1
Introdução

Diariamente a Internet é populada com um grande volume de dados. Em Setembro de 2009, a Internet contava com mais de 1.733 milhões de usuários, que enviavam em média 247 bilhões de e-mails por dia [1]. Havia também 234 milhões de sites, dentre os quais 126 milhões sendo blogs. O YouTube exibia um bilhão de vídeos diariamente, o Flickr detinha 4 bilhões de imagens e o Facebook 2,5 bilhões com uma taxa de crescimento de 30 bilhões por ano. Além de textos, vídeos e imagens, outros dados menos óbvios são os cliques e os carrinhos de compra. Porém, a grande maioria desses dados não está em formatos bem estruturados de onde seria fácil extrair informações.

Das fontes de dados citadas, uma bastante interessante é o blog. Blogs são sites que contêm diários pessoais online com reflexões, comentários e *hyperlinks* geralmente fornecidos pelo autor [2]. Eles contêm uma enorme quantidade de textos nos quais podem ser aplicadas técnicas de Processamento de Linguagem Natural (PLN) para extrair informações.

Entre os blogs, podemos destacar o Twitter. O Twitter é um micro blog. A diferença entre um blog e um micro blog é que o segundo tem o tamanho de seus textos limitado. Em Abril de 2010, o Twitter já registrava 106 milhões de usuários, produzindo 55 milhões de pequenos textos por dia [3]. Todavia, existem certas particularidades que dificultam extrair informações desses textos, também conhecidos como *tweets*. Os *tweets* são escritos de forma a passar o máximo de informação com o mínimo de caracteres. Para isso, abusam do uso de abreviações, contrações, *emoticons* - desenhos em ASCII que tentam expressar o estado do usuário: feliz, triste, irritado - e chegam até a substituir partes ou palavras inteiras por representações fonéticas. Eles são sintaticamente mal formados, o que impossibilita o uso dos processadores linguísticos existentes. Uma solução é considerar essas mensagens como escritas em novas línguas, criadas a partir da língua de origem dos usuários que as escreveram. Batizamos de português-twitter a língua utilizada por brasileiros no Twitter.

O objetivo desta dissertação é construir um anotador morfossintático para o português-twitter, utilizando técnicas de Aprendizado de Máquina.

Um Anotador Morfossintático é um classificador que atribui uma classe a cada *token* de um texto, de acordo com o seu comportamento sintático no contexto. Escolhemos fazer esse classificador, pois a anotação morfossintática é uma tarefa fundamental de PLN, sendo utilizada para a solução de tarefas linguísticas mais complexas, tais como *phrase chunking* e *named entity recognition*. Para construir o anotador de *tweets* em português, utilizamos técnicas de aprendizado supervisionado. Adicionalmente, criamos um corpus anotado, dividido em duas partes: uma para treino e outra para teste. A parte de treino é composta de 39.686 sentenças. A parte de teste é composta por 5.490 sentenças. Como métrica de avaliação, adotamos a acurácia, que mede a percentagem do corpus corretamente anotada. Não temos conhecimento da existência de outros Anotadores Morfossintáticos para o português-twitter. Porém, sabemos que, no estado-da-arte da Anotação Morfossintática para o português, a acurácia é de aproximadamente 96% das anotações, variando de acordo com o conjunto de classes escolhido. Construímos o processador composto de dois estágios, um morfológico e um contextual. Nossos resultados experimentais apresentam uma acurácia de 90,24% para o anotador proposto. Isto corresponde a um aprendizado significativo, pois o sistema inicial tem uma acurácia de apenas 76,58%. Estes resultados são compatíveis com o aprendizado observado nos correspondentes processadores na língua portuguesa.

Essa dissertação está estruturada como descrito a seguir. No Capítulo 2, apresentamos a motivação desta dissertação, destacando a importância do Twitter como fonte de dados. No Capítulo 3, através de exemplos, definimos a função de um Anotador Morfossintático, mostramos as dificuldades de criar um para o português-twitter e quais as suas aplicações. No Capítulo 4, apresentamos os algoritmos utilizados. No Capítulo 5, descrevemos os processos de obtenção e anotação do corpus. No Capítulo 6, apresentamos a solução empregada e os resultados experimentais. No Capítulo 7, apresentamos as conclusões e futuros trabalhos.