Proposta recebida em Outubro 2018 e aceite para publicação em Maio 2019.

Avaliando Atributos para a Classificação de Estrutura Retórica em Resumos Científicos

Evaluating features for rhetorical structure classification in scientific abstracts

Alessandra Harumi Iriguti
Universidade Estadual de Maringá
alehairi@gmail.com

Valéria Delisandra Feltrim
Universidade Estadual de Maringá
vdfeltrim@uem.br

Resumo

A classificação de estrutura retórica é uma tarefa de PLN na qual se busca identificar os componentes retóricos de um discurso e seus relacionamentos. No caso deste trabalho, buscou-se identificar automaticamente categorias em nível de sentenças que compõem a estrutura retórica de resumos científicos. Especificamente, o objetivo foi avaliar o impacto de diferentes conjuntos de atributos na implementação de classificadores retóricos para resumos científicos escritos em português. Para isso, foram utilizados atributos superficiais (extraídos como valores TF-IDF e selecionados com o teste χ^2), atributos morfossintáticos (implementados pelo classificador AZPort) e atributos extraídos a partir de modelos de *word embeddings* (Word2Vec, Wang2Vec e GloVe, todos previamente treinados). Tais conjuntos de atributos, bem como as suas combinações, foram usados para o treinamento de classificadores usando os seguintes algoritmos de aprendizado supervisionado: *Support Vector Machines*, *Naive Bayes*, *K-Nearest Neighbors*, *Decision Trees* e *Conditional Random Fields* (CRF). Os classificadores foram avaliados por meio de validação cruzada sobre três *corpora* compostos por resumos de teses e dissertações. O melhor resultado, 94% de F1, foi obtido pelo classificador CRF com as seguintes combinações de atributos: (i) Wang2Vec–*Skip-gram* de dimensões 100 com os atributos provenientes do AZPort; (ii) Wang2Vec–Skip-gram e GloVe de dimensão 300 com os atributos do AZPort; (iii) TF-IDF, AZPort e *embeddings* extraídos com os modelos Wang2Vec–*Skip-gram* de dimensões 100 e 300 e GloVe de dimensão 300. A partir dos resultados obtidos, conclui-se que os atributos provenientes do classificador AZPort foram fundamentais para o bom desempenho do classificador CRF, enquanto que a combinação com *word embeddings* se mostrou válida para a melhoria dos resultados.

Palavras chave

processamento de língua natural, classificação de estrutura retórica, resumos científicos em português

Abstract

Rhetorical structure classification is a NLP task in which we want to identify the rhetorical components of a discourse and its relationships. In this work, we aimed at automatically identifying categories at the sentential level that make up the rhetorical structure of scientific abstracts. Specifically, the purpose was to evaluate the impact of different sets of attributes on the implementation of rhetorical classifiers for scientific abstracts written in Portuguese. For this, we used superficial features (extracted as TF-IDF values and selected with the χ^2 test), morphosyntactic features (implemented by the AZPort classifier) and features extracted from word embeddings models (Word2Vec, Wang2Vec and GloVe, all of them previously trained). These sets of features, as well as their combinations, were used to train the following supervised learning classifiers: Support Vector Machines, Naive Bayes, K-Nearest Neighbors, Decision Trees and Conditional Random Fields (CRF). We evaluated the classifiers through cross-validation on three *corpora* composed by abstracts of theses and dissertations. The best result, 94% of F1, was obtained by the CRF classifier with the following combinations of features: (i) Wang2Vec–Skip-gram of 100 dimension with features from AZPort; (ii) Wang2Vec–Skip-gram and GloVe of 300 dimension with AZPort features; (iii) TF-IDF, AZPort features and embeddings extracted by Wang2Vec–Skip-gram model with dimensions of 100 and 300, and by GloVe model of dimension 300. From the results, we concluded that the AZPort features were fundamental for the performance of the CRF classifier, while the combination with word embeddings proved valid for improving the results.

Keywords

natural language processing, rhetorical structure classification, scientific abstracts in Portuguese

DOI: 10.21814/lm.11.1.273

1 Introdução

A classificação de estrutura retórica é uma tarefa de Processamento de Língua Natural (PLN) em que se busca identificar os componentes retóricos de um discurso e seus relacionamentos. São essas relações que definem como o conteúdo apresentado está relacionado entre si e como cada parte do texto contribui para satisfazer os objetivos e intenções do autor (Romeiro, 2016). Os componentes retóricos podem ser analisados com uma granularidade mais fina, como no caso da Teoria de Estrutura Retórica (Rhetorical Structure Theory - RST) (Mann & Thompson, 1987, 1988), que identifica relações entre segmentos que podem, por exemplo, compor uma mesma sentença, ou com uma granularidade mais grossa, em que os componentes retóricos são blocos de uma ou mais sentenças que juntos revelam a macro-estrutura de um texto.

A organização retórica em nível de macro-estrutura tem sido especialmente investigada no contexto dos textos científicos. Do ponto de vista linguístico, esses estudos buscam identificar modelos de estrutura retórica que caracterizem os movimentos retóricos observados nas diferentes seções desses textos. Weissberg & Buker (1990), Booth et al. (2005) e Swales & Feak (1994) são exemplos de autores que investigaram estruturas retóricas específicas do gênero científico. Booth et al. (2005) e Weissberg & Buker (1990) propuseram modelos para a estruturação de diversas seções de um texto científico, como o resumo, a introdução e a conclusão. Já Swales & Feak (1994) propuseram um modelo para a estruturação de introduções, que posteriormente foi adaptado por outros pesquisadores para outras seções (Anthony & Lashkia, 2003; Teufel & Moens, 2002). Com exceção do trabalho de Teufel & Moens (2002), esse estudos tiveram como motivação auxiliar a escrita de textos científicos, uma tarefa que é reconhecidamente difícil, especialmente para escritores iniciantes.

Do ponto de vista computacional, os estudos que tratam da classificação retórica de textos buscam construir ferramentas capazes de identificar componentes retóricos de forma automática, tendo como base um modelo de estrutura retórica que pode se aplicar ao texto completo ou apenas a uma de suas seções. Dada a importância dos resumos (abstract) para a indexação de artigos científicos em bases de dados especializadas, bem como para a realização de mapeamentos sistemáticos, várias pesquisas focam a classificação retórica de resumos (Anthony & Lashkia, 2003; Hirohata et al., 2008; Guo et al., 2011; Dayrell et al., 2012; Yepes et al., 2013; Moura, 2018). Assim como noutras áreas do PLN, a maioria dos trabalhos focam a língua inglesa. Para a língua portuguesa, destacam-se os trabalhos de Feltrim et al. (2004) e Andreani & Feltrim (2015), ambos relacionados ao classificador AZPort.

