Open Science Index, Computer and Information Engineering Vol:14, No:11, 2020 publications.waset.org/abstracts/129270.pdf

# Exploring Syntactic and Semantic Features for Text-Based Authorship Attribution

**Authors :** Haiyan Wu, Ying Liu, Shaoyun Shi

**Abstract :** Authorship attribution is to extract features to identify authors of anonymous documents. Many previous works on authorship attribution focus on statistical style features (e.g., sentence/word length), content features (e.g., frequent words, n-grams). Modeling these features by regression or some transparent machine learning methods gives a portrait of the authors' writing style. But these methods do not capture the syntactic (e.g., dependency relationship) or semantic (e.g., topics) information. In recent years, some researchers model syntactic trees or latent semantic information by neural networks. However, few works take them together. Besides, predictions by neural networks are difficult to explain, which is vital in authorship attribution tasks. In this paper, we not only utilize the statistical style and content features but also take advantage of both syntactic and semantic features. Different from an end-to-end neural model, feature selection and prediction are two steps in our method. An attentive n-gram network is utilized to select useful features, and logistic regression is applied to give prediction and understandable representation of writing style. Experiments show that our extracted features can improve the state-of-the-art methods on three benchmark datasets.

**Keywords :** authorship attribution, attention mechanism, syntactic feature, feature extraction
**Conference Title :** ICDAR 2020 : International Conference on Document Analysis and Recognition
**Conference Location :** Jerusalem, Israel
**Conference Dates :** November 26-27, 2020