Estratégias automáticas para análise da concordância da anotação de Sinalizadores Discursivos

Gabriel Sizinio Bomfim Cruz[1], Jackson W. C. Souza[2], Paula C. F. Cardoso[3]

[1]Instituto de Computação – Universidade Federal da Bahia, Salvador/BA

[2]Programa de Pós-Graduação em Língua e Cultura (PPGLinC) – Universidade Federal da Bahia (UFBA) – Salvador/BA

[3]Faculdade de Computação – Universidade Federal do Pará, Belém/PA

{gsizinio, jackcruzsouza}@gmail.com, pcardoso@ufpa.br

Abstract. *The process of annotating a corpus using Rhetorical Structure Theory (RST) has very clear and defined steps, among which the analysis of agreement between annotators stands out. In this work we present two agreement analysis strategies (gold and silver) based on the Krippendorff Alpha measure. The results point to advanced advances in this type of analysis and the possibility of replication by other works in this segment.*

Resumo. *O processo de anotação de um corpus utilizando a Teoria da estrutura retórica (RST) possui etapas bem claras e definidas, dentre as quais destaca-se a análise da concordância entre os anotadores. Neste trabalho apresentamos duas estratégias de análise da concordância (gold e silver) com base na medida de Krippendorff Alpha. Os resultados apontam significativos avanços para esse tipo de análise e a possibilidade de replicação por outros trabalhos nesse segmento.*

1. Introdução

A *Rhetorical Structure Theory* (RST) [Mann e Thompson, 1987] é uma teoria linguística, amplamente utilizada no Processamento de Linguagem Natural (PLN) para análise de discursos. Tal modelo teórico se propõe a descrever a organização estrutural dos textos com base nas relações retóricas que ocorrem entre suas partes. Essas relações, como *Justify*, *Condition* e *Elaboration*, por exemplo, são capazes de evidenciar interpretações e intenções por meio de como diferentes fragmentos do texto se relacionam entre si. Majoritariamente, essas relações têm sido identificadas por meio de marcas explícitas na superfície textual, como Marcadores discursivos (preposições e conjunções) e outros Sinalizadores discursivos (como pontuação e sentido do verbo).

A identificação dessas marcas é feita predominantemente por meio de anotação de *corpus*. Hovy e Lavid (2010) destacam que esse processo possui etapas bem definidas e, dentre elas, são fundamentais as etapas de análise da anotação e avaliação do nível de concordância entre os anotadores, já que, a depender da concordância obtida, o processo de anotação do *corpus* pode seguir para outras etapas. Ainda, segundo os autores, a baixa concordância entre os anotadores indica que não há consistência no trabalho para permitir que os algoritmos de aprendizado de máquina (AM) sejam treinados a partir desse material. Por outro lado, uma concordância alta indica que o processo pode prosseguir para a anotação de uma maior parte do *corpus* e, consequentemente, o material obtido ser usado para treinamento de sistemas baseados em AM. Nesse cenário, dado que análise da concordância é uma questão central para garantir a qualidade dos dados anotados e o prosseguimento do processo de anotação, é necessário garantir que, em sua automação, os resultados possam ser confiáveis.

Diante disso, nosso objetivo neste trabalho é apresentar estratégias automáticas para a medição e análise da concordância entre humanos na tarefa de anotação de Sinalizadores Discursivos (SDs) para as relações retóricas da RST. Mais especificamente, propomos e comparamos duas abordagens distintas para análise da concordância, uma mais restrita (*Gold*) e outra mais flexível (*Silver*), avaliando o alcance de ambas.

Para tanto, este artigo está organizado em quatro seções, além desta Introdução. Na Seção 2 é apresentada a metodologia aplicada para a medição e análise da concordância entre os anotadores e na Seção 3, os resultados obtidos a partir dela. Por fim, na Seção 4, tecemos considerações finais acerca deste trabalho.

