Utilização de Modelo de Extração de Texto em Registros Clínicos

Mateus Nicolas Balestrin[1], **Filipe Mendes Webber**[2], **Carlos Henrique Kuretzki**[3]

[1]Engenharia da Computação – Universidade Positivo (UP)
R. Prof. Pedro Viriato Parigot de Souza, 5300 – 81.280-330 – Curitiba – PR – Brasil

[2]Ciência da Computação – Universidade Positivo (UP)

[3]Engenharia e Ciência da Computação – Universidade Positivo (UP)
Gestão da Informação – Universidade Federal do Paraná (UFPR)

mateusnicolasup@gmail.com, filipewebber@gmail.com, carlos.kuretzki@up.edu.br

Abstract. *This study created a model of supervised machine learning to help health professionals diagnose patients. By using the Watson Natural Language Understanding and Watson Knowledge Studio it was possible to predict, given the symptoms, the probability of each disease. Our model of machine learning was able to extract all the required data directly from the electronic health reports.*

Resumo. *O presente estudo criou um modelo de aprendizado de máquina supervisionado para auxiliar os profissionais de saúde no diagnóstico de doenças. Utilizando o Watson Natural Language Understanding aliado ao Watson Knowledge Studio foi possível prever, a partir dos sintomas, quais as probabilidades de existência de determinadas doenças. O modelo de machine learning criado foi capaz de extrair dados diretamente dos prontuários eletrônicos.*

1. Introdução

Devido ao crescimento exponencial de dados biológicos, são necessárias ações revolucionárias na administração, análise e acessibilidade desses dados. A devida curadoria dessas informações é essencial para as descobertas biológicas e pesquisa biomédica, sendo evidente o seu reflexo no diagnóstico e tratamento de doenças. [Howe et al. 2008]

O presente estudo utilizou um processador de linguagem natural chamado *Watson Natural Language Understanding* (NLU) para processar os registros médicos disponíveis. Como o NLU não possui um modelo padrão adequado para o campo da Medicina, foi necessário utilizar o *Watson Knowledge Studio* (WKS) para criar um modelo customizado.

2. Revisão da Literatura

2.1. Aprendizado de máquina

Aprendizado de máquina (*machine learning*) aborda a questão de como construir computadores que melhoram automaticamente através da experiência. É um dos campos técnicos de crescimento mais rápido, situado no cruzamento da ciência da computação e estatística, e no centro da inteligência artificial e da ciência de dados. O progresso recente no aprendizado de máquina foi impulsionado tanto pelo desenvolvimento de novos algoritmos e teoria de aprendizado quanto pela contínua explosão na disponibilidade de dados online e computação de baixo custo. [Jordan and Mitchell 2015]

2.2. Informática na saúde

Os prontuários eletrônicos são uma fonte promissora para pesquisa. Ainda que esses dados tenham potencial para avanços significativos na prática e pesquisa clínica, o que determinará a sua verdadeira utilidade será a qualidade dessas fontes de dados. Para compreender e caracterizar com precisão os limites dessas fontes, é crucial estabelecer metodologias padronizadas para determinar e reportar a qualidade dos dados. [Kahn et al. 2016]

2.3. Watson Knowledge Studio

O *Watson Knowledge Studio* funciona de forma simples. Após o *upload* de documentos, manualmente são determinadas *regras(entidades e relações)* que serão utilizadas em outros documentos. A partir do aprendizado com os documentos anotados, o WKS consegue treinar um modelo de *machine learning* que extrairá automaticamente as informações.

3. Método

3.1. Amostra

Foram utilizados 100 registros de prontuários eletrônicos de pacientes que realizaram cirurgia bariátrica. Esses documentos são compostos de várias informações, como peso, altura, histórico familiar, hábitos e doenças. Ao fazer a análise dessas informações, será possível determinar tratamentos e procedimentos para o paciente por meio da integração do *Watson Knowledge Studio* com o *Natural Language Understanding*.

