Análise de Métodos para Reconhecimento Automático de Classes Morfossintáticas

A identificação automática das classes morfossintáticas, realizada por sistemas chamados de POS taggers, é importante para identificação de elementos no texto e utilizada por sistemas de processamento linguístico. Para construir tal sistema é preciso de corpora (textos) anotados com informações morfossintaticas, manualmente validados para gerar uma identificação acurada. Existem ferramentas com 96-97% (TreeTagger, WP DV) de acerto para essa tarefa na língua inglesa e temos para o português taxas de acerto de 99% com o parser PALAVRAS (por Eckhard Bick), porém ele não é open source. Dentre os disponíveis para o português, a taxa de acerto é em torno de 97% (Brill's TBL). Foi decidido então o desenvolvimento de um POS Tagger com diferentes abordagens e objetivo de compará-las. Nesta tarefa é importante analisar a acurácia do classificador variando o tamanho do corpus e o número de anotações utilizadas, e as abordagens utilizadas, tais como HMM, Máxima Entropia (ME), Naïve Bayes, Rede Neural. Neste sentido este trabalho analisa duas abordagens para treinamento dos POS taggers: modelo de ME e Perceptron. O modelo de ME combina diversas formas de informação contextual em uma maneira íntegra, e não impõe suposições de distribuição nos dados de treino. Dentre os algoritmos de redes neurais, o Perceptron é um classificador linear que mapeia a entrada para um valor de saída com um valor binário. Para a comparação das abordagens foi utilizado o corpus Amazonia, composto de 275.771 frases, 9683738 palavras e 22 classes morfossintáticas. Foram utilizados os modelos implementados no pacote OpenNLP. Os resultados preliminares obtidos com cross validation 10-fold com o modelo de ME foram de 96.8% de acerto e com o Perceptron de 97% de acerto. Para os resultados finais, deve-se ainda verificar a influência do tamanho do corpus de treinamento e da variação do erro médio. Este trabalho nos indicará qual o modelo tem melhores resultados e consequentemente qual será usado no desenvolvimento do POS Tagger.