible Markup Language*), um padrão para anotação de documentos, aprovado pelo *World Wide Consortium* - W3C[6], que vem sendo mundialmente utilizado para a anotação de córpus.

Em um primeiro momento foi utilizado um conjunto de quatro etiquetas e três atributos para realizar a anotação. A Figura 4.1 apresenta um resumo com a anotação preliminar realizada automaticamente por *scripts* desenvolvidos neste trabalho, os quais removeram anotações não utilizadas neste trabalho do córpus de resumos científicos de Souza e Feltrim (2011) e adicionaram essas anotações XML no córpus jornalístico e nos resumos experimentais.

```
<RESUMO id="3">
<Titulo>Caracterização de Tráfego de Rede</Titulo>
<P id="0">
<S id="3-0">A caracterização de tráfego de rede tem como objetivos principais : garantia de
qualidade de serviço ( QoS ) , planejar e modelar o tráfego , analisar o desempenho de redes ,
analisar a perda de pacotes e estudar o comportamento do usuário .</S>
<S id="3-1">Neste trabalho , serão mostradas algumas técnicas de caracterização de tráfego
como : processos de renovação , processos de Markov , séries temporais e auto-similaridade .</S>
<S id="3-2">Além destes processos , podem ser utilizados os sketches .</S>
<S id="3-3">Estes possuem a característica de fazer a análise dinâmica e em tempo real de
fluxo de dados , o que é muito importante para detectar anomalias .</S>
</P>
</RESUMO>
```

Figura 4.1: Exemplo de resumo com a anotação preliminar.

Na Figura 4.1 é possível observar as etiquetas escritas em azul e os atributos escritos em vermelho. A primeira etiqueta é denotada por "<RESUMO> ... < /RESUMO>" e tem a função de estabelecer o início e fim do resumo, além do atributo "id" que é atribuído de acordo com o nome do arquivo em questão (neste caso, "3.xml") sem a representação de seu formato. A etiqueta "<Titulo> ... < /Titulo>" tem a função de estabelecer o início e o fim do texto referente ao título do resumo. A etiqueta "<P> ... < /P>" tem a função de estabelecer o início e o fim do texto de cada parágrafo do resumo e o atributo "id" se refere a posição do parágrafo, com a contagem iniciando em zero. Por último, a etiqueta "<S> ... < /S>" tem a função de estabelecer o início e fim do texto de cada sentença do

[6] http://www.w3.org/

resumo e possui o atributo "id" o qual possui o valor do "id" da etiqueta "<RESUMO>" seguido da posição da sentença no resumo, com a contagem iniciando em zero.

Essas etiquetas são preliminares no sentido de que se referem a aspectos gerais dos textos. Na seção a seguir, são descritas as marcações de coerência, tanto do córpus jornalístico quanto do córpus de resumos científicos.

4.3 Anotações de Coerência

Como a unidade de marcação de coerência utilizada neste trabalho foi o texto como um todo, um atributo "coerência" foi adicionado a etiqueta "<RESUMO>". Assim como em Burstein et al. (2010), dois valores possíveis foram definidos para esse atributo: "cp", significando "com problemas" de coerência, e "sp", significando "sem problemas" de coerência. Para cada um dos dois corpora (jornalístico e científico) foi realizado um método de anotação de coerência diferente, de acordo com o tipo de experimento pretendido.

O córpus jornalístico foi preparado para ser utilizado em experimentos de ordenação das sentenças semelhantes ao realizado por Barzilay e Lapata (2008). Para cada um dos textos originais foram geradas aproximadamente 20 versões sintéticas contendo permutações aleatórias da ordem das sentenças. Assim como Barzilay e Lapata (2008), 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 anotados como "sem problemas" de coerência e as versões permutadas como "com problemas" de coerência. Como cada texto original formará um par com uma de suas versões permutadas, o resultado foi um córpus com 5.720 pares (286 textos $\times$ 20 versões permutadas). O número de pares formados com textos jornalísticos discriminando o córpus de onde o texto original foi coletado é mostrado na Tabela 4.4. A escolha desse tipo de texto para essa forma de experimento se deve ao fato de se assumir que os textos jornalísticos são bem escritos e sua desordenação aproximaria ao encontro de problemas comuns à falta de coerência.

Tabela 4.4: Total de pares formados com os textos jornalísticos discriminando o córpus de onde o texto original foi coletado.

Córpus	**Textos Originais**	**Total de pares**
CSTNews	136	2720
Summ-it	50	1000
Temário	100	2000
Todos juntos	286	5720

O córpus de resumos científicos foi preparado para experimentos utilizando o julgamento humano acerca do nível de coerência dos resumos. A anotação manual dos

resumos foi realizada nos mesmos moldes do trabalho de Burstein et al. (2010). Desse modo, os anotadores foram instruídos a marcar o resumo como "com problemas" caso ele apresentasse uma quantidade substancial de barreiras na leitura (por exemplo, falta de continuidade anafórica) relativo ao tamanho do texto, demandando maior esforço cognitivo para sua interpretabilidade; caso contrário, os anotadores foram instruídos a marcar o resumo como "sem problemas".

Para medir a concordância entre anotadores, foi realizado um experimento em que dois anotadores treinados anotaram separadamente os 40 resumos experimentais. A concordância entre eles medida por meio da medida *Kappa* foi de 0,70%, valor considerado aceitável e próximo ao valor 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 dos 139 resumos resultou em 117 (84%) resumos marcados como "sem problemas" e 22 (16%) como "com problemas". Vale ressaltar que esse desbalanceamento é característico de corpora manualmente anotados. O mesmo nível de desbalanceamento foi observado nos corpora de redações (*essays*) utilizados por Burstein et al. (2010), conforme comentado na Subseção 2.2.5.

Capítulo

5

Implementação do Modelo Grade de Entidades para o Português

Conforme mencionado na introdução, o objetivo deste trabalho foi investigar a aplicabilidade do modelo grade de entidades na avaliação de coerência de resumos científicos escritos em português. Para tal, foi feita uma implementação do modelo segundo a proposta de Barzilay e Lapata (2008), usando ferramentas de PLN disponíveis para o português. Devido a indisponibilidade de uma ferramenta de resolução automática de correferência para o português, a etapa de identificação de entidades não pode ser implementada conforme o modelo original. 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 consideradas correferentes.

Para avaliar o modelo grade de entidades para o português, dois tipos de experimentos foram realizados, a saber:

1. experimentos de ordenação de sentenças usando o córpus de textos jornalísticos, no mesmo formato dos experimentos realizados por Barzilay e Lapata (2008);

2. experimentos baseados no julgamento de juízes humanos usando o córpus de resumos científicos, em formato semelhante aos experimentos realizados por Burstein et al. (2010).

O objetivo dos experimentos do tipo (1) foi validar a implementação do modelo feita para o português por meio da replicação de experimentos realizados para outras línguas (Barzilay e Lapata, 2008; Filippova e Strube, 2007; Yokono e Okumura, 2010). O objetivo dos experimentos do tipo (2) vem ao encontro dos objetivos deste trabalho e buscou avaliar

o desempenho do modelo grade de entidades no contexto de um classificador capaz de detectar problemas locais de coerência em resumos científicos, uma vez que a motivação para este estudo está na melhoria do módulo de análise de coerência do sistema SciPo.

Este capítulo apresenta a metodologia de trabalho empregada para a implementação do modelo grade de entidades para o português e também descreve os experimentos realizados e os resultados alcançados. Uma visão geral desse processo é mostrada na Figura 5.1, em que o texto com as marcações iniciais em XML é processado por um analisador sintático automático – neste caso, o *parser* PALAVRAS (Bick, 2002) – e com a saída gerada pelo *parser* são encontrados os sintagmas nominais (SNs) e as respectivas funções sintáticas (sujeito (S), objeto (O) e outro (X)). Em uma etapa de pós-processamento, todos os nomes (substantivos e nomes próprios) dos SNs simples e complexos são selecionados. Em seguida, os nomes são lematizados, visando a redução de entidades duplicadas, sendo assim possível a construção da grade de entidades. Por fim, de acordo com as configurações de `Sintático[+/-] Saliência[+/-]`, o vetor de características é extraído, gerando um arquivo de saída. O arquivo de saída é gerado em dois formatos possíveis, que correspondem aos formatos de entrada dos sistemas de aprendizagem de máquina utilizados, a saber: SVM^{rank} (Joachims, 2006) e WEKA – *Waikato Environment for Knowledge Analysis* – (Witten e Frank, 2005).

Na Seção 5.1 é descrito como a grade de entidades é montada e como os vetores de características são extraídos. Os experimentos e os resultados são detalhados na Seção 5.2.