Nesse contexto, o estudo aqui apresentado teve por objetivo avaliar diferentes conjuntos de atributos e algoritmos de classificação aplicados à construção de classificadores retóricos sentenciais para resumos científicos escritos em português. Os *corpora* usados no desenvolvimento foram os mesmos usados por Feltrim et al. (2004) e Andreani & Feltrim (2015), ambos compostos de resumos de teses e dissertações na área de Ciência da Computação, manualmente anotados de acordo com um modelo de sete categorias retóricas.

Na Figura 1, é apresentado um resumo anotado extraído do *corpus* de Feltrim et al. (2004). Como é possível observar, a unidade mínima de anotação é uma sentença e nem todas as categorias retóricas possíveis aparecem no resumo. De fato, embora o modelo retórico preveja sete categorias em uma ordem específica, os resumos não apresentam obrigatoriamente todas as categorias e nem seguem uma ordem estrita entre elas.

Os atributos avaliados no estudo incluíram valores TF-IDF, *word embeddings* e os atributos usados pelo classificador AZPort. Foram avaliados diferentes modelos de *word embeddings*, com variações do número de dimensões e da estratégia de geração dos *embeddings* das sentenças. Os classificadores foram induzidos por diferentes algoritmos supervisionados e com diferentes configurações de atributos. A análise dos resultados mostrou que os atributos usados pelo AZPort foram fundamentais para o desempenho dos classificadores, mas que a combinação com outros atributos, em particular com *word embeddings*, foi benéfica.

O restante deste artigo está organizado como segue. Na Seção 2 é apresentada uma visão geral da área, bem como a descrição dos trabalhos relacionados à classificação retórica de resumos científicos em português. Em seguida, na Seção 3.1 são descritos os *corpora* usados neste estudo, bem como os atributos extraídos e os classificadores avaliados. Os resultados experimentais são apresentados e analisados na Seção 4. O desempenho dos classificadores foi analisado focando nos seguintes aspectos: combinação de atributos, modelos de *word embeddings*, dimensão do vetor de *word embeddings* e estratégia para obtenção do *embedding* de uma sentença. Por fim, na Seção 5 são feitas as conclusões a respeito do estudo e apresentadas direções para trabalhos futuros.

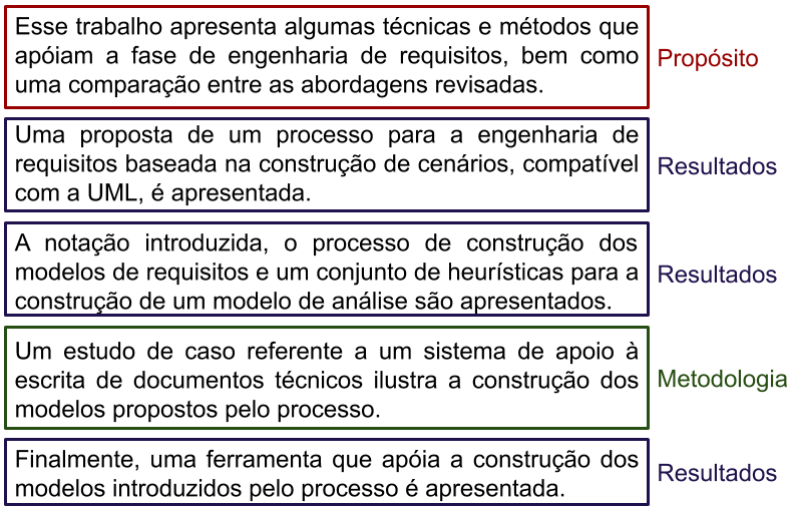

Figura 1: Exemplo de classificação de estrutura retórica das sentenças de um resumo.

2 Trabalhos relacionados

Na literatura, a classificação de estrutura retórica em textos científicos tem sido tratada como um problema de classificação sentencial, no qual se busca associar categorias retóricas às sentenças de um texto. Embora tal tarefa tenha sido abordada como um problema de classificação multirrótulo por Dayrell et al. (2012), a maioria dos trabalhos relacionados busca associar uma única categoria a cada sentença.

Com relação aos métodos de aprendizado, destaca-se o uso de métodos supervisionados e, portanto, dependentes de *corpora* anotados. Exceções são os trabalhos de Guo et al. (2011) e Guo et al. (2013), nos quais foram usados métodos semissupervisionados, e Varga et al. (2012) e Reichart & Korhonen (2012), os quais propuseram soluções não-supervisionadas.

Há uma grande variação com relação aos atributos usados para a classificação, sendo eles, em sua maioria, baseados em informações superficiais de estrutura, léxicas e morfossintáticas (Anthony & Lashkia, 2003; Mullen et al., 2005; Hirohata et al., 2008; Pendar & Cotos, 2008; Merity et al., 2009; Guo et al., 2011; Liakata et al., 2012; Yepes et al., 2013; Fisas et al., 2015). Apenas Teufel & Moens (2002) propuseram o uso de informação semântica, que se deu por meio da identificação de padrões referentes a agentes e ações.

Com relação ao modelo que define as categorias retóricas a serem identificadas pelos classificadores, também há uma grande variação entre os trabalhos encontrados, pois tais modelos costumam ser ajustados ao contexto das aplicações pretendidas por cada pesquisa. Enquanto alguns modelam os movimentos retóricos do texto como um todo (Teufel & Moens, 2002; Merity et al., 2009; Liakata et al., 2012; Varga et al., 2012; Fisas et al., 2015), outros modelam uma seção específica, como o resumo (Anthony & Lashkia, 2003; Hirohata et al., 2008; Guo et al., 2011; Dayrell et al., 2012; Reichart & Korhonen, 2012; Yepes et al., 2013) e a introdução (Pendar & Cotos, 2008).

A seguir são apresentados os trabalhos de Feltrim et al. (2004) e Andreani & Feltrim (2015), pois ambos abordaram a classificação da estrutura retórica de resumos em português. Também é apresentado o trabalho de Teufel & Moens (2002), uma vez que o mesmo serviu de ponto de partida para diversos trabalhos na área, incluindo o de Feltrim et al. (2004).

Teufel & Moens (2002) propuseram a segmentação de um artigo científico em zonas argumentativas, que juntas compõem a sua estrutura retórica. Esse modelo foi chamado de *Argumentative Zoning* (AZ). A motivação para a criação do AZ foi a sua aplicação na sumarização automática de artigos científicos. As zonas, ou categorias, previstas pelo AZ são as seguintes: Objetivo (objetivo específico do artigo), Textual (descrição da estrutura da seção), Próprio (descrição neutra de metodologia, resultados e discussão do artigo), Contexto (conhecimento científico aceito), Contraste (comparações ou diferenças com outros trabalhos, explicitando seus pontos fracos), Base (conformidades com ou continuações de outros trabalhos) e Outro (descrição neutra de trabalhos de outros pesquisadores).

