Extração de Eventos em Notas Clínicas

João Augusto F. Balducci[1], **Saullo H. G. de Oliveira**[1]

[1]Escola Politécnica - Pontifícia Universidade Católica de Campinas

Abstract. *Event Extraction (EE) is the task of identifying and extracting event information from free text. Due to the large number of unstructured text sources, the healthcare sector can benefit from EE to facilitate the interpretation of health records and mitigate medical errors. We therefore propose EEVIN (Extrator de EVentos clÍNicos) an algorithm for the ordered event extraction problem on clinical texts written in Brazilian Portuguese. The solution was compared with state-of-the-art Large Language Models (LLMs) and obtained significant results while presenting reduced computational costs.*

Resumo. *Extração de Eventos (EE) é a tarefa de identificar e extrair informações de eventos em um texto livre. Devido à grande quantidade de fontes de texto não estruturado, a área da saúde pode se beneficiar da EE para facilitar a interpretação de registros clínicos e mitigar erros médicos. Neste trabalho apresentamos o EEVIN (Extrator de EVentos clÍNicos), um algoritmo para a extração de eventos ordenados cronologicamente a partir de textos clínicos escritos em português brasileiro. A solução foi comparada com Large Language Models (LLMs) do estado da arte e obteve resultados significativos com custo computacional mais baixo.*

1. Introdução

Extração de Eventos (EE) é a tarefa de identificar eventos num texto livre. Um evento pode ser definido como uma "mudança de estado" [Li et al. 2024] e geralmente é caracterizado pela presença de um gatilho, argumentos e extensão, ou seja, palavras que anunciam o evento, palavras que interagem com o evento e a parte do texto que o contém. Espera-se então extrair informações sobre os eventos, como por exemplo, o momento no qual o evento aconteceu. Na área médica, a EE pode ser muito útil se aplicada a registros eletrônicos de saúde, uma vez que cerca de 80% dos dados na área da saúde estão em fontes não estruturadas [Juhn and Liu 2020]. Considerando que os prontuários: i) são escritos em texto livre e corrido; ii) sequencialmente no tempo; e iii) não há destaque para acontecimentos específicos [Perera et al. 2013], mas sim uma descrição de fatos relevantes sobre o paciente; é possível desenhar algoritmos para a extração de eventos em prontuários capazes de auxiliar profissionais da saúde em sua tomada de decisão o que melhora a qualidade do serviço prestado, podendo diminuir o uso de insumos [Benício 2020].

Liu et al. 2020 aborda o problema treinando um modelo baseado em BERT que responde perguntas buscando extrair os argumentos dos eventos, enquanto Chen et al. 2015 usa redes neurais convolucionais para identificar gatilhos e classificar argumentos. Ambos os métodos são dedicados a textos escritos em inglês e foram treinados utilizando bases de dados anotadas para EE, como ACE-2005 [Walker, Christopher et al. 2006] e MIMIC-III [Johnson et al. 2016]. Contudo, não foram encontradas propostas dedicadas ao português brasileiro. Neste trabalho, apresentamos o EEVIN (Extrator de EVentos clÍNicos), um algoritmo para a extração de eventos

em textos de prontuários médicos, em português brasileiro, que os organiza em uma linha do tempo de eventos relacionados ao paciente. O EEVIN se baseia em características intrínsecas ao texto clínico e combina técnicas de *Pattern Matching* e Reconhecimento de Entidades Nomeadas num sistema baseado em regras para a extração de eventos.

2. Metodologia

2.1. Base de Dados

Utilizaremos os registros eletrônicos de saúde da base de dados SemClinBr [Oliveira et al. 2022], um corpus semanticamente anotado para NER (do inglês, *Named Entity Recognition*) clínico em português, contendo 1.000 notas clínicas rotuladas, totalizando 65.117 entidades e 11.263 relações anotadas. Diferentemente da língua inglesa, que contém corpus anotados para diversas tarefas e com uma grande quantidade de dados, como [Johnson et al. 2016, Walker, Christopher et al. 2006], o SemClinBR é o único corpus de notas clínicas publicamente disponível em língua portuguesa. Para investigar o desempenho dos métodos para a tarefa de extração de eventos, foram anotadas 50 entradas da SemClinBR, utilizando o seguinte protocolo: i) marcadores temporais no início de frase indicam o gatilho para um novo evento; ii) todo o texto que segue uma marcação temporal até a próxima é considerado extensão do evento; e por fim, iii) cada entidade nomeada biomédica mencionada na extensão de um evento é um argumento do evento.