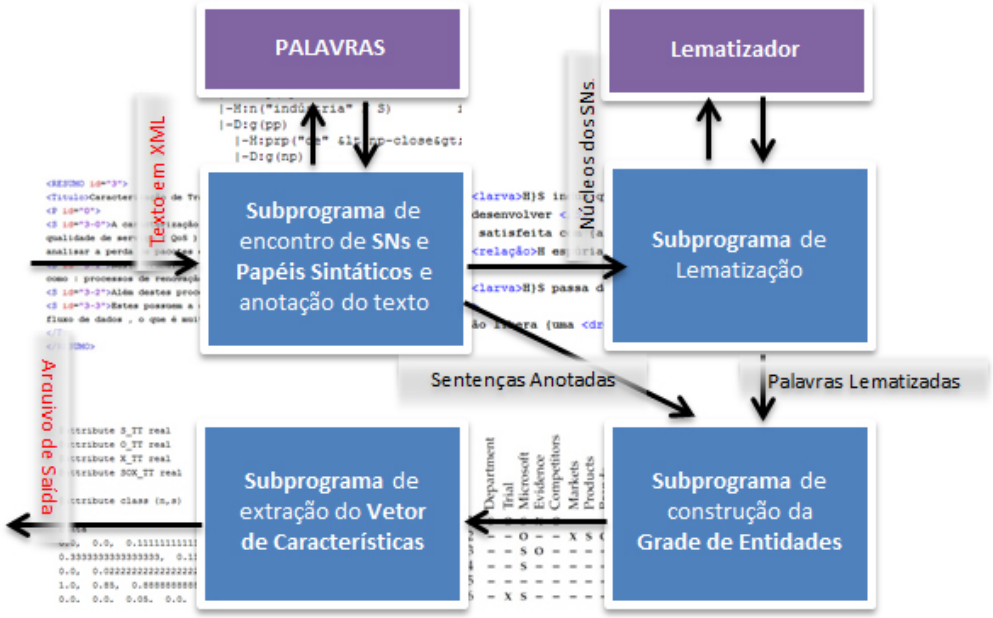


Figura 5.1: Etapas de processamento para a construção do modelo grade de entidades para o português.

5.1 Construção da Grade de Entidades e Extração dos Vetores de Características

5.1.1 Análise Gramatical Automática

Um dos passos iniciais necessários para a etapa de implementação deste trabalho é o uso de ferramentas que realizem análises gramaticais de um texto (*parsers*), visando a extração automática dos SNs e suas respectivas funções sintáticas (sujeito (S), objeto (O), outro (X)). Dado esse objetivo, foi estudada a viabilidade de utilização de três analisadores gramaticais para o português: (1) o Curupira (Martins et al., 2003), (2) o *DepParser* (Branco e Costa, 2010; Silva et al., 2010) e (3) o PALAVRAS (Bick, 2002).

Com relação ao (1) Curupira, o seu principal atrativo é a disponibilidade de uma DLL, o que permite extrair automaticamente sua análise. No entanto, sua saída não identifica diretamente os sujeitos e objetos da sentença, dificultando assim o seu uso no contexto deste trabalho. O (2) *DepParser*, disponível apenas para acesso *online* a partir do *website* da Universidade de Lisboa[1], também não se mostrou adequado devido ao formato em que a saída é disponibilizada. Apenas uma figura mostrando a estrutura da árvore sintática resultante é apresentada. Nesse mesmo *website*, um outro analisador sintático para o português, chamado *LXGram* (Branco e Costa, 2010), está disponível para *download*. Porém, não são disponibilizadas instruções para sua utilização, nem relatos acerca do seu desempenho, o que impossibilitou seu uso neste trabalho. Dado o contexto de aplicação, o *parser* (3) PALAVRAS se mostrou o mais adequado, uma vez que seus formatos de saída permitem a identificação direta dos SNs, bem como de suas funções sintáticas. Sendo assim, se decidiu pelo uso do PALAVRAS.

O PALAVRAS de Bick (2002) é um analisador sintático para o português desenvolvido na Universidade do Sul da Dinamarca[2] e que possui várias formas de visualização da análise gerada. Seu desempenho reflete o estado da arte da análise gramatical para o português e, como dito anteriormente, permite a extração de todos os dados necessários para este trabalho. Uma dificuldade para o uso do PALAVRAS no contexto deste trabalho é o fato do mesmo ter sua utilização via Acesso Remoto ou *download* liberada apenas perante pagamento por tempo de uso. Isso limitaria a implementação proposta, uma vez que seria necessário custeamento contínuo para acesso ao analisador. No entanto, foi possível construir um subprograma que realiza acesso por POST[3] direto ao analisador sintático de acesso gratuíto disponível *online* na página[4] da instituição do

[1] `http://lxcenter.di.fc.ul.pt/services/pt/LXServicesParserDepPT.html`
[2] `http://www.sdu.dk/`
[3] Referência: `http://www.w3.org/Protocols/rfc2616/rfc2616-sec9.html`
[4] `http://beta.visl.sdu.dk/visl/pt/parsing/automatic/parse.php`

autor, permitindo assim a sua utilização na anotação automática dos córpora utilizados neste trabalho.

Essa versão para acesso livre do PALAVRAS disponibiliza duas formas de visualização do resultado da análise: como uma (1) estrutura *flat* ou como uma (2) estrutura em árvore, sendo a forma (2) a utilizada neste trabalho. Um exemplo de saída na forma de estrutura *flat* é mostrado na Figura 5.2 e, de estrutura em árvore visualizada na vertical é mostrado na Figura 5.3. Essa estrutura também pode ser visualizada na horizontal ou em um formato alternativo chamado de *source*. A estrutura em árvore com visualização no formato *source* possui identações e caracteres especiais que facilitam o pós-processamento do resultado, conforme mostrado na próxima subseção.

```
o [o] <artd> DET M S @>N
cachorro [cachorro] <Azo> N M S @SUBJ>
comeu [comer] <vt> <fmc> V PS 3S IND VFIN @FMV
o [o] <artd> DET M S @>N
osso [osso] <mat> N M S @<ACC
```

Figura 5.2: Resultado da análise do PALAVRAS em estrutura *flat*.

```
|-S:g(np)
| |-D:pron(det "o" &lt;artd&gt; DET M S) o
| |-H:n("cachorro" M S) cachorro
|-P:v(fin "comer" &lt;fmc&gt; PS 3S IND VFIN) comeu
|-Od:g(np)
  |-D:pron(det "o" &lt;artd&gt; DET M S) o
  |-H:n("osso" M S) osso
```

Figura 5.3: Resultado da análise do PALAVRAS em estrutura de árvore.

5.1.2 Encontro dos Sintagmas Nominais

Sintagma e Classificação de Sintagmas

Sintagma é o conjunto de palavras que, juntas, possuem um significado na sentença. Dentro desse conjunto, uma palavra é o núcleo do sintagma e as demais mantém uma relação de dependência e ordem em torno desse núcleo. Para exemplificar, na sentença a seguir, os sintagmas são separados por colchetes:

[O cachorro][comeu o osso].

As palavras "o cachorro" constituem um sintagma que corresponde ao sujeito da oração; e as palavras "comeu o osso" constituem um outro sintagma que corresponde ao predicado. A classificação de um sintagma se dá pelo seu núcleo. Se o núcleo do sintagma é um

nome, o sintagma é dito nominal; se o núcleo do sintagma é um verbo, o um sintagma é dito verbal. Desse modo, na sentença acima, o sujeito "o cachorro" é um sintagma nominal (SN), uma vez que o núcleo "cachorro" é um nome, e o predicado "comeu o osso" é um sintagma verbal (SV), uma vez que o núcleo "comeu" é um verbo. Vale destacar que o SV pode ser dividido em sintagmas menores. No exemplo dado, a palavra "comeu" é um verbo transitivo direto e as palavras "o osso" constituem um SN que exerce a função de objeto direto. Isso mostra que a organização dos sintagmas em uma sentença é hierárquica, de modo que um sintagma pode conter um ou mais sintagmas menores, conforme exemplificado na Figura 5.4.

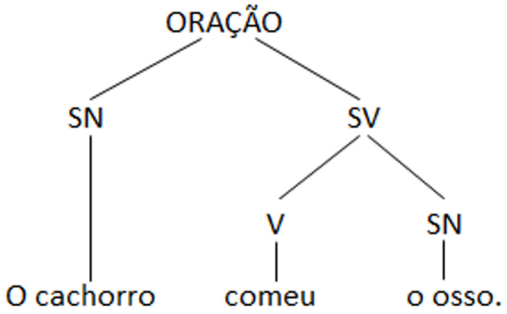


Figura 5.4: Exemplo da estrutura hierárquica dos sintagmas.

O núcleo do sintagma geralmente vem acompanhado de outras palavras que tem a função de especificar ou modificar o núcleo. Palavras que especificam o núcleo são chamadas de determinantes. Palavras que modificam o núcleo são chamadas de modificadores. Os modificadores podem ser uma única palavra ou outro sintagma, em especial, sintagmas preposicionais – sintagmas formados por uma preposição + um SN – funcionam como modificadores de SNs e SVs.

O Sintagma Nominal

Perini (2003) define o SN de maneira muito simples: "é o sintagma que pode ser sujeito de alguma oração". Perini (2010) diz que o SN é composto de uma ou mais palavras e apresenta certas propriedades, a saber:

1. o SN pode ocorrer nas funções sintáticas de sujeito, objeto ou complemento de preposição;

2. semanticamente, o SN pode se referir a uma entidade do mundo (real ou imaginário).

As sentenças abaixo mostram exemplos de SNs separados por colchetes:

[Minha irmã] trabalha [no banco].
[Aquele aluno] obteve [uma nota boa].
[Paulo] chegou.

É importante observar que um SN não precisa conter obrigatoriamente um substantivo como núcleo, já que palavras de outras classes gramaticais podem exercer a função de sujeito. Na sentença abaixo, a palavra "amar", que é um verbo, exerce a função de sujeito:

[Amar] é o verbo que eu gosto.

