Reliable and Cost-Effective Pos-Tagging

Yu-Fang Tsai*, and Keh-Jiann Chen*

Abstract

In order to achieve fast, high quality Part-of-speech (pos) tagging, algorithms should achieve high accuracy and require less manually proofreading. This study aimed to achieve these goals by defining a new criterion of tagging reliability, the estimated final accuracy of the tagging under a fixed amount of proofreading, to be used to judge how cost-effective a tagging algorithm is. In this paper, we also propose a new tagging algorithm, called the context-rule model, to achieve cost-effective tagging. The context rule model utilizes broad context information to improve tagging accuracy. In experiments, we compared the tagging accuracy and reliability of the context-rule model, Markov bi-gram model and word-dependent Markov bi-gram model. The result showed that the context-rule model outperformed both Markov models. Comparing the models based on tagging accuracy, the context-rule model reduced the number of errors 20% more than the other two Markov models did. For the best cost-effective tagging algorithm to achieve 99% tagging accuracy, it was estimated that, on average, 20% of the samples of ambiguous words needed to be rechecked. We also compared tradeoff between the amount of proofreading needed and final accuracy for the different algorithms. It turns out that an algorithm with the highest accuracy may not always be the most reliable algorithm.

Keywords: part-of-speech tagging, corpus, reliability, ambiguous resolution

1. Introduction

Part-of-speech (pos) tagging for a large corpus is a labor intensive and time-consuming task. Most tagging algorithms try to achieve high accuracy, but 100% accuracy is an impossible goal. Even after tremendous amounts of time and labor are spent on the post-process of proofreading, many errors still exist in publicly available tagged corpora. Therefore, in order to achieve fast, high quality pos tagging, tagging algorithms should not only achieve high accuracy but also require less manually proofreading. In this paper, we propose a context-rule