Para a classificação automática, o AZ utilizou um classificador *Naive Bayes* e os atributos extraídos incluíram informações superficiais, gramaticais e semânticas. O *corpus* utilizado para aprendizado foi composto por 80 artigos,

totalizando 12.188 sentenças, e o classificador foi avaliado por meio de validação cruzada de 10 partições. Os resultados gerais obtidos pelo classificador AZ foram 50% de Macro-F, 0,45 de *Kappa* e 73% de acurácia.

Tendo por base o trabalho de Teufel & Moens (2002), Feltrim et al. (2004) propuseram o AZPort (*Argumentative Zoning for Portuguese*), uma adaptação do AZ para a língua portuguesa com foco na classificação retórica de resumos científicos. Uma vez que o AZPort teve como motivação a sua utilização em uma ferramenta de auxílio à escrita, o modelo retórico usado pelo classificador foi adaptado para esse contexto, resultando no seguinte conjunto de categorias: Contexto (B), Lacuna (G), Propósito (P), Método (M), Resultado (R), Conclusão (C) e Estrutura (O).

O AZPort utiliza um conjunto de oito atributos derivados dos atributos propostos por Teufel & Moens (2002), conforme descritos na Tabela 1. Assim como o AZ, o AZPort é um classificador bayesiano, de modo que foi necessário utilizar um conjunto de exemplos para o treinamento do modelo. O *corpus* utilizado para esse fim foi chamado de CorpusDT, sendo composto por 52 resumos que totalizam 366 sentenças, conforme descrito na Seção 3.1. O AZPort foi avaliado por meio de 13 rodadas de validação cruzada de 13 partições, obtendo 60% de Macro-F, 0,65 de *Kappa* e 72% de acurácia.

Andreani (2017) realizou um estudo acerca da aplicação de algoritmos de predição estruturada à tarefa em questão. A fim de encontrar o melhor algoritmo para a classificação de estrutura retórica, os seguintes algoritmos foram avaliados: Modelo Oculto de Markov (HMM), Modelo de Markov de Entropia Máxima (MEMM), *Conditional Random Fields* (CRF) e *Structured Support Vector Machines* (SSVM). Para que fosse possível realizar a comparação com o AZPort, foi usado o mesmo modelo de estrutura retórica (B, G, P, M, R, C e O) e dois *corpora* foram empregados no treinamento e teste dos classificadores: o CorpusDT e o *Corpus* 466 (Andreani & Feltrim, 2015). Assim como o CorpusDT, o *Corpus* 466 é composto por resumos extraídos de teses e dissertações em Ciência da Computação, totalizando 466 sentenças manualmente anotadas.

Os resultados de Andreani (2017) foram calculados a partir de 30 execuções de validação cruzada de 13 partições. Os atributos utilizados incluíram os atributos do AZPort exceto Citação e Histórico, *n-gramas*, segmentação e janela deslizante. O atributo *n-gramas* é composto por valores TF-IDF; o atributo segmentação indica o início, meio e fim de segmentos (sequências de sentenças) com uma mesma categoria retórica; e, por fim, a janela deslizante inclui os atributos das $k = \{0, 1, 2\}$ sentenças vizinhas. Os resultados experimentais mostraram que o melhor desempenho foi obtido pelo classificador CRF, com F1-*score* de 68%, o que representou uma melhoria de 7% em relação ao desempenho do AZPort.

3 Desenvolvimento

Nesta seção são descritos os *corpora* empregados neste estudo, bem como os atributos e classificadores avaliados.

3.1 *Corpora*

Conforme mencionado anteriormente, os *corpora* utilizados neste estudo são constituídos de resumos de teses e dissertações da área de Ciência da Computação, escritos em português do Brasil. Esses resumos foram coletados e anotados como parte dos trabalhos de Feltrim et al. (2004) e Andreani & Feltrim (2015). Todas as sentenças foram anotadas manualmente por três anotadores treinados e com experiência em escrita científica. A concordância entre os anotadores medida por meio da estatística *Kappa* (Siegel & Castellan Jr., 1988) foi de 0,695.

Os *corpora*, que neste estudo foram chamados de 366, 466 e 832, estão anotados com as categorias previstas pelo classificador AZPort, sendo elas: Contexto, Lacuna, Propósito, Resultado, Método, Conclusão e Estrutura. Os *corpora* 366 (chamado por Feltrim et al. (2004) de CorpusDT) e 466 possuem 52 resumos cada, totalizando, respectivamente, 366 e 466 sentenças. O *corpus* 832 corresponde a união dos *corpora* 366 e 466 e, consequentemente, possui 104 resumos, totalizando 832 sentenças.

A distribuição de categorias observada nos *corpora* é mostrada na Tabela 2. Como se pode notar, a distribuição é semelhante nos *corpora* 366 e 466, sendo as categorias Contexto e Resultado as mais frequentes e as categorias Conclusão e Estrutura as menos frequentes. A prevalência da categoria Resultado é comum em *corpora* de resumos científicos, uma vez que os autores buscam enfatizar os resultados encontrados em seus trabalhos (Hirohata et al., 2008; Moura, 2018). Já a alta frequência da categoria Contexto é uma característica particular dos *corpora* usados neste trabalho e se deve ao fato de os resumos terem sido extraídos a partir de teses e dissertações. Os resumos desses tipos de trabalhos tendem a ser mais longos, permitindo que os autores contex-

Atributo	Descrição	Valores possíveis
Tamanho	Qual é o tamanho da sentença em comparação aos limiares 20 e 40 palavras?	curta, média ou longa
Localização	Qual é a posição da sentença no resumo?	primeira, segunda, mediana, penúltima ou última
Citação	A sentença contém citações?	sim ou não
Expressão	Que tipo de expressão padrão a sentença contém?	B, C, G, M, P, R ou *noExpr*
Tempo	Qual é o tempo do primeiro verbo finito da sentença?	IMP, PRES, PAST, FUT, PRES-CPO, PAST-CPO, FUT-CPO, PRES-CT, PAST-CT, FUT-CT, PRES-CPO-CT, PAST-CPO-CT, FUT-CPO-CT ou *noVerb*
Voz	Qual é a voz do primeiro verbo finito da sentença?	passiva, ativa ou *noVerb*
Modal	O primeiro verbo finito da sentença é modal?	sim, não ou *noVerb*
Histórico	Qual é a categoria da sentença anterior?	_, B, C, G, M, O, P ou R

Tabela 1: Conjunto de atributos utilizado do AZPort (adaptado de Feltrim (2004)).

tualizem melhor suas áreas de pesquisa. Com relação a categoria Estrutura, que é minoritária, cabe destacar que é incomum que informações a respeito da organização do texto sejam incluídas em resumos, o que justifica a baixa frequência de sentenças dessa categoria nos *corpora*.

3.2 Atributos

Foram utlizados todos os atributos usados pelo AZPort (Tabela 1) com a adição de um novo atributo que registra a posição relativa da sentença no resumo. Também foram extraídos atributos por meio de TF-IDF e *word embeddings*.

