REM: Resolução de Entidades Multimodal

Paulo Henrique Santos Lima[1], **Leonardo Andrade Ribeiro**[2]

[1]Instituto de Informática (INF) – Universidade Federal de Goiás (UFG)
Goiânia, GO – Brasil

pauloh@discente.ufg.br, laribeiro@inf.ufg.br

Abstract. *The task of Entity Resolution (ER) consists of identifying records that refer to the same real-world entity. While traditional approaches focus on textual data, the growth of multimodal sources demands solutions capable of handling this diversity. This paper proposes an approach for Multimodal Entity Resolution (MER), in which each record is composed of both textual and visual information. The architecture is based on the CLIP model, extended with multimodal fusion mechanisms and loss functions inspired by strategies from Multimodal Entity Linking (MEL). Preliminary results on two public datasets delivered F1-scores of up to 93.6%, indicating that the proposed approach is promising.*

Resumo. *A tarefa de Resolução de Entidades (RE) consiste em identificar registros que se referem à mesma entidade do mundo real. Embora abordagens tradicionais se concentrem em dados textuais, o crescimento de fontes multimodais demanda soluções que lidem com essa diversidade. Este artigo propõe uma abordagem para Resolução de Entidades Multimodal (REM), em que cada registro é composto por informações textuais e visuais. A arquitetura é baseada no modelo CLIP, adaptada com mecanismos de fusão multimodal e funções de perda inspiradas em estratégias de Vinculação de Entidades Multimodal (VEM). Resultados preliminares em dois conjuntos de dados públicos resultaram em F1-score de até 93.6%, indicando que a abordagem proposta é promissora.*

Nível: Mestrado
Estudante: Paulo Henrique Santos Lima
Orientador: Prof. Dr. Leonardo Andrade Ribeiro
Programa: Programa de Pós-Graduação em Ciência da Computação (PPGCC/INF/UFG)
E-mails: pauloh@discente.ufg.br, laribeiro@inf.ufg.br
Data de ingresso no programa: Março de 2024
Previsão de defesa: Março de 2026
Etapas concluídas: Qualificação realizada em Julho de 2025
Publicações relacionadas: ICEIS 2025 [Santana et al. 2025].

1. Introdução

A Resolução de Entidades (RE) é o processo de identificar registros que se referem à mesma entidade do mundo real, estejam eles presentes em um único conjunto de dados ou distribuídos entre múltiplas fontes. Trata-se de um dos principais desafios na integração de dados [Caldeira and Ferreira 2018], e possui um longo histórico de pesquisa, com os primeiros estudos iniciados no final da década de 1950 [Newcombe et al. 1959].

Tradicionalmente, as abordagens de RE focam em dados textuais e estruturados, comumente presentes em tabelas de bancos de dados relacionais [Mudgal et al. 2018]. No entanto, o cenário atual é marcado pela crescente presença de informações multimodais, frequentemente não estruturadas e marcadas por elevado grau de heterogeneidade, envolvendo texto, imagem e áudio [Liu et al. 2024]. Neste cenário, a simples concatenação de atributos textuais pode não ser suficiente para distinguir entidades distintas, sobretudo quando os dados textuais são ambíguos, incompletos ou ruidosos. As imagens, por sua vez, oferecem um contexto visual complementar que pode ser crucial para a desambiguação de entidades. Por exemplo, uma imagem associada a uma pessoa ou objeto pode ajudar a diferenciar registros com nomes ou descrições semelhantes [Song et al. 2024, Liu et al. 2024]. Entretanto, essa oportunidade impõe um desafio significativo para a integração e o gerenciamento de dados heterogêneos em banco de dados.

Para superar esse desafio e aproveitar o potencial das informações multimodais, este artigo foca na tarefa de Resolução de Entidades Multimodais (REM) entre registros contendo informações textuais e visuais. Define-se REM como o processo de identificar registros que se referem à mesma entidade do mundo real, onde cada registro contém informações multimodais. Em cenários multimodais, cada registro é tratado como uma entidade, sendo comparado a outros para identificar os que representam a mesma entidade do mundo real. Exemplos ilustrativos da aplicação de REM são apresentados na Figura 1, que mostra dois cenários típicos: (a) retrato falado de suspeitos sendo comparados com imagem, e (b) exames de imagem médica, como raios X, sendo comparados com laudos clínicos. Outras aplicações incluem o reconhecimento de produtos em lojas virtuais, combinando fotos e descrições para detectar o mesmo produto; a correspondência de perfis em redes sociais distintas, usando imagem de perfil e biografia; e a identificação de registros históricos multimodais relacionados a um mesmo artefato ou figura histórica, mesmo que dispersos em diferentes acervos digitais. Esses casos mostram que a comparação entre modalidades distintas é essencial em muitos cenários, exigindo abordagens de REM que explorem as particularidades de cada tipo de dado.

