Part-of-Speech Tagging with Limited Training Corpora

Robert Staubs
2004-2005
Period 1
TJHSST Computer Systems Laboratory

Abstract

The aim of this project is to create and analyze a part-of-speech (POS) tagger using limited training data and also to analyze the effect of genre on this level of tagging data. The corpus used is of extremely limited size thus offering less occasion to rely entirely upon tagging patterns gleamed from predigested data. The method used to analyze the data and resolve tagging ambiguities is the Viterbi Algorithm for finding internal structures of Hidden Markov Models. Results are analyzed by comparing the system-tagged corpus with a professionally tagged one.

Background

Many different methods of POS tagging have been advanced in the past but no attempts give hope of "perfect" tagging at the current stage. Accuracy of over 90% on ambiguous words is typical for most methods in current use, often well exceeding that. POS taggers cannot at the current time mimic human methods for distinguishing part of speech in language use. Work to get taggers to approach the problem from all the expected human methods—semantic prediction, syntactic prediction, lexical frequency, and syntactical category frequency being the most prominent—have not yet reached full fruition.

The highly accurate results are reached using corpora of very large size which are often prohibitively expensive and difficult to create. A method often used to analyze them is the Viterbi Algorithm which involves creating and combining the probabilities of individual tags for each words based on transitions and lexical properties and choosing the most probable.

Many corpora are divided into genre. Taggers trained for a specific genre tend to have better results but it is unclear how far this extends into smaller corpora.

Hypothesis

Tagging will be less effective; number around around 70-80% for accuracy are expected. The genred items may show a transition point where general tagging is more effective.

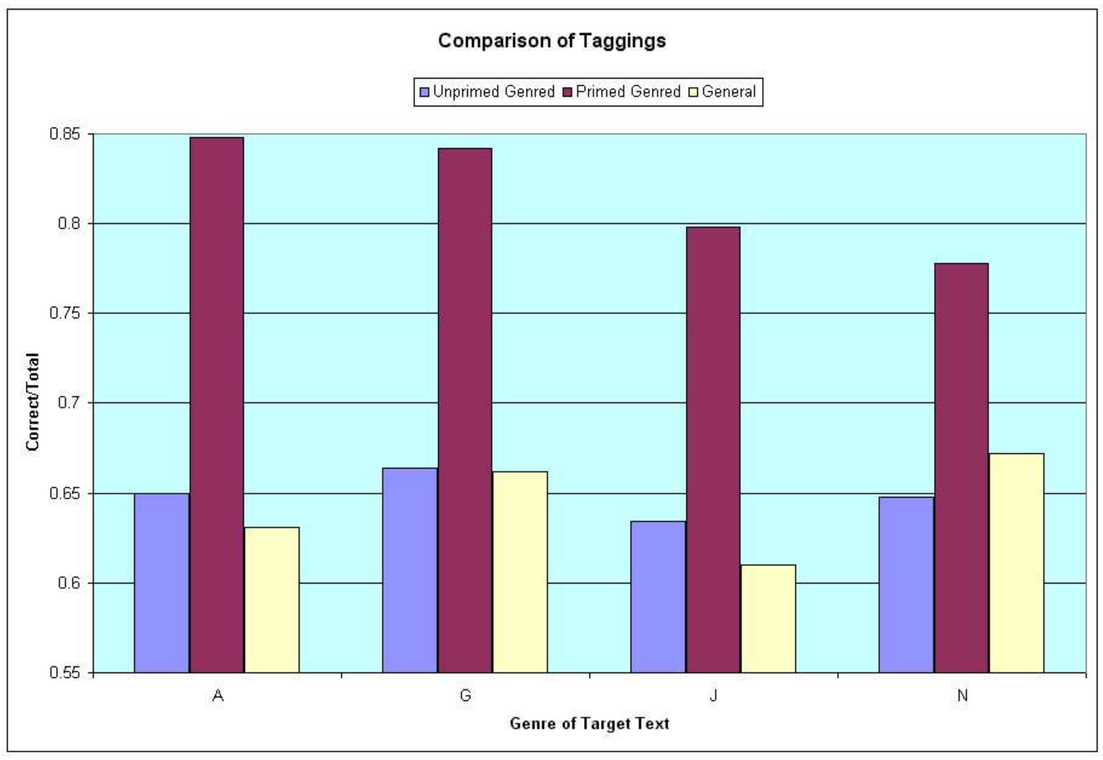
Procedure

Training data consists of: tags represented in the corpus, words represented in the corpus, transitions represented in the corpus, and the frequency of each. Words and tags are read in from the corpus and stored alphabetically series of arrays of structures designed to contain them. This data forms the basis for the statistical information extracted by taggers for making the probabilities for decisions on a unit's tag.

Actual Results

Results ranged from 60-67% accuracy for general tagging, 63-66% for genred tagging, and 77-84% for "primed" genred tagging trained on the target as well as the rest of the genre.

Changes in tagging accuracy with variable amount of data available could not be observed due to the small differences involved.



Comparison of tagging accuracy between test runs.

Conclusion

Tagging accuracy was somewhat less than expected, but it *was* lower as was predicted.

No transition point—where general tagging was clearly more effective—was observed. Instead for one genre it was more effective, for two it was somewhat less effective, and one it was very mildly less effective.

Limited-training tagging is not useful in and of itself but could be used as the basis for larger tagging operations or re-estimation algorithms.

The system could be improved by giving linguistic foreknowledge about morphological and writing rules about word class derivations.