A extração dos vetores com valores TF-IDF (*Term Frequency–Inverse Document Frequency*) foi feita com base em unigramas. Em seguida, os 100 melhores atributos foram selecionados por meio do teste χ^2 (qui-quadrado). A escolha pelo uso dos 100 melhores unigramas se deu por experimentação. Foram avaliadas diferentes configurações de vetores resultantes das combinações de diferentes valores de n (1, 2 e 3) para os n-gramas e diferentes valores de corte (50, 100, 250, 500 e 1000) para o teste χ^2.

Os atributos baseados em *word embeddings* (WE) foram extraídos utilizando-se os modelos CBOW, *Skip-gram* — ambos com as ferramentas Word2Vec (Mikolov et al., 2013) e Wang2Vec (Ling et al., 2015) — e GloVe (Pennington et al., 2014), todos eles previamente treinados para o português por Hartmann et al. (2017). Tais modelos estão disponíveis no repositório NILC–*Embeddings*[1] do NILC (ICMC–USP)[2] e foram gerados a partir de um *corpus* em português brasileiro e europeu, contendo textos de fontes e gêneros variados (Hartmann et al., 2017). Cada modelo está disponível no repositório com vetores de 50, 100, 300, 600 e 1000 dimensões.

Modelos de *word embeddings*, todavia, retornam vetores para palavras e, neste trabalho, a unidade de classificação é uma sentença. Dessa forma, os *embeddings* das sentenças foram gerados por meio de duas estratégias — média simples e média ponderada pelo valor IDF (*Inverse Document Frequency*). Assim, o *embedding* de uma sentença correspondeu à média simples ou à média ponderada dos *word embeddings* das palavras que a formavam.

Além desses atributos serem utilizados de maneira individual, também foram feitas combinações entre eles. Por meio da concatenação dos vetores extraídos de cada um dos atributos, foram feitas as seguintes combinações:

- TF-IDF + atributos AZPort;
- *Word embeddings* + TF-IDF;
- *Word embeddings* + atributos AZPort; e
- *Word embeddings* + TF-IDF + atributos AZPort.

[1] NILC–*Embeddings*. http://www.nilc.icmc.usp.br/embeddings

[2] Núcleo Interinstitucional de Linguística Computacional. http://nilc.icmc.usp.br

Categoria	***Corpus* 366**	***Corpus* 466**	***Corpus* 832**
Contexto	77	179	256
Lacuna	36	36	72
Propósito	65	68	133
Método	45	59	104
Resultado	117	103	220
Conclusão	20	20	40
Estrutura	6	1	7
Total	**366**	**466**	**832**

Tabela 2: Número de sentenças de cada *corpus* (adaptado de Andreani (2017)).

Em todas as combinações de atributos avaliadas, foram adicionados à representação de uma sentença s_i os atributos das sentenças s_{i-1} e s_{i+1} sempre que possível — quando a sentença é a primeira do resumo, não há uma sentença s_{i-1}; e quando a sentença é a última, não há uma sentença s_{i+1}.

3.3 Classificadores

Os seguintes classificadores foram usados: *k-nearest neighbors* (K-NN), *Naive Bayes* — com as variações Gaussiana (G-NB) e Bernoulli (B-NB) —, *Decision Trees* (DT) — com a implementação do algoritmo CART—, *Support Vector Machines* (SVM) — com *kernels* linear e *Radial-Basis Function* (RBF) — e *Conditional Random Fields* (CRF). Com exceção do algoritmo CRF, foram usadas as implementações fornecidas pelas bibliotecas *scikit-learn*[3]. Para o CRF foi utilizada a biblioteca *sklearn-crfsuite*[4], uma versão da ferramenta *CRFsuite*[5] que é compatível com os estimadores da *scikit-learn*.

Para o algoritmo K-NN, foram considerados três vizinhos ($n\_neighbors = 3$) e distância Euclidiana; o restante dos parâmetros foram usados com valores *default*. Os algoritmos bayesianos, G-NB e B-NB, foram utilizados com seus parâmetros *default*. No DT, o único parâmetro com valor alterado foi $random\_state = 0$, o qual determina a semente para gerar números aleatórios.

Para o classificador SVM, foram alterados os seguintes parâmetros: *kernel*, C (parâmetro de penalidade do termo de erro) e *gamma* (coeficiente de *kernel*, no caso, apenas para RBF). O primeiro, *kernel* =‘*linear*’ ou *kernel* =‘*rbf*’, determina se o *kernel* a ser usado é linear ou RBF, respectivamente. No SVM-linear, para os *corpora* 366 e 832, foi usado $C = 100$; e, para o *corpus* 466, $C = 1000$. Já no SVM-RBF, para todos os *corpora*, foram usados $C = 1000$ e $gamma = 0.001$. Esses parâmetros foram escolhidos por meio da execução do algoritmo *Grid Search* para maximizar o desempenho com base nos atributos TF-IDF.

Para o classificador CRF, os parâmetros usados foram os seguintes: *algorithm = ‘lbfgs’, c1 = 0.1, c2 = 0.1, max_iterations = 100, all_possible_transitions = True*. O parâmetro *algorithm* especifica o algoritmo de treinamento, neste caso, gradiente descente usando o método L-BFGS (*default*); os parâmetros *c1* e *c2* definem os coeficientes de regularização L1 e L2, respectivamente; *max_iterations* define o número máximo de iterações para a otimização; e *all_possible_transitions = True* especifica que todas as transições possíveis devem ser geradas, mesmo as que não ocorrem no conjunto de treinamento. O restante dos parâmetros foram usados com seus valores *default*.

4 Resultados

Nesta seção são apresentados os resultados dos experimentos realizados com os *corpora* e os classificadores descritos. Em todos os experimentos, os classificadores foram avaliados por meio de validação cruzada estratificada de 10 partições. Embora a classificação seja feita por sentença, a geração das partições foi feita a partir de um conjunto de resumos, de modo a garantir que todas as sentenças de um mesmo resumo estejam em uma mesma partição. Vale destacar ainda que a mesma divisão de partições foi usada na avaliação de todos os classificadores.

A Tabela 3 apresenta os melhores valores de medida F1 obtidos por cada classificador usando cada uma das combinações de atributos. Nessa tabela e nos gráficos desta seção, RBF corresponde ao classificador SVM com *kernel* RBF e SVM corresponde ao classificador SVM com *ker-*

[3] http://scikit-learn.org

[4] https://sklearn-crfsuite.readthedocs.io/en/latest/

[5] http://www.chokkan.org/software/crfsuite/

Atributos	**RBF**	**SVM**	**K-NN**	**G-NB**	**B-NB**	**DT**	**CRF**
TF-IDF	46%	57%	39%	23%	45%	46%	56%
WE	50%	49%	37%	**74%**	44%	36%	55%
AZPort	66%	66%	56%	26%	64%	**62%**	92%
TF-IDF + AZPort	66%	62%	51%	33%	59%	61%	92%
WE + TF-IDF	54%	54%	38%	33%	49%	43%	58%
WE + AZPort	**71%**	**71%**	**61%**	59%	**68%**	60%	**94%**
Todos os atributos	**71%**	69%	53%	41%	62%	60%	**94%**
Média	61%	61%	48%	41%	56%	53%	77%
Desvio Padrão	10%	8%	10%	19%	10%	11%	20%

Tabela 3: Melhores resultados obtidos por classificador e combinação de atributos.

nel linear. O melhor desempenho obtido por cada algoritmo está destacado em negrito.