Com o avanço dos modelos de inteligência artificial, técnicas como o *Contrastive Language-Image Pre-training* (CLIP) [Radford et al. 2021], têm se mostrado promissor para tarefas de Vinculação de Entidades Multimodal (VEM) [Song et al. 2024, Liu et al. 2024]. O CLIP utiliza transformadores como dois codificadores independentes — um para imagem e outro para texto — treinados para aprender a associação correta entre representações visuais e textuais [Radford et al. 2021]. Modelos aplicados à tarefa de VEM exploram esse alinhamento entre representações visuais e textuais para desambiguar menções e vinculá-las corretamente a entidades em bases de conhecimento. Inspirada por essas abordagens, a arquitetura proposta neste trabalho adapta o modelo CLIP para o contexto de resolução de entidades entre registros multimodais, incorporando mecanismos de fusão e funções de perda apropriadas para a tarefa de REM.

(a) Identificação de suspeito composto por descrição textual e imagem

(b) Exame médico composto por laudo médico e raio-X

Figura 1. Exemplos ilustrativos de registros multimodais para a tarefa de REM.

Embora REM e VEM sejam conceitualmente distintas, ambas lidam com representações heterogêneas em modalidades como texto e imagem. A VEM foca em associar uma menção ambígua a uma entidade em uma base de conhecimento [Chen and Zhang 2024], enquanto a REM busca determinar se dois registros multimodais representam a mesma entidade, sem depender dessa base. Essa proximidade permite adaptar técnicas de VEM para REM, como proposto neste trabalho.

2. Fundamentação Teórica e Trabalhos relacionados

2.1. Fundamentação Teórica

A tarefa de RE tem sido amplamente estudada nas últimas décadas por diversas comunidades científicas. Áreas como Banco de Dados, Recuperação de Informação, Processamento de Linguagem Natural, Aprendizado de Máquina, Web Semântica e Estatística abordam aspectos desse desafio, evidenciando sua importância na integração e análise de dados. Nessas comunidades, o problema de ER é referenciado por uma variedade de termos, incluindo casamento de entidade, casamento de registro e deduplicação. Uma revisão da literatura anterior ao surgimento de técnicas baseadas em *Deep Learning* é apresentada por [Elmagarmid et al. 2007]. Já uma revisão mais recente, com foco em técnicas baseadas em DL, é apresentada por [Barlaug and Gulla 2021].

Com o aumento da disponibilidade de dados multimodais, cresce o interesse em integrar texto e imagem para a desambiguação de entidades. Nesse contexto, surgem abordagens que exploram o potencial de diferentes modalidades de dados, dando origem ao campo de VEM [Sun et al. 2022]. Diferentemente da vinculação de entidades tradicional — que visa vincular menções ambíguas a entidades não ambíguas em uma base de conhecimento — VEM incorpora informações multimodais, geralmente combinando dados textuais e visuais, para realizar o vínculo com entidades em uma base de conhecimento textual ou multimodal [Chen and Zhang 2024].

Com o avanço das técnicas de aprendizado contrastivo, modelos como o CLIP [Radford et al. 2021] ganharam destaque no cenário multimodal. Essa técnica aproxima representações de pares semelhantes e afasta as de pares distintos em um espaço vetorial

compartilhado [Radford et al. 2021]. Em tarefas multimodais, como texto e imagem, o modelo é treinado para alinhar semanticamente elementos correspondentes, como uma imagem de um ”cachorro no jardim”e sua legenda, afastando descrições irrelevantes, como ”helicóptero decolando”. O CLIP implementa esse princípio com codificadores que projetam as representações em um espaço de *embedding* multimodal compartilhado. Esse espaço permite o alinhamento semântico entre texto e imagem, sendo adotado como base em tarefas multimodais, incluindo VEM.

