Ambiente de Treinamento Médico baseado em Casos Clínicos -- Módulo de Análise Visual dos Dados

bolsista: Enzo Hideki Iwata
ra: 215394
Orientador: André Santanchè
Local de Execução: Universidade Estadual de Campinas(UNICAMP)
Vigência: 07/08/2019 - 30/09/2020

1. Introdução e enunciado do problema

O projeto aqui apresentado está inserido em projeto maior multidisciplinar, envolvendo diversos pesquisadores, incluindo médicos que atuam no setor de emergência do Hospital das Clínicas da Unicamp e professores da Faculdade de Ciências Médicas da Unicamp. O problema motivador é a dificuldade de preparação de médicos para atendimento de casos clínicos, em ambiente controlado e livre de pressões de um atendimento real em um hospital. A abordagem escolhida para tratar o problema consiste na concepção e implantação de uma plataforma de aprimoramento baseada na resolução de casos clínicos. Ela inova por ser fortemente guiada pela análise de dados, tanto na construção de casos, quanto no acompanhamento do aluno.

O objetivo específico deste trabalho foi contribuir com técnicas de recuperação de informação e visualização dos dados gerados pelo ambiente para a sua análise. A parte de recuperação de informação envolveu: ligar as respostas com bases de conhecimento na área de saúde e permitir a verificação do quão próximo a resolução submetida está do resultado esperado; combinar a ligação com técnicas de recuperação de informação, que permite que palavras próximas da resposta esperada sejam também consideradas.

A abordagem inicialmente adotada para tornar a interpretação mais precisa foi a técnica da recuperação de informação, contudo tal abordagem se baseia numa comparação exata entre as palavras, no intuito de contornar esse problema utilizamos uma abordagem de interpretação de texto mais recente, baseada em técnicas de aprendizado profundo (deep learning), o BERT. A partir dessa abordagem relacionamos dados inseridos pelo usuário com bases de conhecimento. Dois exemplos são: textos do caso inseridos por autores e informações produzidas pelos alunos.

Como resultados, implementamos um anotador automático, que é capaz de relacionar diretamente termos Mesh com um texto livre (baseado em técnicas de recuperação de informação) e associar termos complexos a rótulos baseada no contexto.

2. Revisão da literatura resumida

Manning et al. foi uma bibliografia básica adotada por reunir os fundamentos de recuperação de informação usados neste projeto (MANNING et al., 2009). Além disso, trabalhamos com o Medical Subject Headings(MeSH) é um dicionário controlado pelo U.S.

National Library of Medicine, também usado para indexar artigos no PubMed. Nesta revisão resumida, daremos atenção ao BERT que foi um dos elementos centrais da pesquisa.

BERT - Bidirectional Encoder Representations from Transformers

O BERT - Bidirectional Encoder Representations from Transformers (Devlin, Jacob, et al) é um método de reconhecimento de linguagem natural apresentado pela Google com o intuito de se compreender melhor textos. Trata-se de um modelo baseado em redes profundas. Ele é primeiramente pré-treinado para a interpretação de textos em geral e depois pode ser especializado para exercer determinada função ou em determinado domínio. No caso deste projeto, a função que utilizamos é a de reconhecimento de entidade nomeada (NER), para assim classificar palavras e agrupá-las em uma área do conhecimento. Dado que o projeto é especializado na área médica, utilizamos uma variante do BERT, o BioBert (Lee, Jinhyuk, et al.), o qual foi pré-treinado com textos relacionados à área biomédica.

O BioBERT utiliza o mesmo esquema do BERT, com a diferença de que enquanto o BERT foi treinado com dados de diversas áreas, o BioBERT foi treinado com dados com foco na área biomédica, dessa forma, gerando um modelo de especializado nos termos da área, e contribuindo ainda mais para o entendimento dos textos da harena.

3. Métodos

Solr e Mesh

Essa primeira parte do projeto iniciou-se com o estudo teórico dos conceitos de recuperação de informação através das bibliografias citadas na revisão da literatura. Após essa etapa, começamos a instalação e estudo de uma das principais ferramentas de recuperação de informação, o Apache Solr (https://lucene.apache.org/solr/). Também foi feita uma análise de bancos de dados médicos e decidimos usar o Medical Subject Headings (MeSH).

A integração entre o MeSH e o Apache Solr envolveu um extenso trabalho de conversão de dados. O formato estruturado adotado pelo MeSH na forma de thesaurus classifica seus termos em uma taxonomia, acrescentando elementos descritivos e correlações entre eles. Esse formato foi convertido em uma estrutura para recuperação de informação, que explora o formalismo e as relações do MeSH para dar pesos diferenciados aos termos, melhorando os resultados da recuperação.