2.2. Proposta: EEVIN

A elaboração dos textos de prontuário clínico possui características úteis para a modelagem do problema de extração de eventos. Assumindo que: i) cada entrada de um prontuário clínico descreve as últimas atualizações sobre um paciente; ii) as entradas são escritas sequencialmente na ordem em que os acontecimentos narrados aconteceram; e iii) cada nova entrada inicia com uma marcação temporal, elaboramos o EEVIN, descrito no Alg. 1.

A função *separarPorGatilhos* identifica os gatilhos dos eventos no texto, juntamente com suas respectivas extensões. Por se tratar de uma narrativa sequencial, os gatilhos marcam o início de um novo registro no texto, e as marcações temporais que iniciam um evento são identificadas com técnicas de *Pattern Matching*. Em seguida, os argumentos do evento são identificados e classificados na função *aplicarNER*. Nesta etapa, o conjunto de modelos de reconhecimento de entidades biomédicas clinicalnerpt [Schneider et al. 2020][1], capaz de classificar as mesmas 13 entidades anotadas na base SemClinBR, é aplicado à extensão do evento. O processo retorna uma lista ordenada de eventos (listaEventos) contendo gatilho, extensão e argumentos, para cada evento presente no texto (*ehr*).

2.3. Experimento

Devido à capacidade dos grandes modelos de linguagem de realizar diversas tarefas sem treinamento especializado (*in-Context learning*), selecionamos alguns modelos atualmente relevantes para comparar o desempenho do EEVIN. São eles: Llama3.3 [Grattafiori et al. 2024], DeepSeek-R1 [DeepSeek-AI et al. 2025] e Qwen2.5 [Yang et al. 2024].

[1]Disponíveis em: `https://huggingface.co/pucpr`

Algoritmo 1 EEVIN

Require: ehr
  $(gatilhos, extensoes) \leftarrow separarPorGatilhos(ehr)$
  $listaEventos \leftarrow []$
  for $gatilho \in gatilhos, extensao \in extensoes$ **do**
    $evento.gatilho \leftarrow gatilho$
    $evento.extensao \leftarrow extensao$
    $evento.argumentos \leftarrow aplicarNER(evento)$
    $listaEventos.add(evento)$
  end for

Tabela 1. Quantidade de eventos identificados por método

	Eventos Identificados
Referência	136
EEVIN	**135**
DeepSeekR1-70B	**137**
Llama3.3-70B	178
Qwen2.5-72B	201

Em experimentos iniciais, foi constatado o grande impacto de diferentes prompts nos resultados parciais de cada modelo, o que inclui o formato de resposta esperado, a maneira pela qual tarefa é explicada e instruções pontuais. O melhor *prompt* [2] foi utilizado em todos os modelos, instruíndo-os a analisar o prontuário e extrair uma linha do tempo de eventos relacionada ao paciente. Apresentamos a composição de um evento como: i) **gatilho:** marcação temporal que determina o início do evento (data, hora, etc.); ii) **argumentos:** entidades que são parte do evento, sendo classificadas em: sintoma, procedimento, medicamento, transtorno, doença, descoberta, atividade de assistência à saúde e resultado de laboratório; e iii) **extensão:** fragmento do texto original que inclui o gatilho e os argumentos, descrevendo todo o evento. Além disso, buscando manter a padronização entre as respostas dos modelos, é solicitado um resultado estruturado como objeto JSON, com instruções exatas sobre a estrutura e os campos.