2.2. Trabalhos Relacionados

Dentre os principais avanços de técnicas baseadas em DL, destaca-se o *DeepMatcher* [Mudgal et al. 2018], que aplica redes profundas em atributos textuais para tarefas RE. Posteriormente, o *Ditto* [Li et al. 2023] introduziu uma arquitetura baseada em *Transformers*, alcançando desempenho do estado da arte em múltiplos *benchmarks*. Trabalhos recentes buscaram aliar eficácia com eficiência computacional [Santana et al. 2025]. No entanto, tais propostas operam exclusivamente sobre dados textuais e ainda demonstram limitações em cenários com atributos incompletos ou ruidosos [Lima et al. 2023].

Trabalhos recentes na área de VEM têm explorado novas arquiteturas e o uso de modelos pré-treinados em larga escala. O *Dual-Way Enhanced Framework* (DWE) [Song et al. 2024] utiliza codificadores como o CLIP para extrair características visuais e textuais. A proposta se destaca por aproveitar atributos visuais de alto nível, como feições faciais e elementos de cena, com o objetivo de enriquecer as representações visuais e promover uma fusão mais eficaz entre modalidade. Além disso, o DWE aplica estratégias de aprimoramento *cross-modal* e alinhamento tanto em nível de atributo quanto de sentença, buscando superar as lacunas semânticas entre as modalidades textual e visual.

Outra abordagem relevante é o UniMEL [Liu et al. 2024], que propõe um arcabouço unificado baseado em *Multimodal Large Language Model*s (MLLMs) para a tarefa de VEM. O modelo processa simultaneamente a imagem e o texto associado à menção, extraindo representações multimodais ricas e semanticamente alinhadas. Para representar as entidades, o UniMEL utiliza *Large Language Models* (LLMs) que integram informações da base de conhecimento. Por fim, a correspondência entre menção e entidade é refinada com re-ranking e seleção final realizada por um LLM com *fine-tune* sobre um conjunto candidato reduzido.

Embora abordagens como o UniMEL explorem o uso de LLMs em VEM, desafios importantes ainda persistem. A aplicação direta de LLMs em tarefas como *blocking* — etapa essencial para reduzir o espaço de busca em RE — permanece limitada, devido ao alto custo computacional envolvido na geração de pares candidatos em larga escala [Freire et al. 2025]. Além disso, o uso de entidades com atributos longos ou ruidosos pode comprometer a eficácia dos LLMs, exigindo estratégias de pré-processamento e seleção de atributos mais informativos para compor os *prompts*. Esses desafios evidenciam a necessidade de arquiteturas que explorem de forma mais eficiente a complementaridade da informações heterogêneas na resolução de entidades multimodais.

3. Arquitetura Proposta

A arquitetura adotada neste trabalho é uma extensão do modelo CLIP com estratégias de fusão e aprendizado inspiradas no DWE [Song et al. 2024], originalmente concebido para tarefas de VEM. Essa combinação visa adaptar ambas as abordagens para REM.

Inicialmente, representações globais e locais são extraídas dos codificadores pré-treinados do CLIP para texto e imagem. Essas representações são então alinhadas por camadas de projeção lineares, que ajustam seus espaços vetoriais em uma dimensão comum. Em seguida, os tokens textuais passam por um bloco de atenção cruzada (*TextFusionBlock)*, onde eles são atualizados com base nos tokens visuais, permitindo que o texto aprenda com a imagem. Em seguida, é utilizado um módulo de atenção multinível (MultiLevelAttention), que combina interações globais, locais e detalhadas (*fine-grained*) entre imagem e texto.

Todas as representações geradas são então concatenadas e encaminhadas para um classificador responsável pela predição final. Durante o treinamento, adota-se uma estratégia de congelamento progressivo dos parâmetros do CLIP, descongelando-os gradualmente após as primeiras épocas. Além da função de perda binária (*BCEWithLogitsLoss*), são incorporadas perdas auxiliares inspiradas no DWE [Song et al. 2024], que orientam o alinhamento e a separação semântica dos *embeddings*, incluindo *CLIPLoss*, *TripletLoss*, *ContrastiveLoss* e *CircleLoss*.

