PRÓ-REITORIA DE PÓS-GRADUAÇÃO E PESQUISA

PROPP - Pesquisa

Dados do Projeto e do(a) Coordenador do Projeto

Título do Projeto	Especificação de Requisitos para o Modelo de Inteligência Artificial da ReINVenTA
Referência da Chamada:	( ) BIC/UFJF e PIBIC/CNPq ( ) PIBIC/CNPq AÇÕES AFIRMATIVAS (X) PROBIC/FAPEMIG ( ) PROBIC-JR/FAPEMIG ( ) Apoio ao Recém-Doutor ( ) Apoio a Grupos de Pesquisa ( ) Apoio à Instalação de Doutores ( ) Cadastro na Propesq
Coordenador do Projeto:	Tiago Timponi Torrent
Equipe:	Natália Sathler Sigiliano, Ely Edison da Silva Matos, Frederico Belcavello
Endereços para contato:	Eletrônico: tiago.torrent@ufjf.br Telefônico: (32) 98803 - 5720
Unidade/Departamento:	Faculdade de Letras / Departamento de Letras
Data:	15/10/2023

1 . Justificativa/Caracterização do Problema

Este projeto caracteriza-se como uma das frentes de atuação da ReINVenTA – Research and Innovation Network for Visual and Textual Analysis of Multimodal Objects –, rede de pesquisa coordenada pela UFJF em parceria com UFMG, UFU e PUC-MG. A ReINVenTA debruça-se sobre o processamento semântico computacional de objetos multimodais (i.e. compostos de modos comunicativos como linguagem verbal, gestos, vídeo, que interagem para a produção de sentido). Para tanto, mobiliza laboratórios e grupos de pesquisa com expertise em Desenvolvimento de Modelos para Compreensão de Língua Natural, Inteligência Artificial, Descoberta de Conhecimento e Tecnologias Assistivas. Mais especificamente, a ReINVenTA investiga aplicações do modelo semântico-computacional da FrameNet à representação semântica de objetos multimodais através da constituição de um dataset semanticamente anotado e do treinamento de modelos de IA neste dataset.

Nesse sentido, parte da premissa de que, assim como palavras podem evocar frames, outros elementos semióticos também o fazem (BELCAVELLO et al., 2020), sendo a representação semântico-computacional de gêneros multimodais o resultado de intrincadas e complexas relações intersemióticas. A FrameNet Brasil vem desenvolvendo, nos últimos anos, os critérios analíticos e o ferramental de software que permitem a extensão do modelo da FrameNet para análise de gêneros multimodais (BELCAVELLO et al., 2022). Na proposta ora apresentada, tais critérios e ferramental serão aplicados à constituição do dataset anotado da ReINVenTA, o qual será posteriormente utilizado para o treinamento de modelos de inteligência artificial multimodais.

Modelos de Inteligência Artificial para a representação semântica de textos e imagens têm sido combinados em tarefas como tradução automática e recuperação de vídeos (AKSOY et al, 2017; BATRA et al., 2018; DELVIN et al., 2015; FANG et al., 2015, NIKOLAUS et al., 2019; SUN et al., 2019). Entretanto, os datasets e os modelos neles treinados apresentam performance restrita quanto à granularidade das representações semânticas, quanto à complexidade das relações estabelecidas pelos elementos componentes das imagens e quanto à complexidade das relações estabelecidas pelas imagens com os textos que as acompanham. Uma alternativa que contempla tal granularidade e complexidade é a FrameNet (BAKER, 2017). Criada por Charles Fillmore em 1997 como uma implementação da Semântica de Frames (FILLMORE, 1982), a FrameNet é um modelo computacional da cognição linguística no qual itens lexicais são modelados em termos dos frames – ou sistemas de conceitos – por eles evocados. O desenvolvimento de um modelo de representação computacional e processamento semântico de gêneros multimodais baseado em frames promove avanços tanto na área de Teoria e Análise Linguística quanto na da Metodologia e Técnicas da Computação, que, em interface, compõem a Linguística Computacional – campo que tem sido um dos mais profícuos da última da última década e que tem levado a um sensível avanço nos modelos linguísticos para Processamento de Língua Natural aplicados a tarefas de impacto cotidiano, como a tradução e a legendagem automáticas, por exemplo. Não obstante os significativos avanços alcançados, referidos modelos linguísticos frequentemente se deparam com os limites impostos pela ambiguidade e indeterminação do significado inerentes a todas as línguas.