O melhor desempenho do SVM-RBF (71%) foi obtido no *corpus* 466 com as seguintes combinações: WE (Wang2Vec–*Skip-gram*–600) com atributos do AZPort e todos os atributos (GloVe–1000 combinado com os atributos do AZPort e TF-IDF). Na primeira combinação, os *embeddings* das sentenças foram calculados por média ponderada e, na segunda, por média simples. O SVM-linear também teve seu melhor desempenho (71%) obtido no *corpus* 466, porém apenas para a combinação de WE com atributos do AZPort. Nesse caso, o modelo de WE usado foi Wang2Vec–*Skip-gram*–600 e os *embeddings* das sentenças foram obtidos pela média simples.

O K-NN obteve o menor dos melhores desempenhos, 63%. Esse resultado foi obtido no *corpus* 466 com a combinação de WE com os atributos do AZPort. Os modelos de WE nesse caso foram dois: Wang2Vec–Skip-gram de dimensão 600 com a média ponderada e GloVe de dimensões 50 e 100 com a média simples.

Os algoritmos bayesianos alcançaram 74% e 68% com as variações gaussiana (G-NB) e Bernoulli (B-NB), respectivamente. Vale notar que o G-NB atingiu o segundo melhor desempenho entre os algoritmos avaliados utilizando apenas os atributos WE. Esse resultado foi alcançado com o modelo Word2Vec–CBOW–1000 com *embeddings* gerados pelas médias ponderada e simples no *corpus* 366 e com o mesmo modelo, mas apenas com a média ponderada, para o *corpus* 466. Já o B-NB atingiu 68% para o *corpus* 466, com os modelos Word2Vec–*Skip-gram*–50 e GloVe–50 usando médias simples e ponderada, respectivamente.

O melhor desempenho do DT (62%) também foi obtido no *corpus* 466, porém utilizando apenas os atributos do AZPort. Esse foi o único classificador cujo melhor resultado que não incluiu WE.

O melhor desempenho observado (94%) foi obtido pelo classificador CRF no *corpus* 466. Esse resultado foi alcançado usando a combinação de WE (Wang2Vec–*Skip-gram*–100 com média simples e Wang2Vec–*Skip-gram* e GloVe, ambos de dimensão 300, com média ponderada) com os atributos do AZPort. O mesmo desempenho foi obtido usando-se todos os atributos (Word2Vec–*Skip-gram* de dimensões 100 e 300 com média ponderada; e GloVe–300 com médias ponderada e simples).

A Tabela 4 mostra os valores de precisão, revocação e F1 obtidos pelo classificador CRF com a combinação WE (Wang2Vec–*Skip-gram*–300 com média ponderada) e AZPort para o *corpus* 466. É possível observar que o classificador mantém o desempenho acima de 85% para todas as categorias exceto Conclusão e Estrutura. No caso da categoria Estrutura, o desempenho foi nulo devido à ausência de sentenças dessa categoria, uma vez que há apenas uma sentença com essa classificação no *corpus* 466. Já no caso da categoria Conclusão, além da sua baixa frequência no *corpus*, existe uma dificuldade por parte do classificador em distinguir as categorias Conclusão e Resultado.

Categoria	**Precisão**	**Revocação**	**F1**
Contexto	99%	100%	100%
Lacuna	100%	100%	100%
Propósito	96%	94%	95%
Método	93%	92%	92%
Resultado	85%	91%	88%
Conclusão	67%	50%	57%
Estrutura	0%	0%	0%
Média	93%	94%	94%

Tabela 4: Precisão, revocação e F1 obtidas pelo classificador CRF usando WE + AZPort sobre o *corpus* 466.

Nas últimas duas linhas da Tabela 3 são apresentadas as médias e desvios padrões dos desempenhos dos algoritmos. Observa-se que o melhor desempenho médio (77%) foi obtido pelo CRF, que também foi o classificador com maior dispersão (desvio padrão igual a 20%). Os algoritmos SVM (RBF e linear) tiveram desempenhos médios iguais (61%), ambos com valores baixos de desvio padrão (10% e 8%, respectivamente).

Ainda na Tabela 3, também é possível observar que os atributos TF-IDF, usados de forma isolada, não trouxeram bons resultados. Além disso, esses atributos não mostraram ter influência no desempenho de algoritmos como SVM-RBF e CRF. Esses dois algoritmos mantiveram o desempenho independentemente da adição dos valores TF-IDF no conjunto de atributos. Ainda, para os classificadores SVM-Linear e B-NB, o desempenho piorou com a adição dos atributos TF-IDF.

A utilização dos atributos do AZPort levaram às melhorias mais significativas em termos dos desempenhos dos classificadores. No caso do CRF, os atributos do AZPort se mostraram fundamentais, já que utilizando apenas esses atributos o CRF atingiu 92% de F1 para o *corpus* 466. Tal valor ficou apenas 2% abaixo do melhores desempenhos observados.

Atributos	**Média**	**Desv.Padrão**
TF-IDF	45%	11%
WE	49%	13%
AZPort	62%	19%
TF-IDF + AZPort	61%	18%
WE + TF-IDF	47%	9%
WE + AZPort	69%	12%
Todos os atributos	64%	17%

Tabela 5: Média e Desvio Padrão dos melhores desempenhos obtidos com cada combinação de atributos.

A Tabela 5 mostra a média e o desvio padrão dos melhores desempenhos obtidos com cada combinação de atributos. Novamente é evidenciada a contribuição dos atributos do AZPort com média superior aos dos outros atributos usados de forma isolada. Quando incluídos no conjunto de atributos, as médias são superiores a 60%, enquanto que, na sua ausência, as médias não alcançam 50%. Vale notar ainda que, embora não tenham tido um bom desempenho médio quando usados de forma isolada, os atributos WE melhoraram o desempenho da classificação quando adicionados aos atributos do AZPort. A maior média de desempenho, 69%, foi obtida por essa combinação.

