Avaliação do Amazon Comprehend Medical: Extrator de informações médicas

Gabriel H. Lins, *Mestrando, Instituto Santos Dumont*

***Abstract*—O Amazon Comprehend Medical(ACM) é um sistema de Deep Learning por meio de programação de linguagem natural que identifica e extrai automaticamente conceitos clínicos de documentos de texto. Ele é conhecido por funcionar com uma boa acurácia nos aplicativos e sites médicos existentes. A aceitação e a confiança em novos produtos de dados baseiam-se na validação independente em vários conjuntos de dados e ferramentas. Este resumo tem como objetivo analisar a tarefa de extração de medicamentos do ACM como uma ferramenta para identificar e classificar informações.**

***Index Terms*—Inteligência artificial, Programação de Linguagem Natural, informações médicas.**

I. Introdução

Devido ao crescente número de instituições de saúde que adotam registros eletrônicos de saúde, a prevalência de texto livre em registros eletrônicos em saúde aumentou.[1] Este tipo de documento é comumente usado para documentação e comunicação. Embora seja possível contratar especialistas para realizar a tarefa, não é raro ocorrer erros humanos na escolha da relevância das informações,[2] sendo assim ainda é necessário que sejam implementados sistemas que possam extrair e interpretar facilmente as informações médicas.[3] Isso pode ser feito por meio do uso de técnicas de processamento de linguagem natural.[4] No campo do extração da informação, existem várias subtarefas e duas delas são: reconhecimento de entidade, que trata do reconhecimento de nomes que são entidades e classificação em categorias predefinidas, como nomes de pessoas, lugares ou organizações e extração de relacionamento, que se concentra na identificação das relações entre as entidades extraídas.[5-6]

Existem basicamente dois tipos de métodos empregados para essas subtarefas. A primeira e predominante abordagem usada no domínio clínico é baseado em regras, que consiste em uma coleção de regras feitas à mão, como expressões regulares e lógicas, que requerem colaboração com especialistas da área. O segundo método é o aprendizado de máquina. Embora subutilizado no domínio clínico, esta abordagem é mais portátil e escalável e, em geral, produz melhores resultados, desde que o modelo é treinado usando um grande conjunto de dados [7]. Alguns sistemas de sucesso, no entanto, utilizam ambos os métodos simultaneamente e, portanto, são chamados sistemas híbridos [8]. Diversas empresas de tecnologia desenvolveram seus próprios sistemas IE. Mais recentemente, Amazon Web Services (AWS) lançou o Amazon Comprehend Medical (ACM). O maquinário da ACM é impulsionado por modelos de aprendizado profundo de última geração [9].Os algoritmos e tecnologias subjacentes que alimentam o ACM já foram implementados, treinados e atualizados rotineiramente conforme os requisitos do usuário final evoluem. Vários modos de acesso, console da web e acesso programático via Python ou Java, atendem a um amplo público com vários níveis de habilidade e facilidade para tecnologia. Os usuários só pagam pelo que eles usarem, sem taxas mínimas e compromissos iniciais (ou seja, US0,01 por unidade, onde 1 unidade = 100 UTF-8 caracteres) [9]. Atualmente, apenas notas de texto clínico escritas em inglês são suportadas por ACM [9].ACM identifica automaticamente entidades, relacionamentos e negação de conceitos (por exemplo, o paciente não tomou o medicamento devido a reações alérgicas conhecidas) por meio da Extração e Identificação de Dados de Informação Médica Protegida de Saúde (PHId), que se concentra em informações de saúde protegidas (PHI) apenas, e Extração de Entidade Médica Nomeada e Relacionamento (NERe), que detecta os seguintes conceitos clínicos: 1. Anatomia: relaciona-se com as partes do corpo e sistemas e a localização / direcionalidade correspondente (por exemplo, dorsal, ventral, proximal e distal). 2. Condição médica: envolve o nome do diagnóstico e a acuidade correspondente (por exemplo, crônica ou aguda), sinais e sintomas. 3. Informações de saúde protegidas: enfoca várias PHI, como nome do paciente, data de nascimento e previdência social número. 4. Nome do teste, nome do tratamento e nome do procedimento: trata dos testes de diagnóstico, intervenções e tratamento procedimentos relacionados a uma condição médica. 5. Medicamentos: inclui o nome (ou seja, nome genérico e marca), dosagem, duração, frequência, forma, frequência, e força.