2. Metodologia

Este trabalho se concentra na medição e avaliação da concordância entre anotadores na tarefa de marcar SDs em textos jornalísticos. A anotação foi feita semiautomaticamente por Cardoso *et. al* (2024) em uma amostra do *corpus* CSTNews [Cardoso *et. al* 2011], em que cada texto, já com as relações RST, teve a indicação de SDs por três anotadores diferentes. A anotação foi realizada a partir da ferramenta rstWeb [Zeldes 2016], seguindo as diretrizes propostas por Dantas *et. al* (2024).

A partir da anotação de SDs realizada, propôs-se um *pipeline* de pré-processamento e construção de um algoritmo para analisar a concordância entre os anotadores. Este *pipeline* envolveu a eliminação de inconsistências nas anotações e a ordenação dos *tokens* indicados pelos anotadores.

A análise da concordância da anotação pode ser subdividida em (i) escolha da medida de análise, (ii) identificação dos *tokens* anotados e (iii) comparação entre as anotações, sobretudo a comparação entre as escolhas dos *tokens* feitas pelos anotadores. Uma medida comumente utilizada é a *Cohen Kappa* [McHugh 2012]; porém, no escopo deste trabalho, utilizá-la criaria a limitação de comparar as anotações somente entre dois anotadores. Como a anotação de cada texto foi realizada por três anotadores diferentes, optou-se pela medida *Krippendorff Alpha* [Krippendorff 2011]. Tal medida estatística foi utilizada para avaliar a confiabilidade entre dois ou mais anotadores, apresentando mais robustez ao trabalhar com dados categóricos e intervalares, como é o caso das anotações de SDs. Seu valor varia de -1 a 1, em que valores próximos a 1 indicam alta concordância entre os anotadores, valores próximos a 0 indicam baixa concordância, e valores próximos a -1 indicam discordância sistemática.

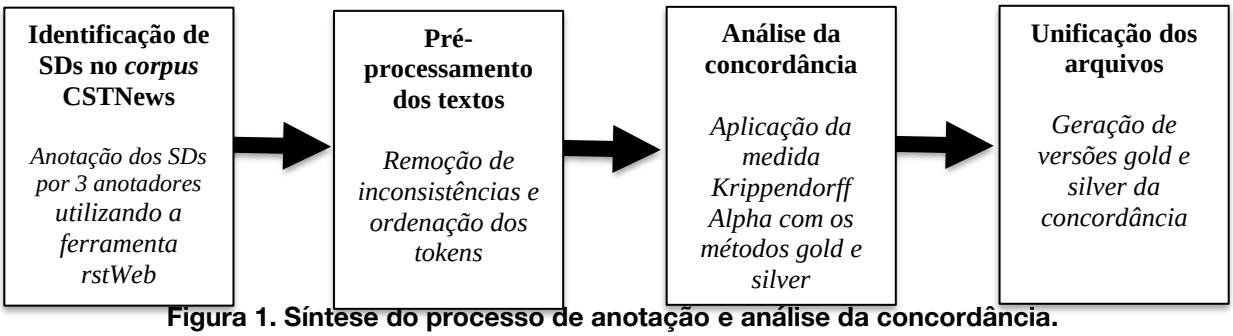

Figura 1. Síntese do processo de anotação e análise da concordância.

Para tanto, implementamos esse cálculo a partir da biblioteca do *Natural Language Toolkit* (NLTK) [Bird, Klein e Loper 2008]. O cálculo foi feito a partir de arquivos .rs3, gerados pela ferramenta rstWeb. Esse arquivo, que é baseado em .xml, está organizado a partir de *tags*, em que cada uma delas se refere a algum aspecto da anotação ou do texto. Nesta pesquisa, as

tags analisadas foram do tipo *signal*, referentes aos SDs indicados por cada um dos sinalizadores em relação ao *Elementary Discourse Unit* (EDU)[1].

3. Resultados

Para análise dos resultados, foram propostas duas análises sobre a concordância. A primeira análise, tida como *gold*, é mais restrita e impõe que os anotadores concordem estritamente na escolha de um *token*, indicando que eles anotaram os mesmos sinalizadores discursivos. Já na segunda, tida como *silver*, é mais flexível, e propõe arbitrariamente a elaboração de uma janela com escopo entre -5 e +5, em relação ao *token*-alvo. Nesta última proposta, os anotadores concordam se um deles anotou um *token* e o outro anotou outro que esteja no espaço da janela de cinco *tokens*, tanto para direita quanto para a esquerda. Na Tabela 1, são apresentados exemplos dessas duas propostas.