Para uma melhor visualização dos resultados obtidos por *corpora*, a Figura 2 mostra um gráfico com os maiores valores de F1 obtidos por cada classificador em cada um dos *corpora*. Contrariando o esperado, os resultados não foram melhores no maior *corpus*. No geral, o que gerou os melhores resultados em todos os classificadores foi o *corpus* 466. Já o *corpus* 366 se mostrou o mais difícil para os classificadores. Apenas o G-NB obteve os melhores resultados no *corpus* 366, junto ao 466. Uma análise mais aprofundada dos *corpora* é necessária para identificar as razões que levaram a esses resultados, mas a dificuldade evidenciada para o *corpus* 366 possivelmente influenciou os resultados obtidos para o *corpus* 832, fazendo com que os mesmos ficassem abaixo dos obtidos para o *corpus* 466 apesar do maior número de sentenças.

Com relação aos modelos de *word embeddings*, o gráfico da Figura 3 mostra as maiores porcentagens obtidas com os cinco modelos avaliados (Word2Vec–CBOW, Word2Vec–*Skip-gram*, Wang2Vec–CBOW, Wang2Vec–*Skip-gram* e GloVe). Nesse gráfico, as porcentagens correspondem às maiores medidas F1 obtidas usando-se apenas WE como atributos.

Na Figura 3 é possível notar que o modelo Wang2Vec–*Skip-gram* foi o melhor modelo apenas para o K-NN, junto com Word2Vec–*Skip-gram* e o outro modelo Wang2Vec. Além disso, foi o pior para os classificadores SVM (RBF e linear) e DT e foi o segundo pior para o CRF. Curiosamente, quatro das combinações nas quais o CRF atingiu seu melhor desempenho, 94%, utilizou o modelo Wang2Vec–*Skip-gram*. Outro ponto a ser destacado é que nenhum classificador apresentou melhor desempenho com o modelo GloVe, mas as outras três combinações em que o CRF obteve 94% incluem os atributos extraídos a partir do modelo GloVe.

O gráfico da Figura 4 mostra um comparativo de desempenho para os diferentes tamanhos dimensionais de WE avaliados. Da mesma forma, as porcentagens correspondem às maiores medidas F1 obtidas usando-se apenas WE como atributos. Para o algoritmo K-NN, a variação da dimensão dos vetores de *word embeddings* não causou variações significantes nos resultados. Já para o DT, o aumento da dimensão piorou o desempenho. Para o restante dos classificadores, pode-se dizer que o aumento da dimensão mostrou melhora no desempenho. O G-NB foi o classificador que mostrou ter maior proporção de melhora em comparação aos outros.

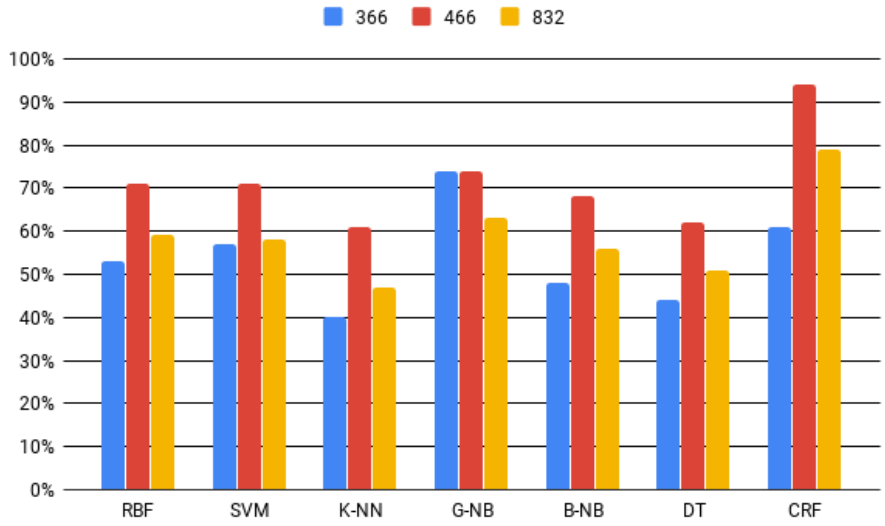

Figura 2: Melhores resultados obtidos para cada *corpus*.

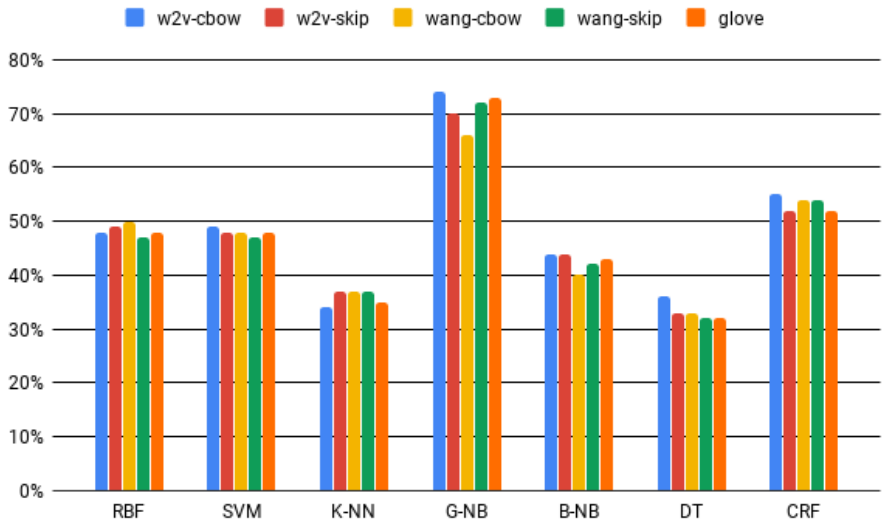

Figura 3: Melhores resultados obtidos em cada modelo de *word embedding*.

O gráfico da Figura 5 mostra um comparativo dos melhores resultados obtidos com as duas formas de representação de *embeddings* para uma sentença (média ponderada pelo IDF e média simples). Nesse gráfico, as porcentagens também correspondem às maiores medidas F1 obtidas usando-se apenas WE como atributos. É possível observar que houve pouca variação no desempenho dos classificadores devido à representação usada. O classificador G-NB se mostrou indiferente quanto à forma de representação, enquanto o restante dos classificadores mostraram resultados levemente superiores usando a média simples. Cabe destacar, no entanto, que quatro das sete combinações em que o CRF obteve 94% usou *embeddings* gerados pela média ponderada.

5 Conclusão

Neste estudo foram avaliados diferentes conjuntos de atributos e algoritmos de classificação aplicados à construção de classificadores retóricos sentenciais para resumos científicos escritos em português. Foram avaliados atributos baseados em TF-IDF, *word embeddings* e os atributos do classificador AZPort, bem como as suas combinações. Com relação aos *embeddings*, foram avaliados os diferentes modelos disponibilizados no repositório NILC–*Embeddings*, bem como duas estratégias de geração de *embeddings* de sentenças a partir de *word embeddings*. As diferentes configurações de atributos foram avaliadas em combinação com sete classificadores distintos, todos eles supervisionados.

