Projeto de Formatura – 2020 – Press Release

PCS - Departamento de Engenharia de Computação e Sistemas Digitais

Engenharia Elétrica – Ênfase Computação

Tema: Reconhecimento de Entidades Nomeadas na Língua Portuguesa

---

O Reconhecimento de Entidades Nomeadas (NER) é um dos problemas da área de Processamento de Linguagem Natural (PLN) que tem como objetivo a localização e classificação de entidades em textos livres escritos em linguagem natural. Entidades nomeadas são termos que possuem nome próprio que são de interesse no texto, comumente as entidades nomeadas reconhecidas em produtos de mercado e pesquisa são pessoas, lugares, organizações, valores monetários e datas, mas outras entidades também podem ser encontradas aplicando as mesmas técnicas de acordo com a necessidade da aplicação.

Embora exista grande esforço e avanço de pesquisa em todas tarefas de PLN, elas são desenvolvidas majoritariamente na Língua Inglesa e nem sempre são diretamente transferíveis para outras linguagens, que possuem características gramaticais e construções sintáticas diferentes. Assim, a pesquisa de técnicas de PLN aplicadas à Língua Portuguesa são importantes para desenvolver soluções e conhecimento público para esse idioma. Além disso, os conjuntos de dados que são fundamentais para o desenvolvimento de pesquisas de PLN, comumente grandes conjuntos de textos anotados manualmente chamados de corpus, disponíveis abertamente para uso científico também são majoritariamente em Língua Inglesa. Portanto, os insumos necessários para tais pesquisas também são escassos para a Língua Portuguesa e trabalhos como este produzem esses dados que poderão ser integrados em avanços futuros na área.

Dada a capacidade de extrair informações relevantes de textos em linguagem natural NER possui diversas aplicações em sistemas que possuem interação com textos escritos por ou para humanos. Algumas das possibilidades de aplicação são: busca de documentos relevantes com base em palavras chave de busca, sistemas de recomendação de conteúdo como notícias, categorização e agrupamento de reclamações ou pedidos em centrais de atendimento ao cliente, seleção de currículos em processos de recrutamento e reconhecimento de linguagem natural para chatbots e assistentes virtuais.

Este trabalho traz duas grandes contribuições para o estado de NER na Língua Portuguesa: um conjunto de dados anotados manualmente para treinamento de modelos de aprendizado de máquina e futuras pesquisas na área e um estudo abrangente sobre como construir modelos de aprendizado de máquina para reconhecimento de entidades nomeadas em português, com análise das melhores técnicas e suas aplicações nessa tarefa.

Foram coletados textos de notícias em português para a criação do corpus disponibilizado, por conta da sua diversidade de assuntos e grande variabilidade de entidades nomeadas presentes. As notícias foram coletadas, tratadas e selecionadas automaticamente, antes da fase de anotação manual dos rótulos de entidades de interesse. Com isso foi produzido um conjunto de dados que contempla aproximadamente 1 milhão de entidades rotuladas.

Com base nas pesquisas de processamento de linguagem natural mais recentes e avançadas na literatura mundial foram desenvolvidos algoritmos de aprendizado de máquina e aprendizado profundo para resolver a tarefa de reconhecimento de entidades nomeadas para a Língua Portuguesa. Os resultados obtidos nos experimentos deste trabalho produziram uma taxa de acerto acima de 95% na classificação das entidades nomeadas propostas usando o conjunto de dados coletado e disponibilizado.

Integrantes: Felipe Coelho de Abreu Pinna, Gabriel Takeshi Medeiros Yamasaki e Ian Alkmin Santos La Rosa

Professor Orientador: Prof. Dr. Ricardo Luis de Azevedo Rocha