6
Modelagem

Esse capítulo apresenta a modelagem do Anotador Morfossintático e os resultados obtidos.

6.1
O Modelo

Nossa modelagem de um Anotador Morfossintático para o português-twitter segue a estratégia de dois estágios proposta por Eric Brill [11] para anotar o POS de um corpus em inglês. Em 2008, Santos [14] a utilizou com sucesso para criar um Anotador Morfossintático para o português, obtendo com ela o estado-da-arte para a tarefa.

O primeiro estágio é um classificador morfológico. Através de informações morfológicas, são aprendidas regras que permitem classificar palavras conhecidas e desconhecidas. Por exemplo, mesmo que a palavra *suavemente* não se encontre no corpus de treino, ele a reconhece como um advérbio devido ao seu sufixo *mente*. No estágio seguinte, o classificador examina o contexto para corrigir erros do estágio anterior.

O classificador morfológico é composto por dois classificadores. O primeiro é um classificador unigrama, treinado para classificar palavras conhecidas. O segundo classificador é treinado para classificar as palavras desconhecidas. Ele é treinado com um TBL, usando os mesmos gabaritos adotados por Brill [11]. Os atributos utilizados na geração das regras a partir dos gabaritos são os seguintes:

- palavra como escrita no *tweet*;

- o POS predito;

- palavra normalizada, retiradas as letras repetidas;

- lista de caracteres da palavra normalizada;

- prefixo de tamanho n;

- prefixo de tamanho n que se removido resulta em uma palavra normalizada do corpus completo;

- lista de prefixos de tamanho n que se adicionados resultam em palavras normalizadas presentes no corpus completo. item - sufixo de tamanho n;

- sufixo de tamanho n que se removido resulta em uma palavra normalizada do corpus completo;

- lista de sufixos de tamanho n que se adicionados resultam em palavras normalizadas presentes no corpus completo;

- lista das palavras normalizadas que aparecem imediatamente depois do *token* no corpus completo;

- lista das palavras normalizadas que aparecem imediatamente antes do *token* no corpus completo.

onde o n utilizado em todos os atributos variou de 1 a 5.

O estágio contextual é implementado utilizando o Entropy Guided Transformation Learning (ETL) [12]. O ETL utiliza a árvore de decisão para resolver o gargalo do TBL, automatizando a geração de gabaritos de regras de correção. Os atributos, fornecidos ao ETL, utilizados na geração das regras são os seguintes:

- palavra como escrita no *tweet*;

- o POS predito;

- palavra normalizada, retiradas as letras repetidas.

A janela, parâmetro que indica o tamanho do contexto ao ETL, varia entre 3, 5 e 7. O treinamento é feito usando o estágio anterior como classificador inicial.

6.2 Experimentos

Para treinar e testar o classificador proposto utilizamos um Pentium Core I7 com 8GB de RAM. Como este tamanho de memória não é suficiente para treinar o classificador por contexto, utilizamos apenas 1/3 do corpus de treino para treiná-lo. Para treinar o classificador morfológico, utilizamos todo o corpus de treino.

Na Tabela 6.1, mostramos os resultados para um classificador unigrama usado como base de comparação (BLS), o Anotador Morfossintático treinado para o português (PT), que utilizamos na geração do corpus na Seção 5.2.1, o classificador morfológico e instâncias do classificador contextual para diferentes tamanhos de janela do ETL.

Modelo	**Janela**	**Acurácia (%)**
BLS	-	76,58
PT	-	82,76
Morfológico	-	86,59
Contextual	3	**90,24**
Contextual	5	90,17
Contextual	7	90,17

Tabela 6.1: Acurácia do corpus de teste.

Os valores dessa tabela são calculados para o corpus de teste. A melhor qualidade obtida está em negrito, correspondendo ao classificador contextual ETL com uma janela de tamanho 3. Considerando que o corpus não foi totalmente revisado por um especialista, o nível máximo de confiança que podemos depositar nele é de 96,6%.