4. Experimentos e Resultados

4.1. Conjuntos de Dados

Os conjuntos de dados utilizados foram o RichpediaMEL e o WikiDiverse [Zhou et al. 2021], usados em trabalhos como DWE e UniMEL. Ambos fornecem menções e entidades do mundo real com descrições textuais e imagens associadas, permitindo a avaliação em contextos multimodais. Para cada conjunto, foi extraído um subconjunto com os atributos relevantes para a tarefa, incluindo o nome da entidade, descrição textual, tipo de instância e a imagem correspondente. Cada registro foi dividido em dois componentes: um textual (nome, descrição e tipo da entidade) e outro visual (imagem associada à entidade). Esses componentes foram alocados em conjuntos separados para a geração de pares positivos (entidade e imagem emparelhadas corretamente) e negativos (incorretamente) para classificação binária supervisionada. Por fim, o conjunto foi dividido de forma estratificada em 60%, 20% e 20% para treino, validação e teste, respectivamente, mantendo equilíbrio entre exemplos positivos e negativos. A Tabela 1 apresenta a distribuição dos pares nos conjuntos.

Dataset	Partição	Pares	Pares Positivos	Pares Negativos
RichpediaMEL 173.538	Treinamento	104.122	52.061	52.061
	Validação	34.706	17.353	17.353
	Teste	34.710	17.355	17.355
WikiDiverse 117.466	Treinamento	70.478	35.239	35.239
	Validação	23.492	11.746	11.746
	Teste	23.496	11.748	11.748

Tabela 1. Estatísticas dos pares gerados para cada conjunto de dados.

4.2. Resultados

Diferente de trabalhos focados em VEM como o DWE [Song et al. 2024] e UniMEL [Liu et al. 2024], que adotam métricas top-k por operarem em cenários de recuperação com conjunto candidato extenso, este trabalho utiliza o *F1-Score* como principal métrica. A escolha se justifica pela natureza da tarefa de REM, formulada como classificação

binária entre pares de registros. Nesse contexto, o objetivo não é ranquear candidatos em relação a uma menção, mas determinar se dois registros se referem à mesma entidade. O desempenho foi acompanhado por 40 épocas, com possibilidade de interrupção antecipada caso não houvesse melhorias no conjunto de validação por 10 épocas consecutivas.

Como ilustrado na Figura 2, o modelo apresentou curva de aprendizado consistente, com ganhos rápidos nas primeiras épocas e posterior estabilização. No Richpedia, o F1-score de validação atingiu pico na época 7 (0.935), mantendo-se estável até a época 17. No teste, o desempenho final foi de 0.936, evidenciando boa generalização. Para o Wikidiverse, observou-se evolução semelhante. O F1-score de validação cresceu até a época 11 (0.931), com leve oscilação nos ciclos seguintes. O modelo alcançou 0.931 no conjunto de teste, indicando robustez diante de dados com maior variabilidade semântica.

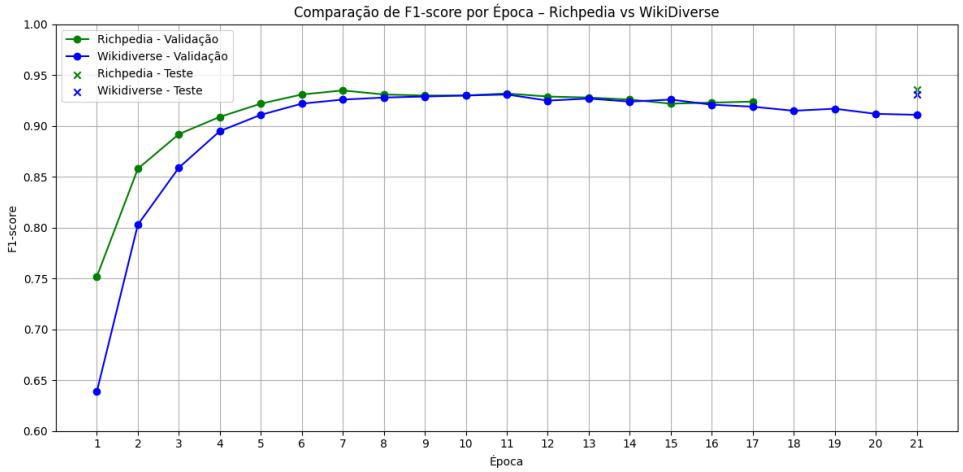

Figura 2. Evolução F1-score por Época – Comparativo Richpedia e WikiDiverse

5. Conclusão

Este trabalho apresentou uma abordagem para REM, explorando a complementaridade entre informações textuais e visuais por meio da arquitetura CLIP estendida com mecanismos de fusão e funções de perda multimodais. Inspirado por estratégias de VEM, o modelo proposto adapta tais técnicas ao cenário específico de correspondência binária entre registros multimodais, típicos de tarefas de REM. A metodologia foi avaliada nos conjuntos RichpediaMEL e Wikidiverse, resultando em F1-scores de 0.936 e 0.931 nos testes, respectivamente. Indicando que a combinação de representações visuais e textuais, aliada ao uso de perdas auxiliares e atenção cruzada, contribui para a desambiguação de entidades em contextos heterogêneos.

