Stylistics for Text Retrieval in Practice
SIGIR 2006 Workshop
Seattle, August 10, 2006

Ozlem Uzuner*, Shlomo Argamon, Jussi Karlgren*****
* University at Albany, State University of New York, USA
** Illinois Institute of Technology, USA
*** Swedish Institute for Computer Science, Sweden
ouzuner@albany.edu, argamon@iit.edu, jussi@sics.se

Abstract

Stylistics for Text Retrieval in Practice has met during SIGIR 2006. With participants from both academia and industry, the workshop spurred interesting discussions on the future of stylistics, its practical uses, and "killer app". The papers presented varied from customer feedback evaluation systems to automatic speech generation.

1 Overview

Recent years have seen an increased attention to various aspects of automatic analysis and extraction of stylistic aspects of natural language texts. This workshop focused on "style", i.e., the manner of expression of the "content" and was open to submissions on modelling, representing, explaining, and utilizing variation in the manner of expression.

Potentially useful applications of stylistic analysis abound, including systems for genre-based information retrieval, authorship attribution, plagiarism detection, context-sensitive text or speech generation systems, organizing and retrieving documents based on their writing style, attitude, or sentiment, quality or appropriateness filters for messaging systems, detecting abusive or threatening language, and more. This year, participants were expected to bring with them a method for applying stylistic analysis to information access tasks. In particular, the organizers encouraged working systems and demonstrations.

Participants were asked to address the following key challenge question in their participation proposals:
* WHAT IS A MEANINGFUL "KILLER APP" FOR STYLISTIC TEXT ANALYSIS? *
and to consider the following questions for discussion in session:

1. How does style relate to other forms of non-topical textual variation?
2. What features are best for different style analysis tasks?
3. Is cross-lingual or 'universal' style analysis possible, and if so, how?
4. How might we develop useful shared resources for moving style research forward?

2 Submissions

The presentations in this workshop encompassed authorship attribution [1, 3, 6], customer review evaluation [4, 5], and speech [2]. Akiva et al. presented "a system for insight retrieval out of web pages using style factors for profile and tone classification", and demonstrated that style features compounded with profile and tone information provides the best results for intelligence [1]. Alias et al. presented a text-to-speech system that can generate natural sounding speech in different "speech styles", or moods [2]. Galitsky et al. presented a "Complaint Engine" suit for mediating consumer disputes. The broad range of applications discussed at this meeting demonstrated the relevance of stylistics to a variety of applications. Participants discussed the possibility of expanding their stylistic research to closely related TREC tasks and to the Digital Humanities communities.

3 Answers?

The question that proved easiest to answer was the question on sharing resources. The participants decided to establish a common data and related work repository to serve both the current community and the future entrants of the field.

In answer to the question on cross-linguality and universality, several of the participants demonstrated technology that was mainly language-independent: in spite of the highly language specific instantiations of feature sets, the feature bundles and dimensions of variations were mostly defined in general terms, readily transferable from one language to another for documents that are within the same genre. This, of course, begs further questions; however, the general sense of the workshop participants is that the techniques and features used are generalizable.

The style features that proved most useful varied with the task. The participants demonstrated systems that took advantage of both conventional and idiosyncratic features, including both linguistic knowledge and surface features. The tasks focused on different types of analysis: obtaining knowledge about text [2] and author [1, 3, 4, 6], and extracting opinions and review information from text [5].

This led us to the question "What is a meaningful killer app?" — the main challenge question of the workshop — which remains unanswered. Several different applications were demonstrated at the workshop and many more mentioned. Given the varying applications, technologies, and research backgrounds of the participants, no single answer could be established. The resolution of this issue will be deferred to future events!

4 References

[1] Navot Akiva, Johnathan Schler. *TrendMine: Utilizing Authorship Profiling and Tone Analysis in Context*. In Proceedings of the ACM SIGIR Workshop on Stylistics for Text Retrieval in Practice. August 2006.

[2] Francesc Alias, Xavier Sevillano, Joan Claudi Socoro. *Text Classification based on Associative Relational Networks for Multi-Domain Text-to-Speech Synthesis.* In Proceedings of the ACM SIGIR Workshop on Stylistics for Text Retrieval in Practice. August 2006.

[3] Shlomo Argamon, Moshe Koppel, James Pennebaker, Jonathan Schler. *A Tool for Automated Authorship Attribution.* In Proceedings of the ACM SIGIR Workshop on Stylistics for Text Retrieval in Practice. August 2006.

[4] Boris Galitsky, Boris Kovalerchuk. *Analyzing Attitude in Customer Emails: A Tool for Complaint Assessment.* In Proceedings of the ACM SIGIR Workshop on Stylistics for Text Retrieval in Practice. August 2006.

[5] Xiao Hu, J. Stephen Downie. *Stylistics in Customer Reviews of Cultural Objects.* In Proceedings of the ACM SIGIR Workshop on Stylistics for Text Retrieval in Practice. August 2006.

[6] George Mikros. *Authorship Attribution in Modern Greek Newswire Corpora.* In Proceedings of the ACM SIGIR Workshop on Stylistics for Text Retrieval in Practice. August 2006.