Conforme mencionado anteriormente, além do núcleo, outras palavras podem fazer parte do sintagma. No caso do SN, os modificadores normalmente são sintagmas preposicionais e adjetivos. Já os determinantes são artigos, pronomes ou numerais. O SN também pode ser modificado por orações subordinadas adjetivas, uma vez que elas exercem papel de adjetivo. Perini (2003) denomina "sintagma complexo" o SN modificado por uma oração subordinada. A sentença a seguir, tirada de dos Santos (2006), exemplifica um SN complexo. O SN está separado por colchetes e a oração subordinada está destacada em negrito:

[O rapaz **que subiu no ônibus**] é meu amigo.

Neste trabalho, a noção de SN complexo é mais abrangente e vem do fato de que SNs acompanhados de modificadores, como orações subordinadas ou sintagmas preposicionais, incluem de modo hierárquico e por vezes recursivo outros SNs. Por exemplo, na sentença anterior, o SN complexo "o rapaz que subiu no ônibus" pode ser desmembrado em dois SNs simples, conforme mostrado abaixo:

[o rapaz] que subiu em [o ônibus] é meu amigo.

A recursividade fica mais evidente no caso dos SNs modificados por sintagmas preposicionais, conforme exemplificado pela sentença abaixo:

[a primeira fábrica de [produção de [remédio de [o governo]]]]. *(SN recursivo)*

Desmembrando o SN complexo acima em SNs simples, o resultado da análise fica como mostrado abaixo:

[a primeira fábrica] de [produção] de [remédio] de [o governo]. *(SNs simples)*

A ideia de desmembramento dos SNs complexos em SNs simples vem do fato de estarmos interessados em detectar todas as possíveis entidades do texto. De fato, essa abordagem é semelhante a adotada por Elsner e Charniak (2011), que consideram como entidades todos os nomes constituintes do SN em vez de considerar como entidade apenas o núcleo.

Pós-processamento da saída do *parser* PALAVRAS

Conforme já mencionado, um dos passos iniciais necessários para a implementação do modelo grade de entidades é a extração automática dos SNs e seus respectivas funções sintáticas, mais especificamente sujeito (S) e objeto (O). Todos os outras funções sintáticas são classificados como outro (X). Essa é uma etapa crucial, uma vez que é a partir dos SNs que se identifica as entidades que comporão a grade. Vale destacar também que o maior interesse se encontra nos SNs que ocorrem nas funções de sujeito e objeto, uma vez que SNs em outras funções são sempre rotulados com o papel outro. As funções sintáticas referentes à sujeito e objeto que o PALAVRAS anota são mostradas na Tabela 5.1, segundo o conjunto de etiquetas VISL *default*[5].

Tabela 5.1: Anotações sintáticas referentes à Sujeito e Objeto do PALAVRAS.

Sigla	Significado em Inglês	Tradução
S	*Subject*	Sujeito
Od	*Direct (accusative) object*	Objeto direto (acusativo)
Oi	*Dative object*	objeto indireto pronominal
Op	*Prepositional object*	Objeto preposicional

O *parser* PALAVRAS identifica todos os SNs de uma sentença, incluindo aqueles que compõem SNs complexos. As funções sintáticas dos SNs também são identificados, bem como os seus núcleos. SNs complexos são apresentados de forma aninhada e hierárquica, o que demandou o pós-processamento da saída do *parser* para a delimitação dos SNs simples. Para exemplificar como a saída do *parser* foi pós-processada, a seguinte sentença será utilizada como exemplo:

"O Klerk teve uma atividade política atuante como ministro da educação."

A saída do *parser* para essa sentença no formato estrutura em árvore e visualização *source* é mostrada na Figura 5.5.

Conforme pode ser visto na Figura 5.5, cada palavra da sentença recebe um conjunto de etiquetas próprias do PALAVRAS. Entre elas, estão as marcações referentes aos sintagmas. Os SNs são anotados com a etiqueta *np*, enquanto os sintagmas preposicionais são anotados com a etiqueta *pp*. A etiqueta relativa à função sintática do sintagma seguida de dois pontos (:) precede as etiquetas que definem o tipo do sintagma. Dado que grande parte das etiquetas geradas pelo PALAVRAS não são de interesse para este trabalho, essa anotação foi convertida para uma anotação simplificada, em que somente os SNs, seus núcleos e os respectivas funções sintáticas são anotados. O processo de simplificação da anotação é exemplificado na sequência.

[5]http://beta.visl.sdu.dk/visl/pt/info/symbolset-manual.html

```
1  S:g(np)
2  =D:pron(det "o" <artd> DET M S)        o
3  =H:prop("Klerk" M S)        Klerk
4  P:v(fin "ter" <fmc> PS 3S IND VFIN)        teve
5  Od:g(np)
6  =D:pron(det "um" <arti> DET F S)        uma
7  =H:n("atividade" F S)        atividade
8  =D:adj("político" F S)        política
9  =D:adj("atuante" F S)        atuante
10 Co:g(pp)
11 =H:prp("como" <right>)        como
12 =D:g(np)
13 ==H:n("ministro" M S)        ministro
14 ==D:g(pp)
15 ===H:prp("de" <sam-> <np-close>)        de
16 ===D:g(np)
17 ====D:pron(det "o" <artd> <-sam> DET F S)        a
18 ====H:n("educação" F S)        educação
```

Figura 5.5: Resultado da análise do *parser* PALAVRAS no formato estrutura de árvore e visualização *source*.

No exemplo da Figura 5.5, o primeiro SN é mostrado na linha 1. A função sintática desse SN é sujeito (S) e seus constituintes – o determinante "o" e o nome próprio (*prop*) "Klerk" – são mostrados nas linhas 2 e 3, respectivamente. Note que o nome "Klerk" foi anotado com a etiqueta H (de *head*), o que o identifica como núcleo do SN. Na linha 5 se encontra o SN "uma atividade política atuante". Sua função sintática é objeto direto (Od) e seu núcleo é o substantivo (n) "atividade". Em seguida, na linha 10, se encontra um sintagma preposicional (SP) que funciona como complemento do objeto (Co). Uma vez que o sintagma preposicional pode conter SNs, ele também é analisado. O núcleo desse SP é a preposição "como" que é seguida pelo SN complexo "ministro da educação". A função sintática do SN complexo é outro (X), já que ele não exerce função nem de sujeito, nem de objeto. Assim, o resultado de uma primeira análise é a seguinte:

{o Klerk}S teve {uma atividade política atuante}Od como {ministro da educação}X

Conforme explicado anteriormente, os SNs complexos são aqueles que são constituídos de sintagmas menores. Dessa forma, sempre que um SN aparecer dentro de outro, caracterizando um SN complexo, ambos serão discriminados como SNs simples e os novos SNs simples herdarão a função sintática do SN complexo que os originou. Esse é o caso do SN "ministro da educação", como pode ser visualizado nas linhas 12 a 18 da Figura 5.5. Seguindo essa abordagem, esse SN complexo é desmembrado em dois SN simples – o SN "ministro" (identificado na linha 12) e o SN "a educação" (identificado na linha 16). A função sintática desses dois SNs simples será outro (X), já que esse é a função sintática do SN complexo, conforme mostrado a seguir:

{o Klerk}S teve {uma atividade política atuante}Od como {ministro}X de {a educação}X

Por fim, a anotação referente aos núcleos dos SNs, já marcados pelo PALAVRAS como H, é refinada visando facilitar o processo de extração de entidades (descrito na próxima subseção). Assim como Elsner e Charniak (2011), neste trabalho se optou por selecionar como entidades todos os substantivos (n) e nomes próprios ($prop$), mesmo quando eles não aparecem como o núcleo do SN. Uma vez que o processo de extração de entidades implementado se baseia nas etiquetas de núcleo (H), esses nomes que não são núcleo de SNs também foram marcados como tal. Dessa forma, a anotação final do exemplo fica como mostrado abaixo:

{o <Klerk>H}S teve {uma <atividade>H política atuante}O como {<ministro>H}X de {a <educação>H}X

5.1.3 Selecionando Entidades para a Grade de Entidades

Conforme descrito no Capítulo 2, a grade de entidades de um texto é uma matriz $G_{i \times j}$, em que cada coluna j representa uma entidade do texto e cada linha i representa uma sentença, sendo que a primeira linha da matriz corresponde a primeira sentença do texto, a segunda linha corresponde a segunda sentença, e assim por diante. Cada célula $c_{i,j}$ dessa matriz é preenchida com a função sintática da entidade j na sentença i. Caso a entidade j não seja mencionada na sentença i, a célula $c_{i,j}$ é preenchida com vazio (-).

Conforme mencionado anteriormente, a seleção de entidades para a composição da grade não pôde agrupar entidades correferentes da mesma forma em que foi proposto no modelo original, devido a indisponibilidade de uma ferramenta de resolução automática de correferência para o português. Assim, neste trabalho, a identificação de entidades seguiu uma abordagem similar a de Elsner e Charniak (2011), em que os SNs com o mesmo núcleo são considerados correferentes.

O processo de seleção de entidades para compor a grade se dá da seguinte maneira: para cada sentença do texto são extraídos os núcleos dos SNs. Para cada núcleo, se verifica se ele já foi incluído na grade de entidades por meio da comparação de palavras, de forma que apenas uma ocorrência de cada entidade é incluída na grade. Para evitar que a mesma palavra gere entidades diferentes quando variar em gênero e/ou número no mesmo texto (por exemplo, "projeto" e "projetos"), foi utilizado o lematizador desenvolvido pelo NILC[6] para converter todas as palavras as suas formas canônicas (lemas) antes de se fazer a comparação. No caso de núcleos compostos, como "linguagem de programação", se optou por verificar se um dos componentes do núcleo candidato já consta na grade, por exemplo, "linguagem". Nesse caso, "linguagem de programação" e "linguagem" serão considerados correferentes e representados na grade como uma única entidade. Essa abordagem foi adotada como uma tentativa de minimizar o número de entidades que, embora correferentes, são tratadas como entidades separadas.