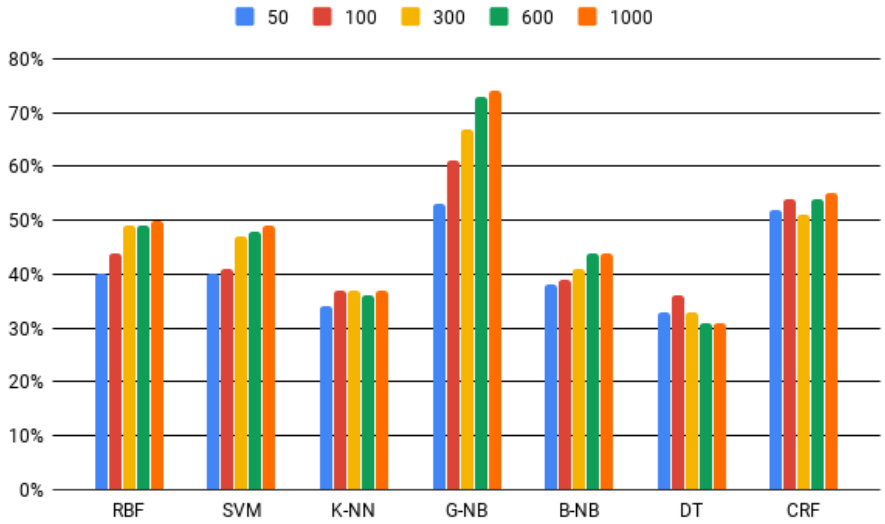

Figura 4: Melhores resultados obtidos em cada dimensão do vetor de *word embedding*.

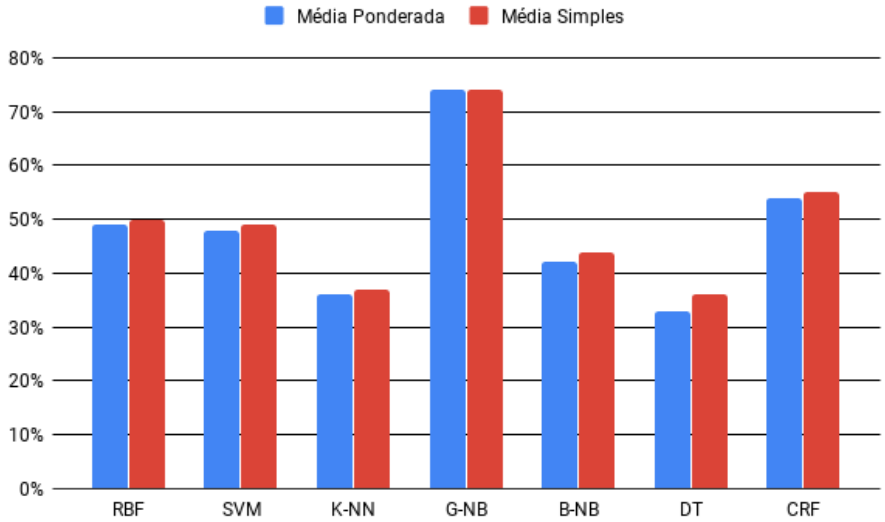

Figura 5: Melhores resultados obtidos em cada forma de representação de *word embedding* para uma sentença.

Dentre os algoritmos avaliados, o que obteve melhor desempenho foi o CRF, confirmando que a classificação retórica é uma tarefa apropriada para algoritmos de rotulação sequencial (Andreani & Feltrim, 2015). No entanto, os resultados se mostraram mais dependentes do conjunto de atributos utilizado.

Observou-se que a utilização de WE, individualmente, não trouxe ganhos significativos nos desempenhos dos classificadores em relação à utilização individual de TF-IDF, com exceção do classificador G-NB, que atingiu o seu melhor desempenho com esses atributos. Já a utilização individual dos atributos do AZPort mostrou desempenho superior ao desempenho com a utilização de TF-IDF e WE em mais de 20% para o CRF e entre 9% a 19% para os classificadores SVM (RBF e linear), K-NN, B-NB e DT. Em especial, o classificador DT obteve seus melhores resultados usando apenas os atributos do AZPort. No caso do classificador CRF, o atributos do AZPort se mostraram os mais efetivos, já que, apenas com eles, o CRF alcançou 92% de F1 para um dos *corpora* usados. Isso mostra que atributos que codificam informações além da superfície do texto, como os do AZPort, são importantes para classificação de estrutura retórica, especialmente quando o conjunto de treinamento é reduzido, como foi o caso deste trabalho.

A combinação de WE com outros atributos mostrou utilidade, uma vez que os classificadores SVM-linear, K-NN e B-NB obtiveram seus melhores resultados com a combinação WE com os atributos do AZPort. Outros classificadores, como SVM-RBF e CRF, obtiveram seus melhores resultados tanto com a combinação de WE com os atributos do AZPort quanto com a combinações de todos os atributos.

Considerando-se os desempenhos obtidos pelos modelos de *word embeddings* quando usados de forma individual, o modelo que obteve o melhor resultado médio foi o Word2Vec–CBOW. Esse resultado está de acordo com o trabalho de Sousa (2016), que também destacou o modelo CBOW como tendo melhor desempenho. Entretanto, o melhor resultado de classificação usando o CRF foi obtido com os modelos Wang2Vec–*Skip-gram* e GloVe. Esse resultado está de acordo com o trabalho de Hartmann et al. (2017) que destacou o desempenho do modelo Wang2Vec. O aumento das dimensões dos modelos de WE trouxe melhorias ao desempenhos dos classificadores, com exceção do K-NN. Essa melhoria ocorreu em maior proporção para o classificador G-NB do que para os outros classificadores.

Para a representação do *embedding* de uma sentença, embora seja comum na literatura realizar a combinação dos *word embeddings* por meio da média ponderada pelo IDF, para este estudo a melhor estratégia foi a combinação pela média simples. Conforme mostrado na Figura 5, a maioria dos classificadores obtiveram melhores resultados com tal estratégia.

Uma vez que os atributos gerados a partir de *word embeddings* contribuíram para a melhoria dos resultados, ainda que de forma discreta, uma das direções para trabalhos futuros é o estudo de outras formas de representação para os *embeddings* das sentenças. Outro ponto para investigações futuras é o treinamento de modelos de *embeddings* com *corpus* de domínio científico, uma vez que os modelos avaliados neste estudo foram treinados com *corpora* de domínios variados. Cabe destacar que, apesar da diferença de domínio, a taxa de palavras não encontradas nos modelos de *word embeddings* variou entre 11% e 13% para os três *corpora* usados no estudo.

Ainda com relação aos atributos, outra vertente de trabalhos futuros é a proposta e a avaliação de atributos que codifiquem informações semânticas, como as fornecidas por analisadores semântico (*Semantic Role Labeling* – SRL).