A aceitação e a confiança em qualquer novo produto de dados, especialmente aqueles aplicados a informações médicas delicadas, dependem de validação independente em conjuntos de dados de referência e/ou ferramentas para estabelecer e confirmar a qualidade esperada dos resultados. Esse o trabalho concentra-se na extração de medicamentos, que busca extrair entidades e relações de medicamentos a partir de anotações clínicas. As entidades incluem atributos como nome, dosagem, frequência, modo, duração, força e forma. Recuperação precisa e a divulgação dessas informações é essencial porque os medicamentos desempenham um papel vital no atendimento ao paciente, especialmente na doença prognóstico e sobrevivência. Erros de medicação podem levar a eventos adversos com medicamentos, que são qualquer dano ou lesão sofrida por paciente devido à intervenção medicamentosa [11]. Esses casos são atribuídos a 7.000-9.000 mortes evitáveis anualmente, e despesas associadas pelo governo dos

Este trabalho teve suporte do Instituto Santos Dumont

Programa de Pós-graduação em Neuroengenharia. Instituto Internacional de Neurociências Edmond e Lily Safra, Instituto Santos Dumont. Av. Alberto Santos Dumont, 1560 – Zona Rural. 59280-000 Macaíba/RN, Brasil (e-mail: gabriel.lins@edu.isd.org.br).

EUA totalizam US40 bilhões. ACM é apoiado por algoritmos de aprendizagem profunda. [13]. Tal configuração supera a principal limitação de acesso limitado a dados médicos para fins de treinamento e garante generalização do sistema nas diferentes especialidades médicas [14]. Finalmente, A função Extração de relações da ACM apresenta um método que combina pontuações de relação de segunda ordem que utiliza um token de contexto que conecta as duas entidades alvo e pontuações de relação de primeira ordem[15-16].

A. Método

O ACM está disponível em vários modos de acesso - console da web e acesso programático via Python ou Java. Experiências em todos os trabalhos foram realizados em Python por meio da biblioteca boto3 para acessar a API NERe. Cada documento é processado um por vez, de forma síncrona, e leva aproximadamente 11 segundos para produzir saídas estruturadas para um único documento com 12.000 caracteres. A saída de ACM inclui o seguinte: a) conceito clínico (por exemplo, anatomia, PHI ou medicação, etc.), b) tipo (o campo onde a entidade pertence), c) texto (entidade extraída real), d) pontuação de confiança (medida quantitativa para determinar o quão certo o sistema está com sua previsão / extração), e) deslocamento (localização da entidade extraída no documento), f) traços (informações compreendido pelo sistema com base no contexto), atributos (informações relacionadas à entidade extraída) [9].

B. Discussão

As notas de texto clínico apresentam entidades de medicamentos por meio de lista (ou seja, enumeração de medicamentos) e / ou por meio de narrativas (ou seja, medicamentos são incorporados em frases, que frequentemente também contêm informações não relacionadas a medicamentos). [17]. Com base em todos os conjuntos de dados de avaliação, uma das deficiências mais perceptíveis do ACM diz respeito à classificação incorreta de certos medicamentos como nomes de tratamento. Alguns exemplos importantes que surgiram incluem oxigênio (O2), glóbulos vermelhos empacotados (PRBCs), quimioterapia, inibidor de ace, diuréticos, beta-bloqueadores (BB) e bloqueadores do receptor de angiotensina II (ARBs). Além disso, ACM teve dificuldades em distinguir nomes de medicamentos que se relacionam com grupos de medicamentos, como casa medicamentos, medicamentos para o coração, medicamentos para a dor, narcóticos e antibióticos. Embora se possa argumentar que isso é uma questão de interpretação, essas nuances podem ter implicações para casos de uso específicos, especialmente quando servem como entradas para aplicativos downstream automatizados. Outro comportamento curioso que surgiu foi a sensibilidade de ACM ao espaço em branco. Isso é especialmente aplicável a medicamentos que são compostos por mais de um substantivo. Por exemplo, se o nome completo do medicamento ”complexo de polissacarídeo de ferro” acontece de ter uma nova linha inserida após o polissacarídeo, a entidade extraída por ACM incluía apenas “complexo de ferro”. Dado o fato de que as notas clínicas são atormentadas por inúmeros problemas sintáticos e gramaticais, esse comportamento pode impactam significativamente o desempenho e geram resultados instáveis.

II. Conclusão

Atualmente, o ACM apresenta algumas limitações técnicas, que podem ser interpretadas como deficiências no tratamento de medicamentos extração de informações. Primeiro, o sistema está perdendo o recurso de detecção do motivo do medicamento. Informações relativas à indicação de medicamentos são incluídas nas notas do texto clínico porque educam o paciente e diminui a inadequada prescrições pelo provedor de saúde e, de maneira geral, melhora a segurança do paciente[24]. Dito isto, é crucial para um extração de informações como o ACM deve ser equipado com esse recurso, especialmente ao extrair informações de medicação. Outra limitação técnica do ACM é o limite de comprimento de documento de 20.000 UTF-8. Ter notas clínicas extensas e ricas em informações são não é incomum em instituições de saúde nos Estados Unidos porque, além de descrever detalhes sobre o atendimento ao paciente, também contém outras informações relacionadas à conformidade e reembolso, que, em última análise, causam o que Kuhn et. al todos chamam como “Note bloat.” [25] Portanto, a restrição de comprimento atual impediria o processamento direto e imediato de tais notas longas. Apesar dessas deficiências que não são incomuns em um produto de dados complexo, os benefícios que o ACM oferece são reais. Por fim, ele fornece uma API limpa que pode ser facilmente escalonada para lidar com grandes cargas de trabalho, garantindo os dados segurança e privacidade. Ao todo, a ACM apresenta uma abordagem promissora com sua simplicidade e usabilidade integradas.