[6]http://www.nilc.icmc.usp.br/nilc/

5.1.4 Extraindo os Vetores de Características a partir da Grade de Entidades

Uma vez que a grade de entidades foi construída, o vetor de características pode ser extraído de acordo com uma das configurações possíveis. No modelo original de Barzilay e Lapata (2008), as configurações possíveis são representadas por `Correferência[+/-] Sintático[+/-] Saliência[+/-]`, representando a consideração (+) ou não (-) de tal aspecto no cálculo 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. Assim, as configurações possíveis neste trabalho são `Correferência-Sintático[+/-] Saliência[+/-]`.

Quando a configuração é `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. Dessa forma, sob essa configuração e considerando apenas transições de tamanho dois, o vetor de características contém as probabilidades de 16 transições possíveis. São elas: SS, SO, SX, S-, OS, OO, OX, O-, XS, XO, XX, X-, -S, -O, -X, –. Vale ressaltar que o aumento no tamanho das transições aumenta bastante o tamanho do vetor de características, o que pode levar a necessidade uma quantidade muito grande de dados de treinamento, dificultando, assim, a etapa de aprendizagem. Quando a configuração é `Sintático-`, as funções sintáticas são desconsiderados e o vetor de características contém as probabilidades das transições possíveis considerando-se apenas a presença (X) ou ausência (-) das entidades nas sentenças. As transições de tamanho dois possíveis nessa configuração são as seguintes: XX, X-, -X, –.

Com relação à saliência, Barzilay e Lapata (2008) explicam que várias classes ou "níveis" de saliência podem ser utilizados, por exemplo, uma classe para entidades mencionadas duas vezes, outra para entidades mencionadas três vezes, e assim por diante. No entanto, como o tamanho do vetor de características também aumenta conforme aumenta o número de classes de saliência, o comum é usar apenas duas classes: entidades salientes e não salientes. De fato, para n transições e k classes de saliência, o tamanho do vetor de características será de $n \times k$. Assim como Barzilay e Lapata (2008), neste trabalho foram consideradas duas classes de saliência, sendo que entidades mencionadas duas ou mais vezes são classificadas como salientes. Dessa forma, quando a configuração é `Saliência+`, a grade de entidade é dividida em duas – uma grade composta apenas de entidades salientes e uma composta de entidades não salientes. As probabilidades das transições são computadas separadamente para cada grade e todas as probabilidades são incluídas no vetor de características que representa o texto em questão. O número n

de transições depende da configuração utilizada para o aspecto sintático – $n = 16$ para `Sintático+` e $n = 4$ para `Sintático-` – e o tamanho do vetor de características é o dobro do tamanho do vetor calculado na configuração `Saliência-`. Quando a configuração é `Saliência-`, nenhuma discriminação de saliência é feita e uma única grade de entidades é usada para o cálculo das probabilidades.

Além dos atributos previstos no modelo original de Barzilay e Lapata (2008), neste trabalho também foi implementada a extração de atributos do tipo *Type/Token* (TT) no mesmo formato utilizado por Burstein et al. (2010) (conforme explicado na Seção 2.2.5, Cap. 2). A escolha pela implementação desses atributos se deu pelo fato do trabalho de Burstein et al. (2010) ter um objetivo próximo ao deste trabalho – avaliar a coerência de redações escritas por estudantes de populações variadas – e também ao fato de que a extração desses atributos não demanda a utilização de ferramentas adicionais. Os outros atributos propostos por Burstein et al. (2010) são calculados por meio de ferramentas de AES proprietárias e desenvolvidas para o inglês ou dependem de conhecimento linguístico não disponível para o português, como é caso da lista de *shell nouns* de Aktas e Cortes (2008).

Os atributos (*_TT) são usados para 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 de todas as funções sintáticas (SOX_TT). Nesse caso, 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 todos os papéis. Quando a configuração é `Sintático-`, apenas um atributo TT é calculado (P_TT) que representa o número de entidades diferentes na grade de entidades dividido pelo número de ocorrências dessas entidades nas sentenças.

5.2 Experimentos e Resultados

Para avaliar o comportamento do modelo grade de entidades para português implementado neste trabalho foram utilizados os corpora jornalístico e científico descritos no Capítulo 4. Conforme já mencionado, dois tipos de experimentos comumente utilizados em trabalhos relacionados foram realizados, a saber: (1) experimentos de ordenação de sentenças usando o córpus jornalístico e (2) experimentos de classificação baseados no julgamento de juízes humanos usando o córpus de resumos científicos.

Os experimentos do tipo (1) buscaram replicar os experimentos de avaliação do modelo grade de entidades 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 a língua portuguesa é semelhante ao comportamento observado para outras línguas. Assim, tanto a criação do córpus jornalístico – composto por pares do tipo {texto original, versão criada artificialmente por permutação aleatória da ordem das sentenças} – quanto o método de aprendizagem – formulação da etapa de aprendizagem como um problema de ranqueamento – seguiu especificações semelhantes as relatadas nos trabalhos relacionados citados acima. Para a etapa de aprendizagem deste trabalho foi utilizado o sistema SVM^{rank} (Joachims, 2006), que implementa o algoritmo SVM (*Support Vector Machine*) para problemas de ranqueamento. O SVM^{rank} é uma melhoria do SVM^{light} (Joachims, 1999) utilizado pela maioria dos trabalhos relacionados.

Os experimentos do tipo (2) buscaram avaliar o desempenho do modelo grade de entidades para a detecção de problemas locais de coerência em resumos científicos escritos em português. Nesse caso, a etapa de aprendizagem foi modelada como um problema de classificação, similar abordagem de Burstein et al. (2010), em que se busca distinguir entre duas classes – textos "com problemas" e textos "sem problemas" de coerência –, sendo que a anotação de coerência foi realizada por juízes humanos conforme descrito na Seção 4.3, Cap. 4. Foram realizados experimentos com três algoritmos de AM conhecidos e disponíveis no ambiente WEKA (Witten e Frank, 2005), a saber: (1) SMO (*Sequential Minimal Optimization*) (Platt, 1998), uma implementação do algoritmo SVM para classificação; (2) J48, uma implementação em Java e de código aberto do algoritmo C4.5 (Quinlan, 1993) que gera árvores de decisão; e (3) *Naïve Bayes*, um algoritmo probabilístico baseado na regra da probabilidade condicional de Bayes.

Os experimentos dos tipos (1) e (2) e os resultados obtidos são descritos a seguir, nas Subseções 5.2.1 e 5.2.2.

5.2.1 Experimento 1: Ordenação de Sentenças

Assim como em Barzilay e Lapata (2008), os resultados obtidos por um modelo baseado em LSA foram utilizados como *baseline*. Tanto o modelo LSA quanto modelo grade de entidades são locais, isto é, são verificadas as transições sentença-a-sentença sem uma análise posterior da estrutura global do texto. Para o cálculo da LSA foi utilizada a implementação de Souza e Feltrim (2012), detalhada na Subseção 2.2.2. Os dois corpora compilados neste trabalho foram utilizados para a criação do espaço semântico e a SVD foi utilizada para redução desse espaço para 200 dimensões, conforme sugerido pelos autores da implementação.

Desse modo, para estimar a coerência de um texto (T), foi utilizada a equação abaixo, que calcula a média dos valores de similaridade ($sim(S_i, S_{i+1})$) calculados por meio de LSA entre duas sentenças adjacentes S_i e S_{i+1}, sendo n o número de sentenças do texto:

$$coher(T) = \frac{\sum_{i=1}^{n-1} sim(S_i, S_{i+1})}{n-1}$$

No caso deste experimento, que avalia a ordem mais coerente das sentenças de um texto, o cálculo do percentual de acertos do modelo LSA pode ser trivialmente inferido pela comparação dos valores de coerência (*coher*) atribuídos aos textos originais com os valores atribuídos as versões com as sentenças desordenadas. O empate de valores foi resolvido aleatoriamente.

Conforme comentado na Subseção 2.2.5, a tarefa de ordenação de sentenças é essencial para uma série de problemas de PLN. A tarefa consiste em determinar a melhor sequência de sentenças que satisfaça determinado conjunto de informações que se deseja avaliar no texto, sendo um fator essencial nos métodos de geração automática de texto, sumarização automática e outros problemas de síntese textual.

Na ordenação de sentenças, o texto é visto como um conjunto de sentenças e o algoritmo de ordenação tenta encontrar uma ordem que maximize a coerência do texto de acordo com alguns critérios, por exemplo, a probabilidade de uma ordem. Assim como em Barzilay e Lapata (2008), neste trabalho essa tarefa foi simplificada, de modo que o texto original foi comparado com suas versões com as sentenças desordenadas. Em vez de tentar encontrar a melhor ordem, o algoritmo deve escolher, entre as ordens apresentadas, aquela que julgar mais coerente.

