

Using Compression to Identify Classes of Inauthentic Texts

Mehmet M. Dalkilic, Wyatt T. Clark, James C. Costello, Predrag Radivojac*

{dalkilic, wtclark, jccostel, predrag}@indiana.edu

School of Informatics, Indiana University, Bloomington, IN 47408

Abstract

Recent events have made it clear that some kinds of technical texts, generated by machine and essentially meaningless, can be confused with authentic, technical texts written by humans. We identify this as a potential problem, since no existing systems for, say the web, can or do discriminate on this basis. We believe that there are subtle, short- and long-range word or even string repetitions extant in human texts, but not in many classes of computer generated texts, that can be used to discriminate based on meaning. In this paper we employ universal lossless source coding to generate features in a high-dimensional space and then apply support vector machines to discriminate between the classes of authentic and inauthentic expository texts. Compression profiles for the two kinds of text are distinct—the authentic texts being bounded by various classes of more compressible or less compressible texts that are computer generated. This in turn led to the high prediction accuracy of our models which support a conjecture that there exists a relationship between meaning and compressibility. Our results show that the learning algorithm based upon the compression profile outperformed standard term-frequency text categorization on several non-trivial classes of inauthentic texts. Availability: <http://www.informatics.indiana.edu/predrag/fsi.htm>.

1 Introduction

When operating over a corpus of text there is a natural presumption that the text is meaningful. This presumption is so strong that neither the tools, like webpage search engines, nor the people who use them take into account whether, for example, a webpage conveys any meaning at all, even though the number of indexable webpages available is so large and growing [4]. And yet, a web search for the nonsensical sentence, “Colorless green ideas sleep furiously,” yields scores of thousands of hits on Google, Yahoo, and MSN. Of course this is no ordinary sentence—it is Noam Chomsky’s famous sentence that he constructed to illustrate that grammar alone cannot ensure meaning [10]. While the sentence is syntactically correct and can be parsed, it does not

possess any real meaning. But the important point is that the sentence *is* meaningless and has become part of the searchable text indistinguishable from any other sentence.

Single sentences can seldom convey enough meaning and are therefore combined into texts or documents to provide some larger, more complex information. According to linguists, texts exhibit not only sentential structure, but also higher levels of structure, for example, the so-called *expository structure* that are meant to be *informative*, that is, scholarly, encyclopedic, and factual as opposed to, say, those intended for entertainment. These higher level distinctions can be somewhat problematic if taken too literally, but are useful nonetheless. We can take other perspectives too: there are global patterns that are *only* manifested when the text is examined in its entirety. For example, one kind of global text pattern is the adherence to a *topic*. Another example is *discourse*—the different kinds of meaning derived solely from the arrangement of sentences.

To make clear the class of problem we are interested in examining, we provide the following definitions:

DEFINITION 1.1. *An authentic text (or document) is a collection of several hundreds (or thousands) of syntactically correct sentences such that the text as a whole is meaningful. A set of authentic texts will be denoted by A with possible sub- or superscripts.*

DEFINITION 1.2. *An inauthentic text (or document) is a collection of several hundreds (or thousands) of syntactically correct sentences such that the text as a whole is not meaningful. A set of inauthentic texts will be denoted by \mathcal{I} with possible sub- or superscripts.*

Now consider a scenario in which inauthentic texts are not human generated, but are automated and embellished further with figures, citations, and bibliographies. Without dedicated human scrutiny such texts can escape identification and easily become part of searchable texts of cyberspace. Such a scenario recently played out when an automated inauthentic text was accepted to a conference without formal review, although we are not aware of the mechanism that led

*To whom correspondence should be addressed.

