PHYS.ORG

Computer scientists quantify elements of writing style that differentiate successful fiction

January 9 2014

Dr. Choi and her colleagues in the College of Engineering and Applied Sciences—Vikas Ashok, a teaching assistant in the Department of Computer Science, and Song Feng, a fifth year PhD student in the same department.

(Phys.org) —Imagine the challenge publishers face, pouring over thousands of manuscripts to determine if a book will be a hit. Stony Brook Department of Computer Science Assistant Professor Yejin Choi thinks she has a tool to bring some science to that art, and she is co-author of a paper, Success with Style: Using Writing Style to Predict the Success of Novels, which was unveiled at the conference on Empirical Methods in Natural Language Processing (EMNLP) 2013.

"Predicting the success of literary works poses a massive dilemma for publishers and aspiring writers alike," Choi said. "We examined the quantitative connection between writing style and successful literature. Based on novels across different genres, we investigated the predictive power of statistical stylometry in discriminating successful literary works, and identified the stylistic elements that are more prominent in successful writings."

Statistical stylometry is the statistical analysis of variations in literary style between one writer or genre and another. The study reports, for the first time, that the discipline can be effective in distinguishing highly successful literature from its less successful counterpart, achieving accuracy rates as high as 84%.

For example, the research indicated that more successful books make more frequent use of discourse connectives (conjunctions such as "and", "but", "or") to join sentences and prepositions. Prepositions, nouns, pronouns, determiners (words that precede nouns to indicate whether the noun is specific or general, e.g. "your letter"), and adjectives are also predictive of highly successful books. Less successful books are characterized by a higher percentage of verbs, adverbs, and foreign words. They also rely more on topical words that could be almost cliché ("love"), typical locations, and extreme ("breathless") and negative ("bruised") words.

Dr. Choi and her colleagues in the College of Engineering and Applied Sciences—Vikas Ashok, a teaching assistant in the Department of Computer Science, and Song Feng, a fifth year PhD student in the same department—found that the less successful books also rely on verbs that explicitly describe actions and emotions ("wanted", "took", "promised", "cried", "cheered"), while more successful books favor verbs that describe thought-processing ("recognized", "remembered") and verbs that simply serve the purpose of quotes ("say").

For practical purposes, the researchers defined "success" by download counts from Project Gutenberg, which houses 42,000 books that are available for free download in electronic format. Dr. Choi and her team scrutinized eight genres—adventure, mystery, historical fiction, fiction, science-fiction, love stories, short stories, and poetry. They also studied a number of books not included at Project Gutenberg, ranging from A Tale of Two Cities by Charles Dickens, through The Old Man and the Sea by Ernest Hemingway, to The Lost Symbol by Dan Brown.

"For a small number of novels, we also considered award recipients—such as Pulitzer and Nobel prizes—and Amazon sales records in order to define a novel's success," Choi says. "Additionally, we extended our empirical study to movie scripts, where we quantified a film's success based on the average review scores at imdb.com."

The researchers took 1000 sentences from the beginning of each book. They performed systematic analyses based on lexical and syntactic features that have been proven effective in Natural Language Processing (NLP) tasks such as authorship attribution, genre detection, gender identification, and native language detection.

"To the best of our knowledge, our work is the first that provides quantitative insights into the connection between the writing style and the success of literary works," Choi says. "Previous work has attempted to gain insights into the 'secret recipe' of successful books. But most of these studies were qualitative, based on a dozen books, and focused primarily on high-level content—the personalities of protagonists and antagonists and the plots. Our work examines a considerably larger collection—800 books—over multiple genres, providing insights into lexical, syntactic, and discourse patterns that characterize the writing styles commonly shared among the successful literature."

The research, supported in part by the Stony Brook University Office of the Vice President for Research, and in part by gift from Google, not only achieved up to 84% accuracy rates in the novel domain and 89% accuracy in the movie sphere. Says Dr. Choi, "It sets forth an understanding of the connection between successful writing style and readability. We also shed light on the connection between sentiment/connotation and literary success, and put forward comparative insights between successful writing styles of fiction and nonfiction."

And if you've had difficulty getting through those dense but highly rated literary tomes, Professor Choi's findings have your back. "We made an unexpected observation on the connection between readability and the literary success—that they correlate into the opposite directions," she says.

Dr. Choi is quick to point out that these findings only demonstrate correlation, not causation, adding, "We conjecture that the conceptual complexity of highly successful literary work might require syntactic complexity that goes against readability."

More information: Success with Style: Using Writing Style to Predict the Success of Novels, by Vikas Ashok, Song Feng, and Yejin Choi, Conference on Empirical Methods in Natural Language Processing (EMNLP) 2013

Provided by Stony Brook University

Citation: Computer scientists quantify elements of writing style that differentiate successful fiction (2014, January 9) retrieved 23 April 2024 from https://phys.org/news/2014-01-scientists-quantify-elements-style-differentiate.html