Em paralelo, pesquisas em Visão Computacional vêm avançando no uso de modelos de aprendizagem de máquina para rotulação automática de entidades e objetos nos mais diferentes contextos e cenários, o que tem fomentado o desenvolvimento acelerado de aplicações que fazem uso de corpora visuais com etiquetas semânticas, como as utilizadas na geração automática de descrições para imagens e aplicações que se beneficiam da detecção da sincronicidade em ferramentas de edição de vídeo. Contudo, as intrincadas relações de sentido que se estabelecem

entre as modalidades visual e textual ainda passam ao largo dos referidos modelos de visão computacional.

É nesse contexto que esta proposta, ao aliar as duas abordagens à luz de um modelo semântico refinado, curado por humanos busca, no campo teórico, estender o arcabouço analítico da FrameNet de modo a prover modelos de análise da comunicação multimodal com um aparato semântico refinado e cognitivamente plausível e, no campo tecnológico, permitir o desenvolvimento de um conjunto de dados inédito, composto de gêneros multimodais e etiquetas semânticas associadas, fundamentando como aplicação computacional um algoritmo de Inteligência Artificial para análise semântica automática de textos multimodais.

Esse projeto, de caráter inovador, tem alta aderência à ciência básica e às áreas de Inteligência Artificial e Tecnologias Assistivas e seus resultados representam avanços nos estudos da semântica multimodal. Dentre seus produtos, conta-se um corpus multimodal anotado (gold standard dataset), o qual permitirá, em fase posterior prevista no cronograma de desenvolvimento da ReINVenTA, o desenvolvimento de uma ferramenta computacional de identificação de frames em gêneros multimodais.

No que concerne ao dataset, foco principal desta proposta, este é constituído de dois segmentos principais (cf. TORRENT et al., 2022):

– Framed Multi 30k: composto de pareamentos de imagens estáticas e legendas, este segmento é desenvolvido a partir da expansão para o português brasileiro do dataset Flickr 30k (YONG et al., 2014), conforme metodologia definida por Elliot et al. (2016) para o Multi 30k. O Framed Multi 30k, faz uso, ainda de *bounding boxes* desenhadas nas imagens e associadas a sintagmas nominais nas legendas para realização da anotação semântica multimodal. Tais *bounding boxes* e associações foram constituídas no dataset Flickr 30k Entities (PLUMMER et al., 2015). Assim, o ponto de partida para as anotações semânticas a serem realizadas neste segmento do dataset é um conjunto de uma imagem acompanhada de legendas a cujos sintagmas nominais se associam regiões da imagem. A Figura 1 exemplifica tal constituição com legendas em inglês.

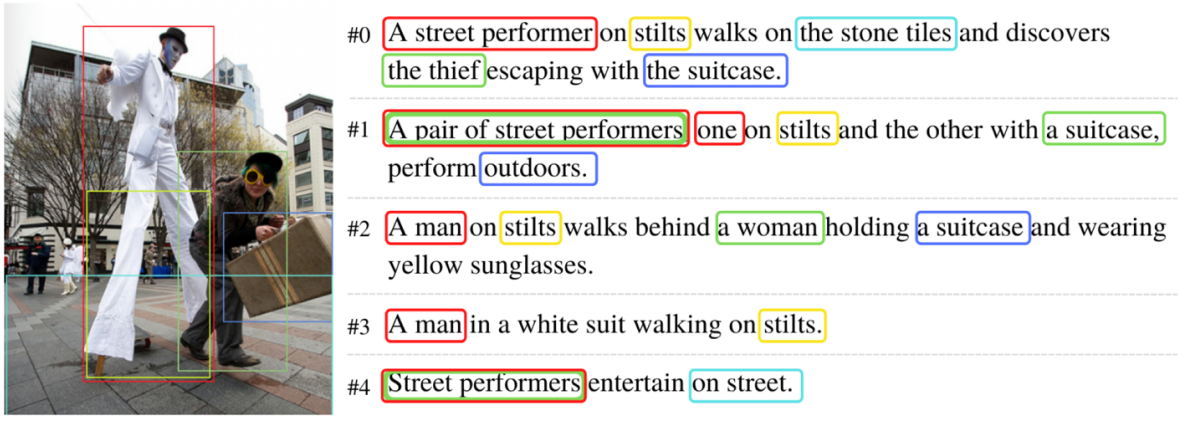

Figura 1: Conjunto de dados a serem anotados no Framed Multi 30k

– Frame^2: composto de vídeos cujas sequências de áudio original e audiodescrição são transcritas. As anotações semânticas são feitas para as modalidades linguística verbal e visual dinâmica, conforme a metodologia definida em Belcavello et al. (2020).