Agradecimentos

As autoras agradecem aos revisores pelas importantes contribuições a este estudo.

Referências

Andreani, Alexandre C. 2017. Predição estruturada aplicada à detecção de estrutura retórica. Dissertação (Pós Graduação em Ciência da Computação), Universidade Estadual do Paraná, Maringá, Brazil.

Andreani, Alexandre C. & Valéria D. Feltrim. 2015. Campos aleatórios condicionais aplicados à detecção de estrutura retórica em resumos de textos acadêmicos em português (conditional random fields applied to rhetorical structure detection in academic abstracts in portuguese). Em *10th Brazilian Symposium in Information and Human Language Technology (STIL)*, 111–120.

Anthony, Laurence & George V. Lashkia. 2003. Mover: a machine learning tool to assist in the reading and writing of technical papers. *IEEE Transactions on Professional Communication* 46(3). 185–193. 10.1109/TPC.2003.816789.

Booth, Wayne C., Gregory G. Colomb, Joseph M. Wiliams & Henrique A. Rego Monteiro. 2005. *A arte da pesquisa.* São Paulo: Martins Fontes.

Dayrell, Carmen, Arnaldo Candido Jr., Gabriel Lima, Danilo Machado Jr., Ann Copestake, Valéria D. Feltrim, Stella Tagnin & Sandra M. Aluísio. 2012. Rhetorical move detection in english abstracts: Multi-label sentence classifiers and their annotated corpora. Em *8th International Conference on Language Resources and Evaluation (LREC)*, 1604–1609.

Feltrim, Valéria D. 2004. *Uma abordagem baseada em corpus e em sistemas de crítica para a construção de ambientes web de auxilio à escrita acadêmica em português*: Universidade de São Paulo, São Carlos, Brazil. Tese de Doutoramento.

Feltrim, Valéria D., Jorge M. Pelizzoni, Simone Teufel, Maria das Graças Volpe das Nunes & Sandra M. Aluísio. 2004. Applying argumentative zoning in an automatic critiquer of academic writing. Em *Advances in Artificial Intelligence – SBIA 2004*, vol. 3171, 214–223. Springer Berlin Heidelberg. 10.1007/978-3-540-28645-5_22.

Fisas, Beatriz, Francesco Ronzano & Horacio Saggion. 2015. On the discoursive structure of computer graphics research papers. Em *9th Linguistic Annotation Workshop (held in conjuncion with NAACL)*, 42–51. 10.3115/v1/W15-1605.

Guo, Yufan, Anna Korhonen & Thierry Poibeau. 2011. A weakly-supervised approach to argumentative zoning of scientific documents. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 273–283.

Guo, Yufan, Ilona Silins, Ulla Stenius & Anna Korhonen. 2013. Active learning-based information structure analysis of full scientific articles and two applications for biomedical literature review. *Bioinformatics* 29(11). 1440–1447. 10.1093/bioinformatics/btt163.

Hartmann, Nathan, Erick R. Fonseca, Christopher Shulby, Marcos Vinícius Treviso, Jessica Rodrigues & Sandra M. Aluísio. 2017. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *CoRR* arxiv:abs/1708.06025.

Hirohata, Kenji, Naoaki Okazaki, Sophia Ananiadou & Mitsuru Ishizuka. 2008. Identifying sections in scientific abstracts using conditional random fields. Em *3rd International Joint Conference on Natural Language Processing*, 381–388.

Liakata, Maria, Shyamasree Saha, Simon Dobnik, Colin Batchelor & Dietrich Rebholz-Schuhmann. 2012. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics* 28(7). 991–1000. 10.1093/bioinformatics/bts071.

Ling, Wang, Chris Dyer, Alan W. Black & Isabel Trancoso. 2015. Two/too simple adaptations of word2vec for syntax problems. Em *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1299–1304. 10.3115/v1/N15-1142.

Mann, William C. & Sandra A. Thompson. 1987. Rhetorical structure theory: A theory of text organization. Relatório Técnico. ISI/RS-87-190 Information Sciences Institute. 10.1515/text.1.1988.8.3.243.

Mann, William C. & Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text* 8(3). 243–281. 10.1515/text.1.1988.8.3.243.

Merity, Stephen, Tara Murphy & James R. Curran. 2009. Accurate argumentative zoning with maximum entropy models. Em *Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, 19–26.

Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR* arxiv:abs/1301.3781.

Moura, Gustavo Bennemann. 2018. Redes neurais recorrentes para a classificação de estruturas retóricas. Dissertação (Pós Graduação em Ciência da Computação), Universidade Estadual do Paraná, Maringá, Brazil.

Mullen, Tony, Yoko Mizuta & Nigel Collier. 2005. A baseline feature set for learning rhetorical zones using full articles in the biomedical domain. *ACM SIGKDD Explorations Newsletter* 7(1). 52–58. 10.1145/1089815.1089823.

Pendar, Nick & Elena Cotos. 2008. Automatic identification of discourse moves in scientific article introductions. Em *3rd Workshop on Innovative use of NLP for Building Educational Applications*, 62–70.

Pennington, Jeffrey, Richard Socher & Christopher Manning. 2014. Glove: Global vectors for word representation. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. 10.3115/v1/D14-1162.

Reichart, Roi & Anna Korhonen. 2012. Document and corpus level inference for unsupervised and transductive learning of information structure of scientific documents. Em *24th International Conference on Computational Linguistics (COLING)*, 995–1006.

Romeiro, Ana Karoline Queiroz. 2016. *Um estudo sobre o uso da teoria da estrutura retórica (rst) para sumarizar a sabedoria da coletividade*: Universidade Federal Fluminense. Tese de Mestrado.

Siegel, Sidney & N. John Castellan Jr. 1988. *Nonparametric statistics for the behavioral sciences*. Berkeley, CA: McGraw-Hill 2nd edn.

Sousa, Samanta de. 2016. Estudo de modelos de word embedding. Bacharel em Ciência da Computação, Universidade Tecnológica Federal do Paraná, Medianeira, Brazil.

Swales, John M. & Christine B. Feak. 1994. *Academic writing for graduate students: Essential tasks and skills: A course for nonnative speakers of english (English for specific purposes)*. Ann Arbor.

Teufel, Simone & Marc Moens. 2002. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational Linguistics* 28(4). 409–445. doi 10.1162/089120102762671936.

Varga, Andrea, Daniel Preotiuc-Pietro & Fabio Ciravegna. 2012. Unsupervised document zone identification using probabilistic graphical models. Em *8th International Conference on Language Resources and Evaluation (LREC)*, 1610–1617.

Weissberg, Robert & Suzanne Buker. 1990. *Writing up research*. Prentice Hall Englewood Cliffs, NJ.

Yepes, Antonio Jimeno, James Mork & Alan Aronson. 2013. Using the argumentative structure of scientific literature to improve information access. Em *Workshop on Biomedical Natural Language Processing*, 102–110.