3.2. Procedimento

Para treinar o modelo de *machine learning*, foram escolhidos 20 dos 100 registros clínicos disponíveis. Foi então criado um conjunto de documentos composto por estes registros convertidos em linguagem natural. Como prática recomendada, o primeiro conjunto para anotação deve ser relativamente pequeno para definir diretrizes de anotação antecipadamente e padronizar o processo. [Fritzner 2017]

3.3. Entidades, relações e regras

Foram criadas várias entidades, visando a diversidade de pacientes que teriam seus registros clínicos analisados. Quanto mais complexo for o documento, maior será o número de entidades. Nota-se que uma entidade pode ser composta de outras entidades, e.g a entidade *doença* pode ser composta de doenças *metabólicas*, *gastrointestinais* e outras.

As relações criam uma ligação entre as entidades encontradas no documento. Como o escopo escolhido foi a área da saúde, foram criadas várias relações com esse viés, e.g relação entre as entidades *pessoa* e *medicamentos*.

Após realizar as análises, classificando os trechos de cada documento com as devidas entidades e relações, os documentos são submetidos e o modelo de aprendizado de máquina é treinado, absorvendo as regras criadas e podendo replicar os conceitos aprendidos em outros documentos.

4. Resultados

O *IBM Watson*, por meio do *IBM Cloud*, é uma ferramenta que presta maior rapidez e facilidade ao aprendizado de máquina, reduzindo as preocupações com os aspectos técnicos. Embora seja de fácil utilização, foram diversas as dificuldades encontradas no desenvolvimento deste projeto.

4.1. Quantidade de documentos

A quantidade de dados anotados necessários para obter resultados satisfatórios varia, dependendo da complexidade e do idioma utilizado. Linearmente, quanto mais documentos forem utilizados para fazer as análises, mais tempo será demandado do usuário.

Para reduzir o tempo gasto com a análise dos documentos recomenda-se fazer as anotações em equipe. Contudo, como podem haver diferenças entre os usuários no modo de anotar os documentos, é fundamental que seja criado um padrão de anotações e que alguém fique encarregado de revisá-las.

4.2. Língua

Apesar do WKS ter suporte para a língua portuguesa, quando utilizado dessa forma não conseguia produzir resultados assertivos. O *software* era capaz de reconhecer a língua, mas não identificava corretamente as doenças. Então traduzimos os termos para o inglês e o programa passou a mostrar devidamente as porcentagens de acerto.

Decidir o que se qualifica como entidades ou relações não é trivial e requer conhecimento médico. Por essa razão, esse passo extra de traduzir os termos para o inglês pode reduzir a qualidade dos resultados obtidos.

5. Conclusão

Uma das principais vantagens do método supervisionado de *machine learning* é a possibilidade de especialistas da área customizarem - através da criação adequada de entidades e relações - o modelo de treinamento, obtendo, desta forma, resultados mais precisos. Essa tarefa se mostrou fundamental para o devido funcionamento do NLU.

Ainda que tenha havido certa dificuldade na utilização do *IBM Watson* na língua portuguesa, os resultados obtidos após a tradução para o inglês foram satisfatórios. O modelo de *machine learning* criado foi capaz de extrair automaticamente as informações dos prontuários eletrônicos e conseguiu prever, a partir dos sintomas, quais as probabilidades de existência de determinadas doenças. Entretanto, como a necessidade de tradução dos termos pode reduzir a qualidade dos resultados, migraremos o projeto para o *TensorFlow* nas próximas etapas da pesquisa.

Referências

Fritzner, J. (2017). Automated information extraction in natural language. Master's thesis, Norwegian University of Science and Technology.

Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., Hill, D., Kania, R., Schaeffer, M., St Pierre, S., Twigger, S., White, O., and Yon Rhee, S. (2008). The future of biocuration. *Nature*, 455:47–50.

Jordan, M. I. and Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260.

Kahn, M., Callahan, T., Barnard, J., Bauck, A., Brown, J., Davidson, B., Estiri, H., Goerg, C., Holve, E., Johnson, S., Liaw, S., Hamilton-Lopez, M., Meeker, D., Ong, T., Ryan, P., Shang, N., Weiskopf, N., Weng, C., Zozus, M., and Schilling, L. (2016). A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *eGEMs*, 4:1244.