A anotação de ambos os datasets será realizada através da Charon (BELCAVELLO et al., 2022), ferramenta de anotação multimodal desenvolvida pela FrameNet Brasil. Apesar de serem realizadas na mesma ferramenta, as anotações seguem procedimentos próprios, definidos conforme a natureza dos objetos multimodais sendo anotados.

Para o Framed Multi 30k, a interface apresenta ao anotador a imagem para a qual foi criada a legenda, a legenda criada e a possibilidade de selecionar *bounding boxes* para anotação, conforme Figura 2. Alternativamente, pode ser apresentado ao anotador apenas a imagem e as *bounding boxes*, sem a presença da legenda. Tal possibilidade permite investigar o enviesamento da anotação causado pela presença da modalidade comunicativa verbal, conforme relatado por Viridiano et al. (2022).

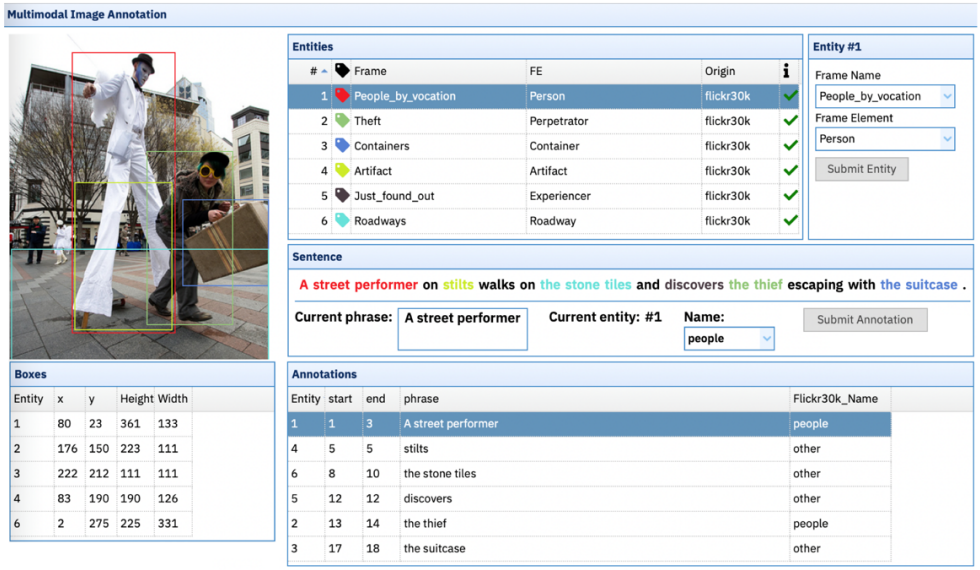

Figura 2: Anotação de pareamentos imagem estática-legenda na Charon

Já para o Frame^2, a interface permite criar *bounding boxes* no vídeo e associá-la a frames e elementos de frame. A Charon inclui, ainda, um algoritmo de visão computacional que pré-processa os vídeos, criando automaticamente uma série de *boxes* para objetos reconhecidos pelo algoritmo na cena. Tais *boxes* podem ser editados, deletados e redesenhados pelo anotador. A interface de anotação de vídeos é mostrada na Figura 3.

A primeira rodada de anotações de ambos os segmentos do dataset está concluída e uma segunda rodada já está em curso. Para o Framed Multi 30k, a primeira rodada de anotação focou na identificação de frames e EFs que representassem as entidades marcadas nas imagens pelas *bounding boxes*. Na etapa atualmente em curso, entre setembro de 2023 e agosto de 2024, as mesmas imagens estão sendo anotadas para frames que representam os eventos nela retratados. Para o Frame^2, a anotação concluída buscou identificar de que maneira as imagens mostradas no vídeo complementavam ou se conjugavam com os frames evocados no áudio. Para a segunda etapa, a anotação está indicando quais frames de evento melhor descrevem cada cena no vídeo.

A partir do dataset anotado, um modelo de Inteligência Artificial (IA) será desenvolvido para a rotulação automática de textos multimodais para frames. A primeira etapa de desenvolvimento de tal modelo é o escopo deste projeto.