Referências

[1] Blumenthal, D. (2010). Launching hitech. New England Journal of Medicine, 362(5), 382-385.

[2] Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., ... Liu, H. (2018). Clinical information extraction applications: a literature review. Journal of biomedical informatics, 77, 34-49.

[3] Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., Chute, C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association, 17(5), 507-513.

[4] Small, S. G., Medsker, L. (2014). Review of information extraction technologies and applications. Neural computing and applications, 25(3-4), 533-548.

[5] Mikheev, A., Moens, M., Grover, C. (1999, June). Named entity recognition without gazetteers. In Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics (pp. 1-8). Association for Computational Linguistics.

[6] Bach, N., Badaskar, S. (2007). A review of relation extraction. Literature review for Language and Statistics II, 2.

[7] Doan, S., Collier, N., Xu, H., Duy, P. H., Phuong, T. M. (2012). Recognition of medication information from

discharge summaries using ensembles of classifiers. BMC medical informatics and decision making, 12(1), 36.

[8] Tang, B., Wu, Y., Jiang, M., Chen, Y., Denny, J. C., Xu, H. (2013). A hybrid system for temporal information extraction from clinical text. Journal of the American Medical Informatics Association, 20(5), 828-835.

[9] Amazon Web Services. Accessed May 15, 2019 from https://aws.amazon.com/comprehend/medical/.

[10] Koda-Kimble, M. A. (2012). Koda-Kimble and Young's applied therapeutics: the clinical use of drugs. Lippincott Williams Wilkins.

[11] Kohn, L. T., Corrigan, J., Donaldson, M. S. (2000). To err is human: building a safer health system (Vol. 6). Washington, DC: National academy press.

[12] Tariq RA, Scherbak Y. Medication Errors. [Updated 2019 Apr 28]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2019 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK519065/

[13] Jin, M., Bahadori, M. T., Colak, A., Bhatia, P., Celikkaya, B., Bhakta, R., ... Doman, T. (2018). Improving hospital mortality prediction with medical named entities and multimodal learning. arXiv preprint arXiv:1811.12276.

[14] Bhatia, P., Arumae, K., Celikkaya, B. (2018). Dynamic Transfer Learning for Named Entity Recognition. arXiv preprint arXiv:1812.05288.

[15] Bhatia, P., Celikkaya, B., Khalilia, M. (2018). End-to-end Joint Entity Extraction and Negation Detection for Clinical Text. arXiv preprint arXiv:1812.05270.

[16] Singh, G., Bhatia, P. (2019). Relation Extraction using Explicit Context Conditioning. arXiv preprint arXiv:1902.09271.

[17] Uzuner, Ö., Solti, I., Cadag, E. (2010). Extracting medication information from clinical text. Journal of the American Medical Informatics Association, 17(5), 514-518.

[18] Patrick, J., Li, M. (2010). High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. Journal of the American Medical Informatics Association, 17(5), 524-527.

[19] Doan, S., Bastarache, L., Klimkowski, S., Denny, J. C., Xu, H. (2010). Integrating existing natural language processing tools for medication extraction from discharge summaries. Journal of the American Medical Informatics Association, 17(5), 528-531.Tao, C., Filannino, M., Uzuner, Ö. (2017). Prescription extraction using CRFs and word embeddings. Journal of biomedical informatics, 72, 60-66.

[20] Tao, C., Filannino, M., Uzuner, Ö. (2017). Prescription extraction using CRFs and word embeddings. Journal of biomedical informatics, 72, 60-66.

[21] Stubbs, A., Buchan, K., Filannino, M., Uzuner, O. (n.d.). 2018 n2c2 shared task track 2 [PDF].

[22] Johnson, A. E., Pollard, T. J., Shen, L., Li-wei, H. L., Feng, M., Ghassemi, M., ... Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. Scientific data, 3, 160035.

[23] Nakayama, Hiroki, doccano, (2019), GitHub repository, https://github.com/chakki-works/doccano

[24] Li, Y., Salmasian, H., Harpaz, R., Chase, H., Friedman, C. (2011). Determining the reasons for medication prescriptions in the EHR using knowledge and natural language processing. In AMIA Annual Symposium Proceedings (Vol. 2011, p. 768). American Medical Informatics Association.

[25] Kuhn, T., Basch, P., Barr, M., Yackel, T. (2015). Clinical documentation in the 21st century: executive summary of a policy position paper from the American College of Physicians. Annals of internal medicine, 162(4), 301-303.