Stage1 - Authorship Identification

Da Eun Kim

Supervisor: Dr.Mark Stevenson
Module: COM3610

11 October, 2019

Chapter 1

Introduction

1.1 Background

Authorship is to ensure the writers who have substantive contributions to the paper have credits as an author and to remind their role to have responsibility on their publication. The task of Authorship Identification is to find the likely authorship in the given documents. According to Bozkurt et al (2007)[4], as the growth in Internet usage, information with unclear writers has been increased. This can be a problem since plagiarism, pseudonym malicious emails in these anonymous information can be a legal consequences. Also, Juola(2012)[5] said it is applicable to detect disputed historical inquires, fake news, journalism and in forensic linguistics.

1.2 Project Description

As the availability of electronic documents and performance of text processing, a various methods of acknowledge the author of a disputed text has become possible. This project will develop a system to identify authorship on group of anonymous texts.

The dataset available is the one provided for the PAN 2012 competition[5]; Author Identification task. There are two tasks to solve. One is traditional closed and open-class problems. With the closed-class, the task is to identify and classify who is the author of an anonymous text. But with the open-class, the actual author can not be exist in the set of candidate authors. The other task is Authorship clustering. With a given text of mixed authorship, the system have to classify which paragraph is from which authors. In this case, it requires to cluster the paragraphs into two clusters; main author who might has authorship and another authors.

1.3 Literature review

1. A survey of Modern Authorship Attribution Methods(2008)

Stamatatos(2008) states that there are main methodological limitations during accuracy evaluations in the early period of the studies. For example, the data is too long (e.g. whole books) or not written in unified style and hard to compare among different methods because of the lack of suitable data. Also, the evaluation method was mainly visual inspection. However, a research on Authorship attribution moves forward as the research in machine learning, information retrieval and natural language processing increases. Also, since the late 1990s the huge amount of electronic texts was generated through the Internet media usage such as emails and blogs.

According to the paper, Authorship Identification problem has been presented in text representation and text classification. Also it is multi-class single-label text categorisation task in a machine learning. There are several tasks related to authorship analysis. For instance, Author verification, Plagiarism detection, Author profiling and Detection of stylistic inconsistencies.

As the project is about identify real writer in anonymous text, the features of writing style which is called stylometric features are essential things to consider. Stamatatos(2008) states three current review of these features. Firstly, lexical and character features which is intuitive methods. Secondly, synthetic and semantic features for more detailed linguistic analysis. Lastly, application-specific features which can be structural, content-specific or language-specific feature.

Chapter 2

Analysis

2.1 possible techniques

(1) Text classification

Information Retrieval - tf-idf structure with support vector machines.

(2)Word Embedding Word embedding - a method that represent word to vector usually used to translate sparse representation to dense representation.

Word2Vec - As vectorise words, it is able to calculate the similarities between words. In word2vec there is CBOW(Continuous bag of words) and skip-gram, two methods.

(3) Featrue Extraction

Stylometry - Features such as number of sentences in an article, words in an article, average number of words in a sentence, commas, colons etc can be used together with one classifier.

Vocabulary Diversity - measuring the diversity of authors' vocabulary.

Bag of Words and frequency of function words - All words can be used as a vector; vector space model. The function words such as particle, pronoun, conjunction are used as features.

2.2 References

1. Stamatatos, E. (2008). A survey of Modern Authorship Attribution Methods. (online) Available at: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.21001 (Accessed 10 Oct. 2019).

2. Bozjurt, I., Baglioglu, O. and Uyar, E. (2007). Authorship Attribution (Accessed 11 Oct. 2019).

3. Juola, P. (2012). An Overview of the Traditional Authorship Attribution Subtask. (online) Available at: https://pan.webis.de/clef12/pan12-

web/author-identification.html (Accessed 11 Oct. 2019).