O campo do Processamento de Língua Natural (PLN) estrutura-se em tarefas – *tasks* – que propõem casos de uso para os modelos de IA desenvolvidos pelos pesquisadores. Tais tarefas incluem, por exemplo, tradução automática multimodal, rotulagem semântica, extração de informação, dentre outras. Assim, para a construção de qualquer modelo de IA, há que se debruçar sobre a especificação dos requisitos de tal modelo, à luz das tarefas que ele objetiva alcançar. Nesse contexto, o trabalho interdisciplinar é fundamental: de um lado, é necessário considerar a área de aplicação do modelo – no caso, a Linguística –, de outro, é necessária a presença de um especialista em tecnologia da informação que possa traduzir os requisitos da área em algo implementável. É na busca por esse ambiente interdisciplinar que este projeto prevê a concessão de duas bolsas: uma a ser concedida a um discente do curso de Letras e outra a ser concedida a um discentes dos cursos de Ciência da Computação ou Sistemas de Informação.

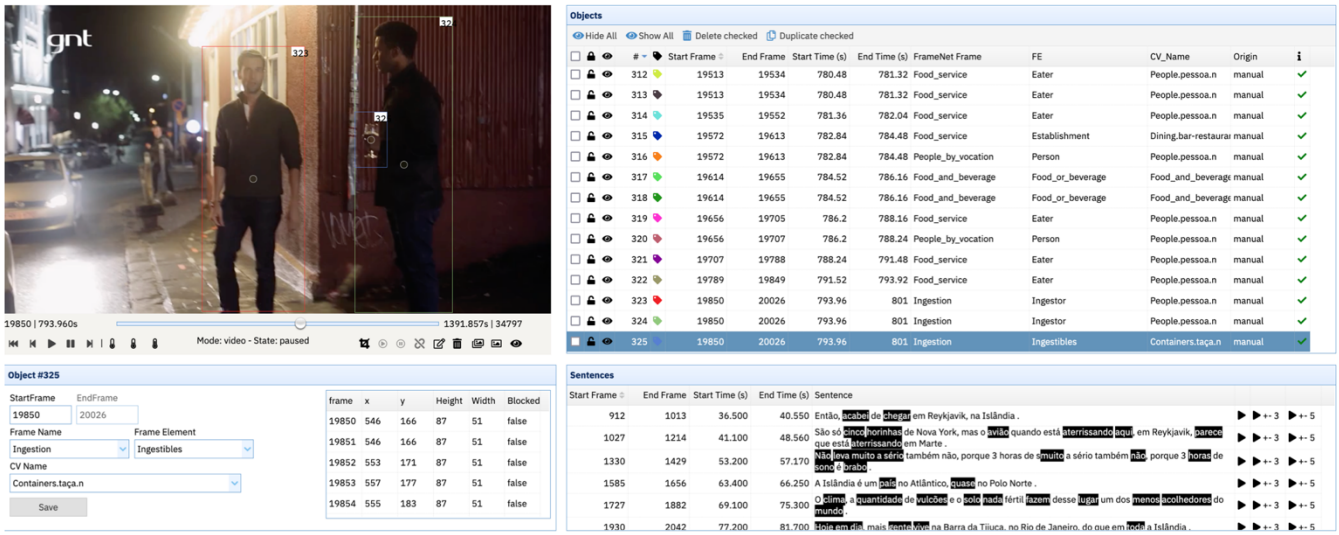

Figura 3: Anotação de vídeos na Charon

2 . Objetivos

Este projeto tem como objetivo principal especificar os requisitos para o modelo de inteligência artificial da ReINVenTA. Para tanto, tem como objetivos correlatos:

- Fazer um levantamento das principais tarefas de PLN multimodal e cotejar sua aplicabilidade ao modelo da Semântica de Frames.
- Definir um conjunto de tarefas computacionais a ser cumprido pelo modelo de IA.
- Produzir o documento de especificação de requisitos do modelo, o qual irá compor o *model card*, uma vez que o modelo esteja implementado.

3 . Metodologia e Estratégias de Ação

Para alcançar os objetivos listados na seção 2, este projeto adotará as seguintes metodologia e estratégias de ação:

- Estudo do dataset multimodal da ReINVenTA: ambos os bolsistas se debruçarão sobre a estrutura de dados e metadados associados ao dataset multimodal da ReINVenTA, porém, com olhares distintos. Enquanto o bolsista de Letras buscará associar as categorias de análise da Linguística à forma como são representadas no dataset, o de TI buscará identificar como tais dados e metadados se estruturam no banco da FrameNet Brasil.
- Levantamento de tarefas de PLN multimodal: os bolsistas conduzirão um levantamento bibliográfico na antologia da Association for Computational Linguistics – ACL – das tarefas de PLN multimodal e de seus requisitos.
- Cotejamento dos dados obtidos no levantamento bibliográfico com o dataset da ReINVenTA: durante esta etapa, o bolsista de Letras irá verificar em que medida as tarefas exploram categorias de análise linguística representáveis no dataset da ReINVenTA, enquanto o bolsista de TI irá compatibilizar os requisitos computacionais das tarefas com a estrutura de dados da FrameNet Brasil.
- Definição das tarefas a serem respondidas pelo modelo de IA: findo o cotejamento, os bolsistas produzirão em conjunto o documento de especificação de requisitos do modelo de IA da ReINVenTA.