Texto	**Tokens anotados**		**Token-alvo**	**Concordância**	
	Anotador A	Anotador B		*Gold*	*Silver*
Eu adoro bolo de chocolate, torta de morango e mousse de limão.	"Eu", "adoro", "chocolate"	"Eu", "adoro", "bolo", "torta", "morango", "mousse"	"chocolate"	"Eu", "adoro"	"Eu", "bolo", "adoro", "torta", "morango"

Tabela 1. Exemplo de comparação entre análises *gold* e *silver*.

A partir da Tabela 1 tem-se que a análise *gold* é mais restrita, uma vez que os anotadores A e B, por exemplo, concordam apenas nos *tokens* "Eu" e "adoro". Já a análise *silver*, por conta da janela de concordância ser mais flexível, em relação ao *token*-alvo "chocolate" (posição 0) foi possível considerar os *tokens* "bolo" (posição -2), "torta" (+2) e "morango" (+4) na análise, pois foram indicados pelo anotador B, e se encontram dentro da janela de concordância em relação à "chocolate". Embora os anotadores não tenham indicado exatamente os mesmos *tokens*, por conta da proximidade, é possível considerar os tokens de maneira mais flexível.

Para operacionalizar a análise, os dados anotados foram organizados de forma a permitir a comparação entre diferentes anotadores. Cada anotação foi processada e armazenada para que a realização da análise da concordância e obtenção do *Krippendorff Alpha* fosse feita.

Na Tabela 2, tem-se a comparação entre os dois métodos, considerando quatro experimentos. Cada texto foi anotado por três anotadores: um mais experiente e outros dois menos experientes. Nos experimentos I e II foram realizados os cálculos do *Krippendorff Alpha* entre anotadores mais e menos experientes separadamente em relação à tarefa de anotação de *corpus*. No Experimento III foi realizado o cálculo entre os anotadores menos experientes. Por fim, no Experimento IV, calculou-se a concordância entre os três anotadores.

Métodos	**Experimentos**			
	I	**II**	**III**	**IV**
Gold	0.477	0.433	0.455	0.455
Silver	0.628	0.680	0.595	0.688

Tabela 2. Exemplo de análise da concordância em um texto do *corpus*.

[1] EDU é a menor unidade de texto que pode ser considerada para análise discursiva. Essas unidades representam segmentos básicos do discurso, como frases ou orações independentes, que contribuem para a estrutura retórica do texto.

Observa-se que a análise *silver* resulta consistentemente em valores de concordância mais altos de *Krippendorff Alpha* em relação à análise *gold*, o que é um reflexo da sua natureza mais flexível da proposta. Além disso, o Experimento IV apresenta os resultados mais elevados, o que sugere algum consenso entre os anotadores.

4. Considerações finais

O processo de avaliação da concordância entre anotadores é um aspecto crítico para garantir a qualidade e a consistência dos dados utilizados no treinamento de modelos de AM. Neste trabalho, optamos por um método diferenciado de análise da concordância, propondo duas abordagens distintas: uma análise mais restrita (*gold*) e outra mais flexível (*silver*). Ainda, nessa pesquisa, procuramos prever quatro cenários distintos, em que fosse possível observar se o perfil de anotadores (com mais ou menos experiência com anotação de corpus) poderia influenciar a concordância.

Os resultados apresentados nas Tabelas 1 e 2 demonstram que a abordagem *silver*, devido à sua flexibilidade, tende a produzir valores de concordância mais elevados em comparação à abordagem *gold*, uma vez que considera uma janela de concordância, permitindo uma avaliação mais inclusiva das anotações. Tal aspecto é especialmente útil em cenários em que pode haver pequenas variações nas escolhas de *tokens*, não comprometendo a qualidade da anotação. A adoção dessas duas abordagens permitirá uma compreensão mais abrangente do alinhamento entre os anotadores, oferecendo uma visão tanto da precisão estrita quanto de uma concordância mais ampla e contextual.