Dessa forma, para cada um dos textos do córpus jornalístico foram criadas 20 versões diferentes com as sentenças desordenadas aleatoriamente, formando 20 pares do tipo {texto original, versão permutada} para cada texto, conforme mostrado da Tabela 4.4. Esse conjunto foi separado aleatoriamente em conjunto de treinamento e conjunto de teste, sendo $\frac{2}{3}$ para treinamento e $\frac{1}{3}$ para testes.

Conforme comentado no início da seção, para a realização deste experimento foi utilizado o sistema SVM^{rank} (Joachims, 2006), o qual possui aprimoramentos que melhoraram a velocidade em relação ao SVM^{light} (Joachims, 1999). O SVM^{rank} exige o uso de um parâmetro obrigatório C, que configura o balanceamento entre erro de treinamento e margem. Neste trabalho, foi utilizado o valor $C = 20$, sendo esse um valor que resultou em valores de ranqueamento com valores decimais maiores, permitindo uma melhor comparação dos resultados, conforme exemplificado na página do desenvolvedor[7].

A métrica de avaliação utilizada neste experimento segue a de Barzilay e Lapata (2008), em que dadas todas as comparações entre pares {texto original, versão permutada}, 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 5.2 é mostrado o percentual de acertos da *baseline* (LSA) e do modelo grade de entidades com atributos *Type/Token*, representado na tabela por suas quatro configurações possíveis (`Sintático[+/-] Saliência[+/-]`). Como os textos jornalísticos são provenientes de três corpora diferentes (CSTNews, Summit e Temário), os resultados são mostrados considerando-se o córpus de origem dos textos originais, além dos resultados calculados

[7] `http://www.cs.cornell.edu/People/tj/svm_light/svm_rank.html`

para o córpus jornalístico como um todo (coluna "Todos juntos"). Os melhores resultados obtidos por cada modelo estão destacados em negrito.

Tabela 5.2: Percentual de acertos da *baseline* (LSA) e do modelo grade de entidades.

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%

Conforme pode ser observado na Tabela 5.2, o modelo grade de entidades superou a *baseline* em todos os casos, com exceção do textos provenientes do córpus Temário, em que a *baseline* superou o desempenho da melhor configuração do modelo grade de entidades em 4%. De fato, os textos do córpus Temário são bem maiores do que os textos provenientes dos outros dois córpora, conforme mostrado na Tabela 4.1, e isso pode ter influenciado o resultado da *baseline*, que é baseada na média de similaridade entre pares de sentenças. No geral, os resultados obtidos pelo modelo grade de entidades para o português são semelhantes aos relatados pelos trabalhos desenvolvidos para outras línguas (Elsner e Charniak, 2011; Filippova e Strube, 2007; Yokono e Okumura, 2010), ficando abaixo apenas dos resultados relatados pelos autores do modelo (Barzilay e Lapata, 2008). A Tabela 5.3 apresenta um resumo dos melhores resultados relatados pelos trabalhos relacionados considerando-se sempre na configuração `Correferência-`.

Tabela 5.3: Resumo dos resultados apresentados para o modelo grade de entidades por trabalhos relacionados.

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

Ainda na Tabela 5.2, observa-se que as configurações `Sintático[+/-]` e `Saliência[+/-]` não resultaram em um padrão que desse condições a um julgamento no sentido de decidir qual a melhor configuração a ser utilizada, variando conforme o córpus utilizado. Esse mesmo comportamento é notado nos trabalhos relacionados e no modelo original.

5.2.2 Experimento 2: Julgamento Humano

O segundo experimento verificou a habilidade do modelo grade de entidades de avaliar a coerência de resumos científicos, distinguindo entre resumos classificados como "com problemas" ou "sem problemas" de coerência. É fato que os dados sintéticos do primeiro experimento aproximam apenas parcialmente os problemas de coerência que leitores humanos encontrariam em um texto. Exemplos de problemas que podem ser encontrados são quebras nas cadeias de correferência ou sentenças que não se relacionam com a anterior por mudança de tópico.

Conforme descrito na Seção 4.3, o córpus de 139 resumos científicos foi manualmente anotado por dois juízes humanos e a concordância entre eles medida em um subconjunto de 40 resumos foi de $K = 0,70$. Dada a natureza do córpus e do objetivo desse tipo de experimento, a etapa de aprendizagem foi modelada como um problema de classificação binária, seguindo a mesma abordagem adotada por Burstein et al. (2010). O córpus é bastante desbalanceado, sendo que a classe majoritária – "sem problemas" – corresponde a 84% do córpus (117 resumos) e a classe minoritária – "com problemas" – corresponde a 16% (22 resumos).

Os experimentos foram realizados no ambiente WEKA com os algoritmos SMO, J48 e *Naïve Bayes*. A escolha pela 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 aprendizagem utilizado por Burstein et al. (2010); e o *Naïve Bayes* foi escolhido por ser um algoritmo de aprendizagem simples, rápido e de larga utilização em tarefas que envolvem classificação de texto. Os resultados foram calculados aplicando-se 10-*fold cross-validation* ao corpus de 139 resumos. Como métrica de avaliação foram adotadas medidas comumente utilizadas na avaliação de classificadores, a saber: acurácia, *Kappa*, *Precison*, *Recall* e *F-measure*. Os resultados em termos das medidas *F-measure* e *Kappa* são mostrados para cada algoritmo de aprendizagem e configuração do modelo grade de entidades na Tabela 5.4. Os valores de *F-measure* apresentados 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* descritos na Subseção 5.1.4. Os resultados listados como *_TT- foram calculados utilizando apenas o modelo grade de entidades.

Tabela 5.4: Resultados do modelo grade de entidades para o córpus de resumos científicos em termos de *F-measure* e *Kappa*.

	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 5.4, os melhores resultados foram obtidos com o algoritmo J48, sendo que o melhor resultado ($K = 0,65$) se aproxima do valor obtido por juízes humanos ($K = 0,70$) e ultrapassa o melhor sistema de Burstein et al. (2010) ($K = 0,61$). 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, 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 classificação e por classe, expressos nas cinco medidas de avaliação utilizadas, são apresentados no Apêndice A.

Com base nos resultados do algoritmo J48 e analisando as diferentes configurações do modelo grade de entidades, fica evidente a contribuição do aspecto saliência no desempenho do modelo. 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 do córpus de treinamento. Uma vez que a configuração `Sintático+` gera um vetor de características 4 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.

Conforme pode ser observado na Tabela 5.4, os atributos *Type/Token* tiveram pouca influência nos resultados, sendo que, em alguns casos, os valores com *_TT+ permanece-

ram iguais aos valores com *_TT-. Uma discreta contribuição dos atributos *_TT pode ser notada nos resultados calculados com o algoritmo *Naïve Bayes*.

Para avaliar o efeito do desbalanceamento do córpus nos resultados, os experimentos com os três algoritmos de aprendizagem 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, novos casos da classe minoritária gerados sinteticamente a partir de casos já existentes são adicionados ao conjunto. Esses novos casos são gerados na vizinhança de cada caso da classe minoritária. 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 5.5 apresenta os resultados após a classe minoritária ter sido aumentada em 400%, valor que deixa o córpus com um balanceamento próximo a perfeito. Assim como na Tabela 5.4, os resultados são mostrados em termos das medidas *F-measure* e *Kappa* para cada algoritmo de aprendizagem e configuração do modelo grade de entidades. Os resultados detalhados por algoritmo de classificação e por classe, expressos nas cinco medidas de avaliação utilizadas, são apresentados no Apêndice A.

Tabela 5.5: Resultados do modelo grade de entidades para o córpus de resumos científicos balanceado com SMOTE em termos de *F-measure* e *Kappa*.

	Naïve Bayes		**SMO**		**J48**	
***_TT-**	**F-meas.**	**Kappa**	**F-meas.**	**Kappa**	**F-meas.**	**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

Conforme pode ser observado na Tabela 5.5, os resultados usando *oversampling* com SMOTE foram melhores para os três algoritmos testados, sendo que o J48 continuou apresentando o melhor resultado, especialmente em termos da 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 5.4), 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 5.6.

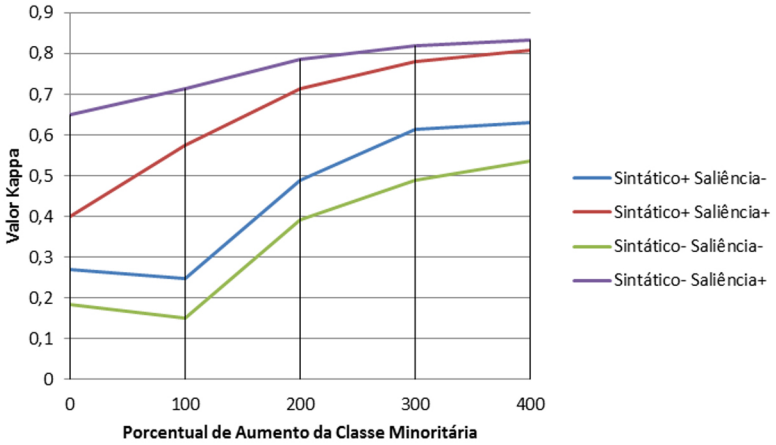


Figura 5.6: Variação dos valores da medida *Kappa* de acordo com o percentual de *oversampling* (SMOTE).

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 5.6 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* com SMOTE são mostrados na segunda parte.

Tabela 5.6: Medidas do J48.

