7
Conclusão

O Twitter é um microblog que produz uma enorme massa de dados. Em Abril de 2010 já eram registrados 55 milhões de novas mensagens por dia. Entretanto, todo esse potencial é subutilizado. Muito conhecimento pode ser extraído se forem utilizados processadores de linguagem natural nessas entradas. Contudo, as linguagens utilizadas no Twitter são coloquiais e sintaticamente mal formadas, o que impossibilita o uso dos processadores existentes. Nessa dissertação, desenvolvemos um anotador morfossintático para o português-twitter, linguagem originada do português, utilizada por brasileiros no Twitter. Essa é uma tarefa fundamental no Processamento de Linguagem Natural, sendo utilizada para a solução de tarefas lingüísticas mais complexas.

Os corpora de treino e teste foram criados a partir de um processo que reduz o esforço de anotação por humanos, exigindo apenas fluência no idioma alvo. Para obter um corpus dourado é necessária a revisão das anotações de POS por um especialista, caso contrário, a confiança a ser depositada nele é a acurácia do anotador automático utilizado no processo. No nosso caso, aproximadamente 96%.

O anotador gerado a partir do corpus de treino apresentou uma acurácia de 90,24% no corpus de teste, com uma confiança de 96,6%. Não temos conhecimento da existência de outros Anotadores Morfossintáticos para o português-twitter, para comparar com nossos resultados. Entretanto, isto corresponde a um aprendizado significativo, pois o sistema inicial tem uma acurácia de apenas 76,58%. Além disso, o corpus traduzido pode ser anotado para outras tarefas do português-twitter. Finalmente, ressaltamos que a metodologia apresentada pode ser aplicada para outras linguagens do Twitter, como o Inglês-Twitter, por exemplo.

Como trabalhos futuros, com o intuito de melhorar a acurácia e aumentar a confiança, pretendemos revisar o corpus de treino e teste com o auxílio de linguistas.

Considerando que nosso *tokenizador* utiliza expressões regulares, novos *emoticons* poderiam ser tratados de forma errada. Sendo assim, iremos testar

a possibilidade de utilizar o ETL para *tokenizar* os *emoticons*.

Pretendemos também substituir o classificador unigrama utilizado na Seção 5.2.1 por um classificador ETL. Isso por que, quando um *token* tem mais de uma tradução possível, o classificador unigrama irá sempre sugerir a tradução mais frequente. Por exemplo, o *token* '+' aparece traduzido no corpus 70% como 'mais', 24% como '+', 3% como 'e', 3% como 'mas'. Consequentemente, o classificador irá traduzir, erroneamente, em dois dos três casos abaixo, o *token* '+' para 'mais'.

> "Somando as contas da viagem e das compras, deu: 2000+500=2500"

> "Bem q queria, + não vai dar para estudar + pq eu tenho niver p/ ir."

Utilizando o ETL, podemos construir um classificador que considera o contexto e, com isso, traduza corretamente o primeiro e o segundo *token* '+' para '+' e 'mas' respectivamente, facilitando a etapa de revisão da tradução.

Por último, pretendemos buscar por novos atributos que melhorem o desempenho do modelo criado.