Antes de usarmos diretamente o Apache Solr tentamos fazer o acesso pelo Python através de uma biblioteca chamada Pysolr (https://github.com/django-haystack/pysolr), a qual faz uma conexão com o Solr. Contudo a adição de novos documentos era muito lenta, assim um banco de dados como o MeSH demorava muito para ser adicionado. Optou-se pela conexão diretamente com o Apache Solr. A adição foi bem mais rápida e eficaz nesse caso. Para usarmos diretamente o Solr foi necessário o estudo de XPath (https://www.w3schools.com/xml/xpath_intro.asp) e XQuery (https://www.w3schools.com/xml/xquery_intro.asp) para a conversão de dados do MeSH para o Solr, pois o formato de dados que é adicionado no Solr é XML.

Desse modo, partimos do MeSH em formato de XML e o transformamos para que o Solr o interpretasse e dessa forma criássemos o nosso próprio “dicionário” com funcionalidades de

recuperação de informação. O dicionário criado deu foco em campos necessários para a recuperação de informação, assim descartando informações que só acarretaria em mais tempo de processamento sem contribuição na recuperação.

Os campos escolhidos para se indexar no dicionário foram:

- DescriptorUI: ID associado a um elemento do MESH;
- ConceptName: nome de um elemento do MESH, como Doenças, por exemplo;
- ConceptUI: ID relacionada ao nome de um elemento do MESH;
- EntryTerm: termos associados ao elemento, como sinônimos;
- Annotation: informações adicionais sobre o elemento;
- ScopeNote: definição do elemento;
- DateCreated: data em que o elemento foi adicionado.

Observando o trabalho do Solr com o Mesh, percebemos que ele sozinho não era suficiente para o que buscamos, com isso, resolvemos partir para um abordagem diferente baseada no processamento de linguagem natural, o BERT.

BERT

Numa segunda etapa do projeto, decidimos explorar um caminho diferente para realização da recuperação de informação. Ela se baseou no reconhecimento de entidades nomeadas (NER), classificando a palavra em um contexto, em vez de usar métricas de recuperação de informação em documentos do MeSH. Para realização dessa etapa, foi realizado um estudo teórico sobre o BERT, envolvendo tanto sua arquitetura quanto sua aplicação através das bibliografias citadas na revisão da literatura. Para os propósitos da pesquisa, foi usado uma variante do BERT, o BioBERT, referenciado anteriormente.

Como o treinamento final do BioBert necessita de uma série de dados já rotulados foi necessário um exaustivo trabalho de coleta de bancos de dados. O trabalho com os dados foi principalmente buscar os bancos de dados públicos e organizar a rotulação dada. Para organizar as rotulações foi preciso resumir as rotulações à rótulos mais gerais, já que alguns dados possuíam rótulos muito específicos, o qual poderiam ser resumidos para rótulos mais gerais.

A normalização feita foi baseada nos estudos de cada banco de dados e o que eles atribuem para cada rótulo.

O tratamento dos dados para entrada do modelo foi feito seguindo os princípios do BioBERT, portanto, adicionamos um token específico para o final das frase (‘PAD’), depois tokenizamos as frases com o próprio tokenizador gerado pelo pré treinamento do BioBERT.

O processo de treinamento foi feito usando o Python, já que é uma linguagem que tem suporte para treinar e testar o modelo, e ferramentas que ajudam no trabalho de técnicas de aprendizado profundo (técnica em que se baseia o BERT).

A utilização do BioBERT, após seu estudo, começou com o carregamento dos parâmetros gerados no pré-treinamento, para isso usamos as bibliotecas torch (https://pytorch.org/) e transformers (https://huggingface.co/transformers/). Já com os dados rotulados a mão

precisamos fazer algumas conversões de palavras para index, pois as entradas do modelo são numéricas e não em texto, e por fim fazemos o treinamento do modelo para realizar a tarefa de reconhecimento de entidades.