Sem oversampling						
	Sem Problema (117)			Com problema (22)		
*_TT-	Precision	Recall	F-measure	Precision	Recall	F-measure
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-						
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

Com o objetivo de observar a influência do tamanho do córpus nos resultados foi realizado um experimento em que aumentando-se 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 aumento de tamanho, 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 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 5.7.

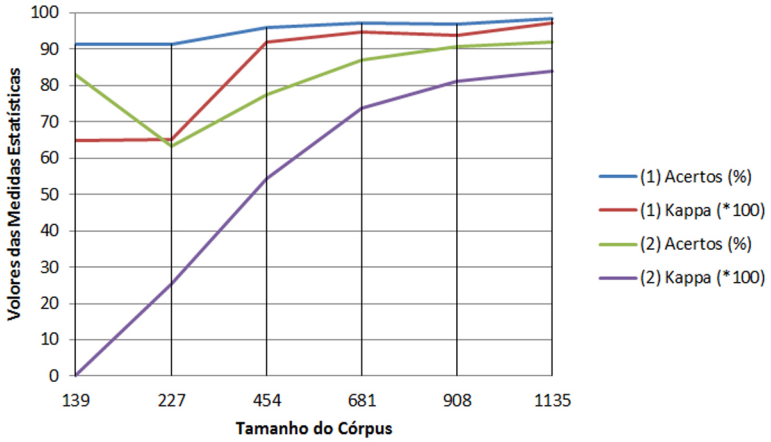


Figura 5.7: Variação dos valores de acurácia e medida *Kappa* de acordo com o aumento artificial do tamanho do córpus.

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 5.7, 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 classes – o desempenho do modelo grade de entidades se estabiliza. Demonstra também a configuração utilizada no cenário (1) é a melhor configuração para esse córpus, independente do seu tamanho.

Capítulo

6

Conclusões

Este trabalho teve como objetivo implementar e avaliar o modelo grade de entidades proposto por Barzilay e Lapata (2008) para a língua portuguesa, visando sua aplicação na avaliação de coerência em resumos científicos. A motivação para este trabalho 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 (2012), visando melhorar os resultados obtidos no âmbito da detecção de quebras de linearidades entre sentenças adjacentes de um resumo.

Para a implementação do modelo foi desenvolvido um subprograma que recupera a análise realizada pelo *parser* PALAVRAS (Bick, 2002) *online*, anotando no texto as identificações de sintagmas nominais (SNs), as funções sintáticas desses SNs e os respectivos núcleos. Processamento adicional foi necessário para desmembrar os SNs complexos identificados pelo *parser* em SNs simples, a fim de se selecionar também os substantivos que estavam nos SNs complexos, semelhante a abordagem adotada por Elsner e Charniak (2011). A partir desses dados, as entidades puderam ser extraídas para a construção da grade de entidades.

Para a seleção de entidades, como não havia disponível para o português um sistema de resolução automática de correferência para transformar cada grupo de entidades correferentes em uma única entidade, apenas palavras com o mesmo lema foram consideradas correferentes e tratadas como uma mesma entidade. Também foi feito um tratamento das palavras compostas, visando eliminar entidades duplicadas.

Além dos atributos extraídos a partir da grade de entidades previstos no modelo original, neste trabalho também foram incluídos atributos relacionados à variabilidade de vocabulário, mais especificamente, foi utilizado o modelo *Type/Token* para extrair atributos adicionais de forma similar ao relatado por Burstein et al. (2010).

Para realização dos experimentos, primeiramente foi utilizado um córpus jornalístico para a tarefa de ordenação de sentenças, buscando reproduzir os mesmo cenário de testes empregado pelos trabalhos relacionados encontrados na literatura. Isso permitiu a comparação dos resultados deste trabalho com outras implementações para outras línguas, mostrando que os resultados são próximos aos relatados para as línguas inglesa e superiores ao encontrado para a língua japonesa e alemã. O modelo de entidades também superou o modelo baseado em LSA nos testes realizados com o mesmo córpus, validando a utilização do modelo para os fins propostos neste trabalho.

Os experimentos com córpus de resumos científicos modelado como um problema de classificação mostraram que o uso do modelo de grade de entidades no contexto de um classificador para a dimensão quebra de linearidade é viável. O melhor resultado ($K = 0,65$) foi alcançado com o algoritmo J48, implementação do C4.5 disponível no ambiente WEKA, com o modelo grade de entidades na configuração `Sintático-Saliência+`, sendo esse resultado bem próximo do obtido por dois juízes humanos ($K = 0,70$) em um experimento de anotação realizado com um subconjunto de 40 textos extraídos do córpus de resumos científicos. Esse resultado também é superior ao relatado por Burstein et al. (2010) para seu melhor sistema ($K = 0,61$), que além de contar com a resolução automática de correferência, utiliza atributos adicionais extraídos por ferramentas de AES.

Para continuidade da pesquisa, se vê a necessidade de compilação e anotação manual de um córpus maior e mais balanceado para a realização dos testes de forma a obter resultados isentos das possíveis influências que a técnica de *oversampling* possa exercer. Também se faz necessária a busca por ferramentas para a resolução de correferência para o português, principalmente a pronominal, sendo que os pronomes não foram levados em consideração para seleção de entidades neste trabalho devido a indisponibilidade de ferramentas de apoio. Além disso, 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. Observa-se também a importância de um estudo que vise encontrar os *ShellNouns* para a língua portuguesa, uma vez que esse estudo pode ser útil não só no contexto deste trabalho, mas também em outras pesquisas na área de PLN. 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, por exemplo, os índices calculados pela Coh-Metrix-Port referentes a aspectos de coerência. Por fim acredita-se que este estudo possa beneficiar outras ferramentas que realizam o processamento de textos em português, como os sumarizadores automáticos, sendo que a utilização do modelo em outros contextos de aplicação constituem uma das fontes de trabalhos futuros.

Com isso este trabalho eleva os estudos de avaliação de coerência para o português ao estado da arte dos estudos em outras línguas. Além disso, este trabalho dá continuidade aos estudos da coerência textual em textos científicos realizados no Departamento de Informática da Universidade Estadual de Maringá. De maneira geral, espera-se que este trabalho contribua com o desenvolvimento científico na área de PLN, principalmente em relação às ferramentas de auxílio à escrita voltadas para a língua portuguesa, no qual existe uma carência significativa de desenvolvimento.

Referências

Abraços, J.; Lopes, J. G. Extending drt with a focusing mechanism for pronominal anaphora and ellipsis resolution. In: *Proceedings of the 15th conference on Computational linguistics - Volume 2*, COLING '94, Stroudsburg, PA, USA: Association for Computational Linguistics, 1994, p. 1128 – 1132 (*COLING '94*,).

Aktas, R. N.; Cortes, V. Shell nouns as cohesive devices in published and esl student writing. Journal of English for Academic Purposes, 2008, p. 3—-14.

Aluisio, S. M.; Oliveira, J. O. N. A detailed schematic structure of research paper introductions: An application in support-writing tools. In: *Proceedings of XII Congress of SEPLN*, 1996.

Aluísio, S. M.; Barcelos, I.; Sampaio, J.; Oliveira, J. O. N. How to learn the many unwritten "rules of the game"of the academic discourse: A hybrid approach based on critiques and cases to support scientific writing. In: *Proceedings of the IEEE International Conference on Advanced Learning Technologies*, ICALT '01, Washington, DC, USA: IEEE Computer Society, 2001, p. 257–260 (*ICALT '01*,).

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

Beaugrande, R.; Dressler, W. U. *Introduction to textlinguistics*. Londres/New York: Longman, 1981.

Bick, E. *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, 2002.

Branco, A.; Costa, F. A deep linguistic processing grammar for portuguese. In: *Pardo et al. (eds.), Computational Processing of Portuguese*, LNAI 6001, Springer, 2010, p. 86–89 (*LNAI 6001*,).

Burstein, J.; Chodorow, M.; Leacock, C. 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*, 2003, p. 3–10.

Burstein, J.; Tetreault, J.; Andreyev, S. 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, Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, p. 681–684 (*HLT '10*,).

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

Charolles, M. *Introdução aos problemas da coerência dos textos: abordagem teórica e estudo das práticas pedagógicas.* Campinas: Pontes, 39–90 p., 1978.

Chaves, A. R.; Rino, L. H. The mitkov algorithm for anaphora resolution in portuguese. In: *Proceedings of the 8th international conference on Computational Processing of the Portuguese Language*, PROPOR '08, Berlin, Heidelberg: Springer-Verlag, 2008, p. 51–60 (*PROPOR '08*,).

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; Kegelmeyer, W. P. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 2002.

Collovini, S.; Carbonel, T. I.; Fuchs, J. T.; Coelho, J. C.; Rino, L.; Vieira, R. Summ-it: um corpus anotado com informações discursivas visando sumarização automática. In: *TIL 2007*, 2007.

Cuevas, R. R.; Paraboni, I. A machine learning approach to portuguese pronoun resolution. In: *Proceedings of the 11th Ibero-American conference on AI: Advances in Artificial Intelligence*, IBERAMIA '08, Berlin, Heidelberg: Springer-Verlag, 2008, p. 262–271 (*IBERAMIA '08*,).

Deerwester, S.; Dumais, S. T.; Furnas, G. W.; Landauer, T. K.; Harshman, R. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, v. 41, n. 6, p. 391–407, 1990.

van Dijk, T. A. Studies in the pragmatics of discourse. In: *American Anthropologist*, 1983, p. 190–192.

van Dijk, T. A.; Kintsch, W. *Strategies in discourse comprehension.* New York: Academic Press, 1983.

