Capítulo 1

Processamento de Linguagem Natural via Aprendizagem Profunda

Bárbara Stéphanie Neves Oliveira, Luís Gustavo C. do Rêgo, Lucas Peres, Ticiana L. Coelho da Silva, José Antônio F. de Macêdo

Abstract

Humans need to communicate. Out of this basic need combined with the Web, a vast amount of text has been generated on a daily basis. Given the presence of a lot of information allocated in different resources, it becomes vital to enable machines to understand spoken and written texts. This chapter presents how Deep Learning techniques can solve Natural Language Processing (NLP) tasks (e.g., Text Classification and Sentence Summarization), aiming to benefit from the computational power currently available and the low need for feature engineering when using these models. Initially, some essential concepts about NLP and Deep Learning are presented. Then, different pre-processing and textual representation techniques are explained to be used as input in Deep Learning models. Finally, it is shown how to apply the knowledge acquired in real applications of NLP.

Resumo

Seres humanos precisam de comunicação. A partir da junção dessa necessidade básica com a Web, uma grande quantidade de texto tem sido gerada diariamente. Dada a presença de muitas informações alocadas em diferentes meios, torna-se vital permitir que máquinas compreendam textos falados e escritos. Este capítulo apresenta como técnicas de Aprendizagem Profunda podem ser utilizadas na resolução de tarefas de Processamento de Linguagem Natural (PLN), como Classificação e Sumarização de Sentenças, visando o benefício do poder computacional disponível atualmente e da baixa necessidade de engenharia de features na utilização destes modelos. Inicialmente, são apresentados alguns conceitos importantes sobre PLN e Aprendizagem Profunda. Em seguida, diferentes técnicas de pré-processamento e representação textuais são explicadas a fim de serem usadas como entrada em modelos de Aprendizagem Profunda. Por fim, é mostrado como aplicar os conhecimentos adquiridos em aplicações reais do PLN.

1.1. Introdução

A comunicação, como uma necessidade básica da condição humana, juntamente com a existência da *Web* permitem que uma vasta quantidade de textos escritos e falados seja gerada diariamente. Dado o conteúdo textual presente em mídias sociais, aplicativos de bate-papo, *e-mails*, análises de produtos, artigos de notícias, trabalhos de pesquisa e *ebooks*, tornou-se vital a existência de um processamento automático de textos a fim de oferecer assistência ou tomar decisões para diversas tarefas diárias.

A capacidade de entender textos ou áudios em linguagem natural por uma máquina é um problema que vem sendo investigado há muito tempo [Chollet 2021]. As primeiras tentativas de construção de sistemas de Processamento de Linguagem Natural (PLN, de *Natural Language Processing* ou NLP) foram feitas através da análise intrínseca de linguagens que são naturalmente moldadas por um processo de evolução (por isso o termo "natural"). O PLN moderno envolve não apenas a habilidade de entendimento de uma linguagem como também possibilita, de forma automática, a extração de informações por meio de tarefas, tais como Classificação de Textos, Reconhecimento de Entidades Nomeadas (NER, de *Named Entity Recognition*), Desambiguação do Sentido das Palavras, e *Part-of-Speech* (POS) *tagging*.

Modelos de Aprendizagem Profunda, do inglês *Deep Learning* (também conhecida como Aprendizado Profundo ou Redes Neurais Profundas), aprendem vários níveis de representação de complexidade/abstração dos dados de forma crescente. Vários fatores evidenciam porque esses modelos têm sido amplamente usados em tarefas de PLN: (i) exigem pouca engenharia de *features*; (ii) produzem representações vetoriais que capturam similaridades de unidades linguísticas (palavras, por exemplo) presentes em textos, permitindo que sistemas de PLN possuam uma espécie de dependência de conhecimento; (iii) permitem aprendizado não supervisionado ou semi-supervisionado, o que é importante quando se tem um grande volume de dados e nenhum rótulo; (iv) aprendem vários níveis de representação, possibilitando que o nível mais baixo geralmente possa ser compartilhado entre diferentes tarefas; e (v) naturalmente lidam com a recursividade da linguagem humana, sendo capazes de capturar informações de forma sequencial.

O uso de modelos de Aprendizagem Profunda no PLN iniciou-se com a investigação da capacidade de compreensão de linguagem por Redes Neurais Recorrentes (RNNs, do inglês *Recurrent Neural Networks*) e redes LSTM (*Longest Shortest Term Memory*) [Hochreiter and Schmidhuber 1997]. Essas duas arquiteturas dominaram o PLN de forma geral de 2015 a 2017, uma vez que processam textos de comprimento variável. Os modelos LSTM bidirecionais, em particular, definiram o estado da arte em muitas tarefas importantes, desde Sumarização de Textos até Tradução Automática. Contudo, por volta de 2017 e 2018, uma nova arquitetura surgiu para "substituir" as RNNs: o *Transformer* [Vaswani et al. 2017], que permitiu um progresso considerável do PLN em um curto período de tempo.

O objetivo deste capítulo não é expor todas essas aplicações e arquiteturas de forma abrangente. Em vez disso, o foco está em como aplicar de forma prática representações textuais existentes e obtidas através de técnicas de Aprendizagem Profunda para resolver problemas de PLN. Este capítulo aborda ainda as diferentes etapas de processamento de textos realizada antes de treinar uma rede neural para uma tarefa de PLN. Ao