Authorship Detection in the Romanian Language

Chibici Tiberiu Alexandru

June 3, 2017

Contents

Abstract

Authorship detection is an important problem in natural language processing, which involves attributing a document to an author through automated means. In often cases, such as in archeology, the author might not be known, but building a profile of the potential author could prove useful. This paper presents the implementation of an authorship detection system, tailored for the Romanian language.

Notice This document is a draft. Some information might be wrong or incomplete.

1 Introduction

Authorship attribution is an important problem in natural language processing. It has applications in various fields, such as in archeology, journalism, or even law (knowing the source of a ransom note might help save lives).

There are several problems that authorship detection involves:

1. Identifying the author of a text: in this case, we have a database of authors and some of their works. In this case, given an unknown text, we are trying to attribute it to a known author.

2. Profiling the author of a text: in this case, the author is unknown. We want to build a profile of the author based on the text. This means finding things such as the ones enumerated below. Of course, some of these attributes might be impossible to find just from a single text, but we might be able to give a rough estimate.

   (a) the time when the author lived

   (b) an estimate of the age of the author when he wrote the text

   (c) the location where he was born or where he lived, whether he came from a rural or urban background

   (d) the gender of the author

   (e) what kind of studies the author had

   (f) what occupations the author had

3. Verifying authorship of a text: we want to verify if a text is correctly attributed to a specific author (for example, the authorship of some of the books in the *Bible* is contested. Traditionally, the *Letter to the Hebrews* is attributed to *Paul*, but most researchers reject this view). Also, we want to detect instances of plagiarism.

In the following sections we will look into each of these problems in greater detail. I will present a survey of existing research on this topic in section .... In section ...I will present my own research and methodology.

Add chapter number here

This introduction could be improved.

Same here

2 Previous work

This is a test [1].

3 Contents

4 Conclusion

References

[1] Sample D. Sample. The Comprehensive Tex Archive Network (CTAN). *TUGBoat*, 14(3):342–351, 1993.

Things left to do