3. Resultados e discussão

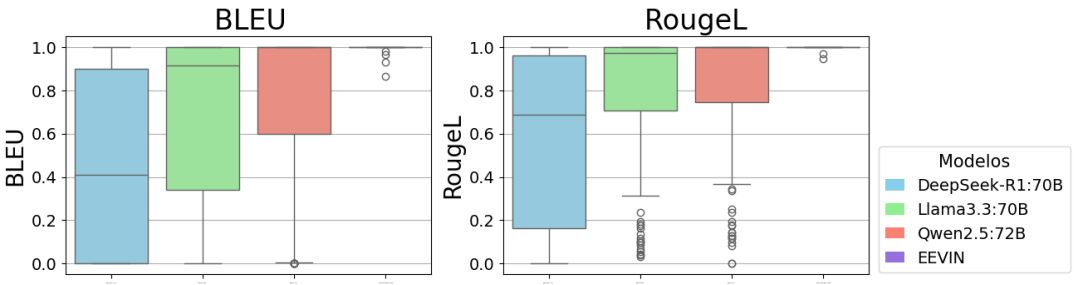

Figura 1. BLEU (à esquerda) e Rouge-L (à direita), por modelo.

A avaliação dos modelos utilizou como base de dados 50 registros eletrônicos de saúde da SemClinBR anotados manualmente para essa tarefa por um dos autores, se-

[2]Disponível em: `https://github.com/Joniiss/timeline-extraction-prompt`

guindo o protocolo descrito na seção Base de Dados. A Tab. 1 exibe a quantidade de eventos detectada por cada modelo. EEVIN encontra 135 de 136 eventos, o que era esperado. Todos os LLMs retornam uma quantidade de eventos superior à quantidade verdadeira, indicando que tais modelos tendem a fragmentar um evento em vários. Esse fenômeno é mais acentuado nos modelos Llamma e Qwen, que extraem 178 e 201 eventos, respectivamente.

Para verificar a presença de alucinação nos resultados obtidos com LLMs, comparamos a extensão de cada evento extraído com a porção do texto original em cada evento do mesmo prontuário. Foram calculadas as métricas ROUGE-L e BLEU por evento, comparando a extensão de um evento predito pelo modelo contra o evento original, indicando a correspondência entre eles. A extensão dos eventos deve coincidir com o texto descrito nos prontuários; logo, se ROUGE-L e BLEU são 1, não houve alucinação. É importante observar que, por ser uma implementação algorítmica das regras da anotação dos dados, é esperado que EEVIN tenha resultados excelentes nessas métricas.

Apesar de identificar a quantidade de eventos mais próxima da realidade, DeepSeek-R1:70B obteve resultados inferiores em todas as métricas. Ao inspecionar saídas geradas pelo modelo, percebemos uma frequente alteração e ocultação de palavras, afetando negativamente seu desempenho. Llama3.3:70B foi capaz de manter o texto original sem alterações, exceto pela ocultação do gatilho no campo de extensão do evento e fragmentação de eventos em alguns casos, impactando as métricas ROUGE e BLEU. Qwen2.5:72B obteve os melhores resultados nas métricas, mas também apresentou eventos fragmentados e alucinações nas respostas. Mesmo assim, como indicado na Fig. 1, o modelo obteve a melhor proporção de eventos perfeitos, indicados por 1.0 em Rouge-L (51,51%, seguido por Llama3.3:70B com 41,35% e 18,18% de DeepSeek-R1:70B) e em BLEU (50,75%, e 41,35% e 15,7% de Llama3.3:70B e DeepSeek-R1:70B, respectivamente). Ressaltamos também a diferença entre as médias de tempo gasto por execução dos algoritmos em cada nota clínica. Dentre os LLMs, Llama3.3:70B obteve o menor tempo, de 59,2s, enquanto DeepSeek-R1:70B e Qwen2.5:72B responderam, em média, em 92,7s e 161,6s, respectivamente. Por outro lado, devido ao seu custo computacional ser consideravelmente menor, EEVIN obteve um tempo médio de 37,6s.