4 . Resultados e os impactos esperados

Ao final do projeto, terão sido entregues os seguintes produtos:

- Especificação de requisitos do modelo de IA da ReINVenTA;
- Participações em eventos;
- Publicação de artigo com os resultados finais do projeto.

5 . Cronograma

As atividades previstas para o projeto serão executadas conforme o cronograma a seguir:

ATIVIDADES	MESES											
	01	02	03	04	05	06	07	08	09	10	11	12
Estudo do dataset multimodal	■	■										
Levantamento de tarefas de PLN multimodal			■	■	■							
Cotejamento de dados						■	■	■				
Produção da especificação de requisitos									■	■	■	■
Participação em eventos								■		■		■
Publicação de artigo com os resultados finais do projeto												■

6. Orçamento

O projeto insere-se na ReINVenTA – Research and Innovation Network for Text and Visual Analysis of Multimodal Objects –, a qual recebeu recentemente financiamento da FAPEMIG, no âmbito da Chamada nº 07/2021, da ordem de R$ 1.000.000,00.

7. Referências Bibliográficas

AKSOY, E. E. et al. Unsupervised linking of visual features to textual descriptions in long manipulation activities. *IEEE Robotics and Automation Letters*, v. 2, n. 3, p. 1397-1404, 2017.

BAKER, C. F. FrameNet: Frame Semantic Annotation in Practice. In: IDE, N. & PUSTEJOVSKY, J. *Handbook of Linguistic Annotation*. Dordrecht: Springer, 2017. p. 771-811.

BATRA, V.; HE, Y.; VOGIATZIS, G. Neural caption generation for news images. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC 2018). Miyazaki, 2018.

BELCAVELLO, F.; VIRIDIANO, M.; COSTA, A.; MATOS, E.; TORRENT, T. Frame-Based Annotation of Multimodal Corpora: Tracking (A) Synchronies in Meaning Construction. In: *Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet*. Paris: ELDA, 2020. p. 23-30.

BELCAVELLO, F.; VIRIDIANO, M.; MATOS, E.; TORRENT, T. Charon: a FrameNet Annotation Tool for Multimodal Corpora. In: *Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) @LREC 2022*. Marseille: ELDA, 2022. P. 91–96.

DEVLIN, J. et al. Language Models for Image Captioning: The Quirks and What Works. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Beijing, 2015. p. 100-105.

ELLIOTT, D.; FRANK, S.; SIMA'AN, K.; SPECIA, L. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*. Berlin, Germany: Association for Computational Linguistics, 2016. p. 70–74.

FANG, H. et al. From captions to visual concepts and back. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Boston, 2015. p. 1473-1482.

FILLMORE, C. J. Frame Semantics. In THE LINGUISTIC SOCIETY OF KOREA (org.). *Linguistics in the Morning Calm*. Seoul: Hanshin, 1982, p. 111-137.

NIKOLAUS, M. et al. Compositional Generalization in Image Captioning. In: *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. 2019. p. 87-98.

PLUMMER, B. A.; WANG, L.; CERVANTES, C. M.; CAICEDO, J. C.; HOCKENMAIER, J.; LAZEBNIK, S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-

to-sentence models. In: *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015. p. 2641–2649.

SUN, C. et al. Videobert: A joint model for video and language representation learning. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019. p. 7464-7473.

TORRENT, T. T.; MATOS, E. E. D. S.; BELCAVELLO, F.; VIRIDIANO, M.; GAMONAL, M. A.; COSTA, A. D. D.; MARIM, M. C. Representing context in FrameNet: A multidimensional, multimodal approach. *Frontiers in Psychology*, 13, 2022.

YOUNG, P.; LAI, A.; HODOSH, M.; HOCKENMAIER, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2, 2014. p.67–78.

VIRIDIANO, M.; TORRENT, T. T.; CZULO, O.; LORENZI, A.; MATOS, E.; BELCAVELLO, F. The Case for Perspective in Multimodal Datasets. In: *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*. Marseille, France: ELRA, 2022. p. 108-116