Como trabalhos futuros, pretende-se explorar mecanismos de *blocking* para redução do espaço de busca e investigar a integração de *embeddings* de linguagem multimodal oriundos de modelos generativos como LLMs. Além disso, pretende-se estudar estratégias inspiradas em abordagens como o Ditto [Li et al. 2023], que incorpora conhecimento de domínio e aumento de dados, a fim de incorporar características semânticas mais especializadas e aumentar a representatividade dos dados de treinamento.

Agradecimentos Esta pesquisa foi apoiada pela CAPES e LaMCAD/UFG.

Referências

Barlaug, N. and Gulla, J. A. (2021). Neural Networks for Entity Matching: A Survey. *ACM Trans. Knowl. Discov. Data*, 15(3):52:1–52:37.

Caldeira, L. and Ferreira, A. (2018). Melhorias no Processo de Blocagem para Resolução de Entidades Baseadas na Relevância dos Termos. In *Proceedings of the Brazilian Symposium on Databases*, pages 61–72.

Chen, D. and Zhang, R. (2024). Building Multimodal Knowledge Bases With Multimodal Computational Sequences and Generative Adversarial Networks. *Trans. Multi.*, 26:2027–2040.

Elmagarmid, A. K., Ipeirotis, P. G., and Verykios, V. S. (2007). Duplicate Record Detection: A Survey. *IEEE Trans. Knowl. Data Eng.*, 19(1):1–16.

Freire, J., Fan, G., Feuer, B., Koutras, C., Liu, Y., Peña, E., Santos, A. S. R., Silva, C. T., and Wu, E. (2025). Large Language Models for Data Discovery and Integration: Challenges and Opportunities. *IEEE Data Engineering Bulletin*, 49(1):3–31.

Li, Y., Li, J., Suhara, Y., Doan, A., and Tan, W. (2023). Effective Entity Matching with Transformers. *VLDB Journal*, 32(6):1215–1235.

Lima, P. H. S., Santana, D. R., Martins, W. S., and Ribeiro, L. A. (2023). Evaluation of Deep Learning Techniques for Entity Matching. In *International Conference on Enterprise Information Systems*, pages 247–254.

Liu, Q., He, Y., Xu, T., Lian, D., Liu, C., Zheng, Z., and Chen, E. (2024). UniMEL: A Unified Framework for Multimodal Entity Linking with Large Language Models. In *Proceedings of CIKM*, pages 1909–1919.

Mudgal, S., Li, H., Rekatsinas, T., Doan, A., Park, Y., Krishnan, G., Deep, R., Arcaute, E., and Raghavendra, V. (2018). Deep Learning for Entity Matching: A Design Space Exploration. In *Proceedings of the SIGMOD Conference*, pages 19–34. ACM.

Newcombe, H. B., Kennedy, J. M., Axford, S. J., and James, A. P. (1959). Automatic Linkage of Vital Records. *Science*, 130(3381):954–959.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *CoRR*, abs/2103.00020.

Santana, D. R., Lima, P., and Ribeiro, L. (2025). EM-Join: Efficient Entity Matching Using Embedding-Based Similarity Join. In *International Conference on Enterprise Information Systems*, pages 402–409.

Song, S., Zhao, S., Wang, C., Yan, T., Li, S., Mao, X., and Wang, M. (2024). A Dual-Way Enhanced Framework from Text Matching Point of View for Multimodal Entity Linking. In *Proceedings of AAAI*, pages 19008–19016.

Sun, W., Fan, Y., Guo, J., Zhang, R., and Cheng, X. (2022). Visual Named Entity Linking: A New Dataset and A Baseline. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of EMNLP*, pages 2403–2415.

Zhou, X., Wang, P., Li, G., Xie, J., and Wu, J. (2021). Weibo-MEL, Wikidata-MEL and Richpedia-MEL: Multimodal Entity Linking Benchmark Datasets. pages 315–320.