Future Short Term Goals of Research in Computational Analysis of Stylistics in Text

Shlomo Argamon, Jussi Karlgren and James G. Shanahan

1 Introduction

The first workshop on stylistic analysis of text for information access was held on the day following the 2005 SIGIR conference. This workshop addressed the automatic analysis and extraction of stylistic aspects of natural language texts. Style, roughly defined as the 'manner' in which something is expressed, as opposed to the 'content' of a message is usually disregarded by information access applications as having no bearing on the target notion of relevance: systems have typically focused on the "factual" aspect of content analysis.

The goal of improving the textual analysis of information access systems is a motivating factor for stylistic research. In addition, readers, authors, and information specialists of whatever persuasion are aware of stylistic variation. This provides us with the added philological motivation for research: that of understanding text, readers, and authors better.

The program for this workshop was tight and full of presentations: this was an exploratory meeting, with presentations ranging extensively across various examples of non-topical analysis of text. The data sets used, the features extracted, the target dimensions aimed at, and the computational schemes employed varied widely, attendant to the impressive variation in application.

Speaking generally, taking first steps in stylistic analysis of text is quite easy:

- select computable textual features;
- combine them judiciously;
- model the choice space;
- compare results from measurements on texts under consideration to some norm or norms.

This is a process which is familiar to any practitioner of information access research. The challenge, returning to the motivations mentioned above, is to ensure that the analysis has reliable predictive power for the application under consideration, and that the results have adequate explanatory altitude to provide purchase for further study and generalization.

2 Evaluation and Application

Evaluation was naturally at the forefront of the presentations. The various application areas motivated several different approaches to evaluation, from the relatively clear case of authorship attribution and forensic applications to the less clear cut ones one of mood classification of blog posts. For any information access application, the evaluation must be both operationally quantifiable and related to some formalization of user needs --- one of the projects presented explicitly gathered user opinions for an information retrieval system which utilized stylistic analysis for presentation of results.

3 Feature Rally

The crucial methodological difference between stylistic analysis and topical information retrieval is that of feature extraction. The features studied are different than those studied in topical analysis of text -- in the workshop we addressed this in a Feature Rally session, where participants were invited to present their favourite feature in a few minutes.

4 Common Resources

To better synchronize the efforts of resources, the workshop decided to establish a clearinghouse for common resources, a mailing list, and a bibliography of previously published research. First and foremost, the proceedings of this first workshop on textual stylistics in information access will be made publicly available.

Any interested parties are welcomed to contact the organizers for further information!

5 Future Meeting

At the end of the proceedings, the workshop ended with the consensus that future meetings are in order. A workshop is provisionally planned for the next SIGIR, including a common task for participants to provide an embryo of a common evaluation scheme. Anyone interested in participation should contact the organizers of this years workshop to find out more!