Ben-Gurion University

Faculty of Natural Sciences

Department of Computer Sciences

Authorship Attribution in Modern Hebrew

In partial fulfillment of requirements
for Master of Science Degree

Presented By David Gabay
Abstract

This thesis deals with a text classification problem: the identification of the author of a text by its style. Given a text whose author is unknown, and a set of candidates with sample texts, we need to find the true author of the text. The authorship attribution problem has usages in the humanities, in forensic linguistics and in intelligence.

The corpora on which this study was done are written in Modern Hebrew. Hebrew presents many challenges for this task: complex morphology, high level of ambiguity and general lack of processing tools. The corpus at the focus of the research is comprised of posts from web-blogs. The texts in this corpus are shorter than in most authorship attribution studies: most of them are a few hundred words long, the shortest - only a few dozens words long. A secondary corpus was comprised of literary works from early Modern Hebrew literature. Each corpus contained text by nine different authors.

Two general approaches were taken. The first considers a text as a sequence of characters, and uses general methods from information theory to compute distances between texts. In this approach, two methods were employed: one that is based on Markov chains and another that is based on text compression. In the second approach texts are represented as vectors of frequencies of linguistic elements, and are then classified using machine learning algorithms. The features that were tried in this work were lexical (prominent words, whose selection can be done in different ways) and morphological – parts of speech, or other morphological characteristics such as the construct state. The machine learning tool that was used was SVM.

We have found the second approach performs better, reaching an accuracy level of over 98% on the literature corpus and 74% on the blog corpus. This approach was also less affected by variations in the topic of the texts. The first approach, which does not separate content from style, failed to classify texts written by authors on subjects that are atypical for them.

Both approaches did not perform well on very short texts. Attempting to solve this by choosing only features that exist in the anonymous texts did not yield good results.

Combining lexical and morphological features considerably improved the classification, even though the morphological disambiguation tool that was used is not
perfect. The relatively simple measure of segmenting the prefixes had a positive effect, in any variation of the second approach.

The study has demonstrated that authorship attribution in Hebrew is achievable in texts of sufficient length and that style identification requires morphological analysis.
# Table of Contents

1. Introduction .................................................................................................................. 3  
2. Objects and Assumptions .............................................................................................. 5  
3. Characteristics of the Corpora and Evaluations of the classification ...................... 6  
4. Difficulties in the Task of Authorship Attribution ................................................... 8  
5. Text Representation and Learning Methods ............................................................. 10  
   5.1. Hebrew Text Representation ................................................................................. 11  
   5.2. Learning Methods ................................................................................................. 11  
6. Previous Works ........................................................................................................... 13  
   6.1. Early Works and Stylometry in the Humanities ................................................... 13  
   6.2. Advanced Studies ................................................................................................. 14  
7. Methods based on Information Theory ....................................................................... 17  
   7.1. Markov Chains (Khmelev Method) ...................................................................... 17  
      7.1.1. Method Description and Previous Works ..................................................... 17  
      7.1.2. Mathematical Background ............................................................................ 18  
      7.1.3. Linguistic Reasoning ...................................................................................... 19  
      7.1.4. Previous Results ............................................................................................. 20  
      7.1.5. Results ............................................................................................................ 21  
      7.1.6. Analysis .......................................................................................................... 24  
   7.2. Attribution by Text Compression ........................................................................... 26  
      7.2.1. Method Description and Previous Works ..................................................... 26  
      7.2.2. Mathematical Background ............................................................................ 27  
      7.2.3. Results ............................................................................................................ 28  
8. Parts of Speech, Function Words and Content Words ............................................. 30  
   8.1. Text Representation .............................................................................................. 30  
   8.2. Learning Method – SVM ...................................................................................... 32  
      8.2.1. Multiclass classification ............................................................................... 33  
   8.3. Experiments, results and Analysis ....................................................................... 35  
      8.3.1. Bigrams .......................................................................................................... 35  
      8.3.2. Non-Linear Kernel Functions ....................................................................... 35  
      8.3.3. Ben-Yehuda Corpus ......................................................................................... 36  
      8.3.4. Function Words Selection by the Anonymous Text .................................. 38  
9. Conclusion .................................................................................................................. 39