ELLIOT, S. Intellimetric: From here to validity. In: *Shermis, M.; Burstein, J., eds. Automatic Essay Scoring: A Cross-Disciplinary Perspective.*, Hillsdale, NJ: Lawrence Erlbaum Associates, 2003.

ELSNER, M.; CHARNIAK, E. 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, Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, p. 125–129 (*HLT '11*,).

FELTRIM, V. D. *Uma abordagem baseada em córpus e em sis-temas de crítica para a construção de ambientes web de auxílio à escrita acadêmica em português.* Tese de Doutoramento, Instituto de Computação e Matemática Computacional – Universidade de SãoPaulo, São Carlos, SP, 2004.

FELTRIM, V. D.; ALUÍSIO, S. M.; NUNES, M. G. V. Analysis of the rhetorical structure of computer science abstracts in portugese. In: *Archer D, Rayson P, Wilson A, McEnery T (eds) Proceedings of Corpus Linguistics*, 2003, p. 212—-218 (*UCREL Technical Papers*, v.16).

FELTRIM, V. D.; TEUFEL, S.; NUNES, M. G. V.; ALUÍSIO, S. M. 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, 2006, p. 233–246.

FILIPPOVA, K.; STRUBE, M. Extending the entity-grid coherence model to semantically related entities. In: *Proceedings of the Eleventh European Workshop on Natural Language Generation*, ENLG '07, Stroudsburg, PA, USA: Association for Computational Linguistics, 2007, p. 139–142 (*ENLG '07*,).

FRANZKE, M.; KINTSCH, E.; CACCAMISE, D.; JOHNSON, N.; DOOLEY, S. Summary street[r]: Computer support for comprehension and writing. In: *Journal of Educational Computing Research*, 2005, p. 53–80.

GOLUB, G.; REINSCH, C. Singular value decomposition and least squares solutions. *Numerische Mathematik*, v. 14, p. 403–420, 10.1007/BF02163027, 1970.

GRAESSER, A. C.; MCNAMARA, D. S.; LOUWERSE, M. M.; CAI, Z. Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments and Computers*, v. 36, p. 193–202, 2004.

GROSZ, B. J.; WEINSTEIN, S.; JOSHI, A. K. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, v. 21, p. 203–225, 1995.

HASLER, L. An investigation into the use of centering transitions for summarisation. In: *Proceedings of the 7th Annual Colloquium for the UK Special Interest Group in Computational Linguistics (CLUK)*, 2004, p. 100–107.

HASLER, L. Centering theory for evaluation of coherence in computer-aided summaries. In: *Proceedings of the 6th Language Resources and Evaluation Conference (LREC'08)*, 2008.

HIGGINS, D.; BURSTEIN, J.; MARCU, D.; GENTILE, C. Evaluating multiple aspects of coherence in student essays. In: *Human Language Technologies: The 2004 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics*, 2004.

HINKEL, E. *Teaching academic esl writing: Practical techniques in vocabulary and grammar.* Mahwah, New Jersey: Lawrence Erlbaum Associates, 2004.

JOACHIMS, T. Making large-scale support vector machine learning practical. In: SCHÖLKOPF, B.; BURGES, C. J. C.; SMOLA, A. J., eds. *Advances in kernel methods*, Cambridge, MA, USA: MIT Press, p. 169–184, 1999.

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

KARAMANIS, N. *Entity coherence for descriptive text structuring.* Tese de Doutoramento, Institute for Communicating and Collaborative Systems – School of Informatics – The University of Edinburgh, 2004.

KINTSCH, W. The potential of latent semantic analysis for machine grading of clinical case summaries. *J. of Biomedical Informatics*, v. 35, n. 1, p. 3–7, 2002.

KOCH, I. G. V. *A cosão textual.* São Paulo: Contexto, 1994.

KOCH, I. G. V.; TRAVAGLIA, L. C. *A coerência textual.* São Paulo: Contexto, 2003.

KRIEGSMAN, M.; BARLETTA, R. Building a case-based help desk application. *IEEE Expert: Intelligent Systems and Their Applications*, v. 8, n. 6, p. 18–26, 1993.

LANDAUER, T.; FOLTZ, P.; LAHAM, D. An introduction to latent semantic analysis. *Discourse processes*, v. 25, p. 259–284, 1998.

LANDAUER, T. K.; LAHAM, D.; FOLTZ, P. W. *Automated essay scoring and annotation of essays with the intelligent essay assessor.* Mahwah, NJ: Lawrence Erlbaum Associates, 2003.

Lin, Z.; Ng, H. T.; Kan, M.-Y. 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, Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, p. 997–1006 (*HLT '11*,).

Ljungstrand, P.; Johansson, H. *Intranet indexing using semantic document clustering.* Dissertação de Mestrado, Department of Informatics – Göteborg University, Göteborg, Sweden, 1998.

Mani, I. *Natural language processing.* John Benjamins Publishing Company, 2001.

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

Martins, R.; Nunes, G.; Hasegawa, R. Curupira: a functional parser for brazilian portuguese. In: *Proceedings of the 6th international conference on Computational processing of the Portuguese language*, PROPOR'03, Berlin, Heidelberg: Springer-Verlag, 2003, p. 179–183 (*PROPOR'03*,).

McNamara, D.; Louwerse, M.; McCarthy, P.; Graesser, A. Coh-metrix: Analysis of text on cohesion and language. v. 47, p. 292–330, 2010.

McNamara, D.; Ozuru, Y.; Graesser, A.; Louwerse, M. Validating coh-metrix, p. 573–578. 2006.

Miller, T. Essay assessment with latent semantic analysis. *Journal of Educational Computing Research*, v. 29, n. 4, p. 495–512, 2003.

Miltsakaki, E.; Kukich, K. Evaluation of text coherence for electronic essay scoring systems. *Natural Language Engeneering*, v. 10, p. 25–55, 2004.

Orăsan, C. PALinkA: a highly customizable tool for discourse annotation. In: *Proceedings of the 4th SIGdial Workshop on Discourse and Dialog*, Sapporo, Japan, 2003, p. 39 – 43.

Palomar, M.; Martínez-Barco, P. Computational approach to anaphora resolution in spanish dialogues. *Journal of Artificial Intelligence Research*, v. 15, p. 263 – 287, 2001.

Pardo, T. A. S. *Métodos para Análise Discursiva Automática.* Tese de Doutoramento, Universidade de São Paulo, São Carlos, 2005.

Perini, M. A. *Gramática descritiva do português.* São Paulo: Editora Ática, 2003.

Perini, M. A. *Gramática do português brasileiro.* São Paulo: Editora Ática, 2010.

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

QUINLAN, R. *C4.5: Programs for machine learning*. San Mateo, CA: Morgan Kaufmann Publishers, 1993.

RINO, L. H. M.; PARDO, T. A. S. *A coleção temário e a avaliação de sumarização automática*, v. 1. Lisboa, Portugal: IST Press, 1–17 p., 2007.

ROSSI, D.; PINHEIRO, C.; FEIER, N.; VIEIRA, R. Resolução de correferência em textos da língua portuguesa. 2001.

DOS SANTOS, C. N. *Aprendizado de máquina na identificação de sintagmas nominais: o caso do português brasileiro*. Dissertação de Mestrado, Instituto Militar de Engenharia, Rio de Janeiro, RJ, 2006.

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

SCARTON, C. E.; ALMEIDA, D. M.; ALUÍSIO, S. M. Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: adaptando as metricas do coh-metrix para o português. In: *Proceedings of STIL 2009*, 1 CD-ROM v1., 2009.

SCHWARM, S. E.; OSTENDORF, M. Reading level assessment using support vector machines and statistical language models. In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, MI, 2005, p. 523—-530.

SILVA, J.; BRANCO, A.; CASTRO, S.; REIS, R. Out-of-the-box robust parsing of portuguese. In: *Proceedings of the 9th International Conference on the Computational Processing of Portuguese*, PROPOR'10, 2010, p. 75—-85 (*PROPOR'10*,).

SILVA, J. F.; ROSA, J. L. G. Crt-ml - coreference resolution tool through machine learning. In: *Proceedings of PROPOR 2010, 9th International Conference on Computational Processing of the Portuguese Language, Extended Activities Proceedings*, PROPOR '10, 2010 (*PROPOR '10*,).

SOUZA, J. G.; GONÇALVES, P. N.; VIEIRA, R. Learning coreference resolution for portuguese texts. In: *Proceedings of the 8th international conference on Computational Processing of the Portuguese Language*, PROPOR '08, Berlin, Heidelberg: Springer-Verlag, 2008, p. 153–162 (*PROPOR '08*,).

Souza, V. M. A.; Feltrim, V. D. An analysis of textual coherence in academic abstracts written in portuguese (to be publised). In: *Proceedings of the Sixth Corpus Linguistics Conference: CL 2011*, Birmingham, UK, 2011.

Souza, V. M. A.; Feltrim, V. D. A coherence analysis module for scipo: providing suggestions for scientific abstracts written in portuguese. *Journal of the Brazilian Computer Society*, p. 1–15, 2012.

Strube, M.; Ponzetto, S. P. Wikirelate! computing semantic relatedness using wikipedia. In: *Proceedings of the 21st National Conference on Artificial Intelligence*, Boston, Mass., 2006, p. 1219—-1224.