4. Considerações finais

Os resultados iniciais demonstraram a superioridade de EEVIN tanto em métricas qualitativas como em custo computacional. Os LLMs investigados apresentaram principalmente dois problemas: a fragmentação de eventos e alucinações na extensão do evento. Sendo baseado em regras, EEVIN não está sujeito à alucinações desse tipo. Como continuação deste trabalho, pretende-se: i) ampliar os experimentos à toda a base de dados SemClinBR; ii) analisar a complexidade computacional; iii) analisar a viabilidade de utilização clínica do EEVIN; e iv) inspecionar a pegada de carbono das soluções comparadas, utilizando bibliotecas como CodeCarbon [3] ou EcoAI [4].

Agradecimentos

Agradecemos à PUC-Campinas, ao AIoT Lab Brasil pelos recursos computacionais, e ao CNPq (#177575/2024-7) pelo financiamento do projeto.

[3]CodeCarbon - `http://codecarbon.io/`

[4]Eco2AI - `https://github.com/sb-ai-lab/Eco2AI`

Referências

[Benício 2020] Benício, D. H. P. (2020). Aplicação de mineração de texto e processamento de linguagem natural em prontuários eletrônicos de pacientes para extração e transformação de texto em dado estruturado. Master's thesis, Universidade Federal do Rio Grande do Norte.

[Chen et al. 2015] Chen, Y., Xu, L., Liu, K., Zeng, D., and Zhao, J. (2015). Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks. In Zong, C. and Strube, M., editors, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–176. Association for Computational Linguistics.

[DeepSeek-AI et al. 2025] DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., et al. (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

[Grattafiori et al. 2024] Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., et al. (2024). The llama 3 herd of models.

[Johnson et al. 2016] Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035.

[Juhn and Liu 2020] Juhn, Y. and Liu, H. (2020). Artificial intelligence approaches using natural language processing to advance ehr-based clinical research. *Journal of Allergy and Clinical Immunology*, 145(2):463–469.

[Li et al. 2024] Li, Q., Li, J., Sheng, J., Cui, S., Wu, J., Hei, Y., Peng, H., Guo, S., Wang, L., Beheshti, A., and Yu, P. S. (2024). A survey on deep learning event extraction: Approaches and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 35(5):6301–6321.

[Liu et al. 2020] Liu, J., Chen, Y., Liu, K., Bi, W., and Liu, X. (2020). Event Extraction as Machine Reading Comprehension. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1641–1651. Association for Computational Linguistics.

[Oliveira et al. 2022] Oliveira, L. E. S. E., Peters, A. C., da Silva, A. M. P., Gebeluca, C. P., Gumiel, Y. B., Cintho, L. M. M., Carvalho, D. R., Al Hasan, S., and Moro, C. M. C. (2022). SemClinBr - a multi-institutional and multi-specialty semantically annotated corpus for Portuguese clinical NLP tasks. *Journal of Biomedical Semantics*, 13(1):13.

[Perera et al. 2013] Perera, S., Sheth, A., Thirunarayan, K., Nair, S., and Shah, N. (2013). Challenges in understanding clinical notes: Why NLP engines fall short and where background knowledge can help. In *Proceedings of the 2013 International Workshop on Data Management & Analytics for Healthcare - DARE '13*, pages 21–26. ACM Press.

[Schneider et al. 2020] Schneider, E. T. R., de Souza, J. V. A., Knafou, J., Oliveira, L. E. S. e., Copara, J., Gumiel, Y. B., Oliveira, L. F. A. d., Paraiso, E. C., Teodoro, D., and Barra, C. M. C. M. (2020). BioBERTpt - a Portuguese neural language model for clinical named entity recognition. In Rumshisky, A., Roberts, K., Bethard, S., and

Naumann, T., editors, *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 65–72, Online. Association for Computational Linguistics.

[Walker, Christopher et al. 2006] Walker, Christopher, Strassel, Stephanie, Medero, Julie, and Maeda, Kazuaki (2006). ACE 2005 Multilingual Training Corpus.

[Yang et al. 2024] Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F., Dong, G., Wei, H., Lin, H., Tang, J., Wang, J., Yang, J., Tu, J., et al. (2024). Qwen2 technical report.