Explanation in Computational Stylometry

Walter Daelemans

CLiPS, University of Antwerp, Belgium
walter.daelemans@ua.ac.be

Abstract. Computational stylometry, as in authorship attribution or profiling, has a large potential for applications in diverse areas: literary science, forensics, language psychology, sociolinguistics, even medical diagnosis. Yet, many of the basic research questions of this field are not studied systematically or even at all. In this paper we will go into these problems, and suggest that a reinterpretation of current and historical methods in the framework and methodology of machine learning of natural language processing would be helpful. We also argue for more attention in research for explanation in computational stylometry as opposed to purely quantitative evaluation measures and propose a strategy for data collection and analysis for achieving progress in computational stylometry. We also introduce a fairly new application of computational stylometry in internet security.

1 Meta-knowledge Extraction from Text

The form of a text is determined by many factors. Content plays a role (the topic of a text determines in part its vocabulary), text type (genre, register) is important and will determine part of the writing style, but also psychological and sociological aspects of the author of the text will be sources of stylistic language variation. These psychological factors include personality, mental health, and being a native speaker or not; sociological factors include age, gender, education level, and region of language acquisition.

Writing style is a combination of consistent decisions in language production at different linguistic levels (lexical choice, syntactic structures, discourse coherence, ...) that is linked to specific authors or author groups such as male authors or teenage authors. It remains to be seen whether this link is consistent over time and whether there are style features that are unconscious and cannot be controlled, as some researchers have argued. The basic research question for computational stylometry seems then to describe and *explain* the causal relations between psychological and sociological properties of authors on the one hand, and their writing style on the other. These theories can be used to develop systems that generate text in a particular style, or perhaps more usefully, systems that detect the identity of authors (authorship attribution and verification) or some of their psychological or sociological properties (profiling) from text.

A limit hypothesis arising from this definition is that style is unique for an individual, like her fingerprint, earprint or genome. This has been called the *human stylome hypothesis*:

A. Gelbukh (Ed.): CICLing 2013, Part II, LNCS 7817, pp. 451–462, 2013.

> '(...) authors can be distinguished by measuring specific properties of their writings, their stylome as it were.' [1]

Reliable authorship attribution and profiling is potentially useful in many areas: literary science, sociolinguistic research, language psychology, social psychology, forensics, medical diagnosis (detecting schizophrenia and Alzheimer's), and many others. In Sect. 3 we describe results in the context of an internet security case study as an example of useful computational stylometry. However, the current state of the art in computational stylometry seems not advanced enough to always guarantee the levels of reliability expected.

There are many excellent introductions to modern computational methods in stylometry [2–5] describing the methods and feature types used. Feature types include simple character n-grams, punctuation, token n-grams, semantic and syntactic class distributions and patterns, parse trees, complexity and vocabulary richness measures, and even discourse features.

Computational stylometry should be investigated in a Natural Language Processing (NLP) framework, more specifically as one of three levels of text understanding. The goal of text understanding is to extract knowledge from text and present it in a reusable format. NLP has seen significant progress in the last decade thanks to a switch to statistical and machine learning based methods in research and increased interest because of commercial applicability (Apple's SIRI and Google translate are only two examples of recent high impact commercial applications of NLP). The three types of knowledge we distinguish that can be extracted from text are: (i) objective knowledge (answering the who, what, where, when, ... questions), (ii) subjective knowledge (who has which opinion about what?), and (iii) metaknowledge (what can we extract about the text apart from its contents, mainly about its author?). Computational stylometry belongs in the latter category.

Core research in NLP addresses the extraction of objective knowledge from text: which concepts, attributes, and relations between concepts can be extracted from text, including specific relations such as causal, spatial and temporal ones. Research is starting also on the Machine Reading loop (how to use background knowledge in text analysis and conversely how to build up background knowledge from text). See work on Watson for state of the art research at this first level [6]. In addition to the extraction of objective knowledge, the large amount of text produced in social networks has motivated research to focus also on the extraction of subjective knowledge (sentiment and opinion). Never before have so many non-professional writers produced so much text, most of it subjective and opinionated (reviews, blogs, e-mail, chat, ...) [7]. Extraction of *metaknowledge* is conceptually a different type of knowledge extraction from text than the other two types. Where objective and subjective knowledge extraction try to make explicit and structure knowledge that is present in unstructured textual information, metaknowledge concerns knowledge about the author of the text (psychological and sociological properties, and ultimately identity), so outside the text. Recent advances in knowledge extraction from text at all these three levels have been made possible thanks to the development of robust and fairly