Taboada, M.; Mann, W. C. Rhetorical structure theory: looking back and moving ahead. *Discourse Studies*, v. 8, p. 423–459, 2006.

Tetreault, J. R. A corpus-based evaluation of centering and pronoun resolution. *MIT Press Journal – Computational Linguistics*, v. 27, p. 507 – 520, 2001.

Teufel, S.; Moens, M. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational Linguistics*, v. 28, n. 4, p. 409–445, 2002.

Wade-Stein, D.; Kintsch, E. Summary street: Interactive computer support for writing. *Cognition and Instruction*, v. 22, n. 3, p. 333–362, 2004.

Walker, M.; Joshi, A.; Prince, E. *Centering theory in discourse.* Clarendon Press Oxford, 1998.

Witten, H. I.; Frank, E. *Data mining - practical machine learning tools and techniques.* Morgan Kaufmann – Elsevier, 2005.

Yokono, H.; Okumura, M. Incorporating cohesive devices into entity grid model in evaluating local coherence of japanese text. In: Gelbukh, A., ed. *Computational Linguistics and Intelligent Text Processing*, v. 6008 de *Lecture Notes in Computer Science*, Springer Berlin/Heidelberg, p. 303–314, 2010.

Apêndice

A

Resultados do Modelo Grade de Entidades para o Experimento (2): Julgamento Humano

Tabela A.1: Resultados para o córpus anotado manualmente usando o algoritmo Naïve Bayes.

			Sem Problema (117)			Com problema (22)			Média ponderada (139)		
*_TT-	Acertos (%)	Kappa	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
Sintático+ Saliência-	61,151	0,211	0,944	0,573	0,713	0,265	0,818	0,400	0,836	0,612	0,663
Sintático+ Saliência+	73,381	0,053	0,851	0,829	0,840	0,200	0,227	0,213	0,748	0,734	0,741
Sintático- Saliência-	66,187	0,211	0,917	0,658	0,766	0,273	0,682	0,390	0,815	0,662	0,707
Sintático- Saliência+	82,739	0,168	0,860	0,949	0,902	0,400	0,182	0,250	0,788	0,827	0,799
*_TT+											
Sintático+ Saliência-	69,065	0,262	0,930	0,684	0,788	0,302	0,727	0,427	0,831	0,691	0,731
Sintático+ Saliência+	77,698	0,114	0,858	0,880	0,869	0,263	0,227	0,244	0,764	0,777	0,770
Sintático- Saliência-	70,504	0,223	0,904	0,726	0,806	0,289	0,591	0,388	0,807	0,705	0,740
Sintático- Saliência+	82,734	0,168	0,860	0,949	0,902	0,400	0,182	0,250	0,788	0,827	0,799

Tabela A.2: Resultados para o córpus anotado manualmente usando o algoritmo SMO.

			Sem Problema (117)			Com problema (22)			Média ponderada (139)		
*_TT-	Acertos (%)	Kappa	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
Sintático+ Saliência-	84,173	0,000	0,842	1,000	0,914	0,000	0,000	0,000	0,709	0,842	0,769
Sintático+ Saliência+	85,612	0,144	0,854	1,000	0,921	1,000	0,091	0,167	0,877	0,856	0,802
Sintático- Saliência-	84,173	0,000	0,842	1,000	0,914	0,000	0,000	0,000	0,709	0,842	0,769
Sintático- Saliência+	83,453	-0,014	0,841	0,991	0,910	0,000	0,000	0,000	0,708	0,835	0,766
*_TT+											
Sintático+ Saliência-	83,453	-0,014	0,841	0,991	0,910	0,000	0,000	0,000	0,708	0,835	0,766
Sintático+ Saliência+	85,612	0,144	0,854	1,000	0,921	1,000	0,091	0,167	0,877	0,856	0,802
Sintático- Saliência-	84,173	0,000	0,842	1,000	0,914	0,000	0,000	0,000	0,709	0,842	0,769
Sintático- Saliência+	84,892	0,127	0,853	0,991	0,917	0,667	0,091	0,160	0,823	0,849	0,797

Tabela A.3: Resultados para o córpus anotado manualmente usando o algoritmo J48.

			Sem Problema (117)			Com problema (22)			Média ponderada (139)		
*_TT-	Acertos (%)	Kappa	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
Sintático+ Saliência-	82,014	0,256	0,877	0,915	0,895	0,412	0,318	0,359	0,803	0,820	0,810
Sintático+ Saliência+	89,362	0,515	0,906	0,975	0,939	0,769	0,455	0,571	0,885	0,894	0,882
Sintático- Saliência-	83,453	0,183	0,862	0,957	0,907	0,444	0,182	0,258	0,796	0,835	0,804
Sintático- Saliência+	91,367	0,650	0,934	0,966	0,950	0,778	0,636	0,700	0,909	0,914	0,910
*_TT+											
Sintático+ Saliência-	81,295	0,271	0,882	0,897	0,890	0,400	0,364	0,381	0,806	0,813	0,809
Sintático+ Saliência+	88,653	0,494	0,906	0,966	0,935	0,714	0,455	0,556	0,876	0,887	0,876
Sintático- Saliência-	83,453	0,183	0,862	0,957	0,907	0,444	0,182	0,258	0,796	0,835	0,804
Sintático- Saliência+	91,367	0,650	0,934	0,966	0,950	0,778	0,636	0,700	0,909	0,914	0,910

Tabela A.4: Resultados para o córpus anotado manualmente e balanceado com SMOTE usando o algoritmo Naïve Bayes.

			Sem Problema (117)			Com problema (110)			Média ponderada (227)		
*_TT-	Acertos (%)	Kappa	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
Sintático+ Saliência-	72,687	0,460	0,877	0,547	0,674	0,656	0,918	0,765	0,770	0,727	0,718
Sintático+ Saliência+	76,212	0,525	0,794	0,726	0,759	0,733	0,800	0,765	0,765	0,762	0,762
Sintático- Saliência-	64,317	0,294	0,750	0,462	0,571	0,594	0,836	0,694	0,674	0,643	0,631
Sintático- Saliência+	65,198	0,290	0,603	0,949	0,738	0,860	0,336	0,484	0,728	0,652	0,615
*_TT+											
Sintático+ Saliência-	77,533	0,554	0,884	0,650	0,749	0,709	0,909	0,797	0,799	0,775	0,772
Sintático+ Saliência+	79,736	0,596	0,838	0,752	0,793	0,762	0,845	0,802	0,801	0,797	0,797
Sintático- Saliência-	71,366	0,433	0,842	0,547	0,663	0,649	0,891	0,751	0,749	0,714	0,706
Sintático- Saliência+	88,987	0,780	0,934	0,846	0,888	0,851	0,936	0,892	0,894	0,890	0,890

Tabela A.5: Resultados para o córpus anotado manualmente e balanceado com SMOTE usando o algoritmo SMO.

			Sem Problema (117)			Com problema (110)			Média ponderada (227)		
*_TT-	Acertos (%)	Kappa	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
Sintático+ Saliência-	73,568	0,478	0,913	0,538	0,677	0,658	0,945	0,776	0,790	0,736	0,725
Sintático+ Saliência+	83,260	0,663	0,784	0,932	0,852	0,909	0,727	0,808	0,845	0,833	0,830
Sintático- Saliência-	68,723	0,383	0,871	0,462	0,603	0,618	0,927	0,742	0,748	0,687	0,670
Sintático- Saliência+	63,436	0,253	0,589	0,957	0,730	0,865	0,291	0,435	0,723	0,634	0,587
*_TT+											
Sintático+ Saliência-	83,260	0,667	0,916	0,744	0,821	0,773	0,927	0,843	0,846	0,833	0,832
Sintático+ Saliência+	85,612	0,144	0,854	1,000	0,921	1,000	0,091	0,167	0,877	0,856	0,802
Sintático- Saliência-	73,128	0,466	0,804	0,632	0,708	0,681	0,836	0,751	0,745	0,731	0,729
Sintático- Saliência+	86,784	0,735	0,860	0,889	0,874	0,877	0,845	0,861	0,868	0,868	0,868

Tabela A.6: Resultados para o córpus anotado manualmente e balanceado com SMOTE usando o algoritmo J48.

			Sem Problema (117)			Com problema (110)			Média ponderada (227)		
*_TT-	Acertos (%)	Kappa	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
Sintático+ Saliência-	80,617	0,612	0,812	0,812	0,812	0,800	0,800	0,800	0,806	0,806	0,806
Sintático+ Saliência+	90,393	0,808	0,915	0,899	0,907	0,893	0,909	0,901	0,904	0,904	0,904
Sintático- Saliência-	77,093	0,544	0,849	0,675	0,752	0,716	0,873	0,787	0,785	0,771	0,769
Sintático- Saliência+	91,189	0,824	0,929	0,897	0,913	0,895	0,927	0,911	0,913	0,912	0,912
*_TT+											
Sintático+ Saliência-	80,617	0,612	0,807	0,821	0,814	0,806	0,791	0,798	0,806	0,806	0,806
Sintático+ Saliência+	90,393	0,808	0,915	0,899	0,907	0,893	0,909	0,901	0,904	0,904	0,904
Sintático- Saliência-	76,652	0,536	0,848	0,667	0,746	0,711	0,873	0,784	0,782	0,767	0,764
Sintático- Saliência+	91,630	0,833	0,922	0,915	0,918	0,910	0,918	0,914	0,916	0,916	0,916