A automatização do processo de análise de concordância, como implementado neste trabalho, representa um avanço significativo, permitindo uma avaliação mais rápida e objetiva da qualidade das anotações. O uso das bibliotecas NLTK e lxml para manipulação e análise dos dados mostrou-se extremamente eficaz, o que possibilitou a criação de *pipelines* de processamento que podem ser reutilizados em outros contextos.

Por fim, as metodologias e ferramentas desenvolvidas neste trabalho tem potencial para servir de base para futuras pesquisas. Dessa maneira, a possibilidade de automatizar a análise da concordância acelera o processo de validação das anotações e garante maior objetividade e confiança dos resultados. Tais aspectos são de suma importância em um processo de anotação de um *corpus* e, consequentemente, o uso dos dados no treinamento de modelos de AM.

Em trabalhos futuros, caberá um aprofundamento nas análises e métodos propostos inicialmente nesta pesquisa. Destaca-se o fato de haver, como demonstrado, impacto no relaxamento da análise por meio do método *silver*, quando comparado ao método *gold*. Ainda, as versões unificadas dos textos anotados poderão ser analisadas no desempenho da função de adjudicator, ao passo que, ao invés de eliminar os trechos em discordância, o método *silver* pode ser utilizado para validar esses mesmos trechos.

Agradecimentos

Este trabalho foi realizado no âmbito do Centro de Inteligência Artificial da Universidade de São Paulo (C4AI -http://c4ai.inova.usp.br/), com o apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (processo FAPESP #2019/07665-4) e da IBM. Este projeto também foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei N. 8.248, de 23 de outubro de 1991, no âmbito do PPI-Softex, coordenado pela Softex e publicado como Residência em TIC 13, DOU 01245.010222/2022-44. Além disso agradecemos ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo financiamento e suporte.

Referências

Bird, S., Klein, E., Loper, E. (2008). *NLTK documentation*. Online. Disponível em: https://www.nltk.org/

Cardoso, P. C., Maziero, E. G., Jorge, M. L. C., Seno, E. M., Di Felippo, A., Rino, L. H. M., ... & Pardo, T. A. (2011). CSTnews-a discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese. In *Proceedings of the 3rd RST Brazilian Meeting* (pp. 88-105). Cuiabá/MT.

Cardoso, P.C.F., Souza, J.W.C., Rodrigues, R. Dantas, E., Cruz, G.S.B., Bárbara, L. de J. S., Gama, N. S., Almeida, T. J. A. Pereira, M.A. 2024. A Linguagem em foco: Anotação de Sinalizadores Discursivos em *corpus* jornalístico. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*. Porto Alegre: SBC.

Dantas, E., Bárbara, L.J.S., Pereira, M.A., Gama, N.S., Almeida, T.J.A., Souza, J.W.C., Cardoso, P.C.F., Rodrigues, R. (2024). *Manual de anotação de sinalizadores discursivos em textos jornalísticos*. São Carlos: Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. Disponível em https://repositorio.usp.br/item/003207370

Hovy, E., Lavid, J. (2010). Towards a Science of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics. *International Journal of Translation*, 22, p.13-36.

Krippendorff, K. (2011). *Computing Krippendorff's Alpha-Reliability*. Departmental paper, Annenberg School for Communication, University of Pennsylvania.

Mann, W.C., Thompson, S. A. (1987). *Rhetorical structure theory: Description and construction of text structures*. In: Natural language generation: New results in artificial intelligence, psychology and linguistics. Dordrecht: Springer Netherlands. p. 85-95.

McHugh, M.L. (2012). *Interrater reliability*: The kappa statistic. Biochemia Medica, 22(3), 276–282

Passonneau R. (2006) Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In: *Proceedings of the international conference on language resources and evaluation* (LREC). Genoa/Italia: European Language Resources Association. p. 831-836.

Zeldes, A. (2016) rstWeb-a browser-based annotation interface for Rhetorical Structure Theory and discourse relations. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations.* p. 1-5.