Banco de Dados usados para o treinamento/avaliação do BioBERT

- **Anatomical entity mention recognition(AnatEM):** entidades anatômicas anotadas
- **BioCreative II gene mention(BC2GM):** entidades genéticas anotadas
- **BioCreative IV Chemical and Drug(BC4CHEMD):** entidades químicas anotadas
- **BioCreative V Chemical Disease Relation(BC5CDR):** entidades químicas e doenças anotadas
- **National Center for Biotechnology Information disease(NCBI-disease):** doenças anotadas
- **Biomedical Natural Language Processing 9(BioNLP9):** proteínas anotadas
- **Biomedical Natural Language Processing 11(BioNLP11):** proteínas, genes e organismos anotados
- **Biomedical Natural Language Processing 13(BioNLP13):** anatomias, câncer, células, químicos, genes, órgãos, organismos, patologias, proteínas e tecidos anotados
- **Colorado Richly Annotated Full Text(CRAFT):** células, genes, químicos e taxonomias anotadas.

4. Discussão e Conclusão

Após os estudos experimentais do Solr e do BERT, percebemos que há um potencial interessante na aplicação de ambos de forma complementar do nosso problema. O primeiro adota estratégias de recuperação de informações e análise de documentos, se baseando inteiramente, por exemplo, na quantidade de aparições da palavra no contexto e quão rara ela é. O segundo se baseia numa interpretação contextual do texto, buscando fazer um reconhecimento da entidade melhor para cada situação.

O resultado deste projeto será aplicado no projeto maior, o Harena, na recuperação de informação e reconhecimento de entidades nomeadas dentro de textos produzidos por médicos.

5. Referências Bibliográficas

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* (Vol. 1, pp. 4171–4186). Association for Computational Linguistics (ACL).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 2017-December, pp. 5999–6009). Neural information processing systems foundation.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

Doğan, R. I., Leaman, R., & Lu, Z. (2014). NCBI disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, *47*, 1–10. https://doi.org/10.1016/j.jbi.2013.12.006

Pyysalo, S., & Ananiadou, S. (2014). Anatomical entity mention recognition at literature scale. *Bioinformatics*, *30*(6), 868–875. https://doi.org/10.1093/bioinformatics/btt580

Smith, L., Tanabe, L. K., Ando, R., Kuo, C. J., Chung, I. F., Hsu, C. N., … Wilbur, W. J. (2008, September 1). Overview of BioCreative II gene mention recognition. *Genome Biology*. https://doi.org/10.1186/gb-2008-9-s2-s2

Krallinger, M., Rabal, O., Leitner, F., Vazquez, M., Salgado, D., Lu, Z., … Valencia, A. (2015). The CHEMDNER corpus of chemicals and drugs and its annotation principles. *Journal of Cheminformatics*, *7*. https://doi.org/10.1186/1758-2946-7-S1-S2

Peng, Y., Rios, A., Kavuluru, R., & Lu, Z. (2018). Extracting chemical-protein relations with ensembles of SVM and deep learning models. *Database*, *2018*(2018). https://doi.org/10.1093/database/bay073

Li, J., Sun, Y., Johnson, R. J., Sciaky, D., Wei, C. H., Leaman, R., … Lu, Z. (2016). BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database : The Journal of Biological Databases and Curation*, *2016*. https://doi.org/10.1093/database/baw068

Kim, J., Pyysalo, S., Tomoko, O., Bossy, R., Nguyen, N., & Tsujii, J. (2011). Overview of BioNLP Shared Task 2011. In *Proceedings of the BioNLP Shared Task 2011 Workshop* (Vol. 2009, pp. 1–6).

Nédellec, C., Bossy, R., Kim, J.-D., Kim, J. J., & Ohta, T. (2013). Overview of BioNLP shared task 2013. In *BioNLP Shared Task 2013 Workshop* (pp. 1–7).

Bada, M., Eckert, M., Evans, D., Garcia, K., Shipley, K., Sitnikov, D., … Hunter, L. E. (2012). Concept annotation in the CRAFT corpus. *BMC Bioinformatics*, *13*(1). https://doi.org/10.1186/1471-2105-13-161

Dasiopoulou, S., Giannakidou, E., Litos, G., Malasioti, P., & Kompatsiaris, Y. (2011). A survey of semantic image and video annotation tools. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6050*, 196–239. https://doi.org/10.1007/978-3-642-20795-2_8

Keim, D., Andrienko, G., Fekete, J. D., Görg, C., Kohlhammer, J., & Melançon, G. (2008). Visual analytics: Definition, process, and challenges. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 4950 LNCS, pp. 154–175). https://doi.org/10.1007/978-3-540-70956-5_7

Lipscomb, C. E. (2000). Medical Subject Headings (MeSH). *Bulletin of the Medical Library Association*.

Manning, C. D., Raghavan, P., & Schutze, H. (2008). *Introduction to Information Retrieval*. *Introduction to Information Retrieval*. Cambridge University Press. https://doi.org/10.1017/cbo9780511809071