1-1-2021

Lost in Translation: What Linguistic Measurements Best Measure Text Quality of Online Listings

Yang Sun
California Northstate University

Shaonan Tian
San Jose State University, shaonan.tian@sjsu.edu

Ming Zhou
San Jose State University, ming.zhou@sjsu.edu

Recommended Citation

Yang Sun, Shaonan Tian, and Ming Zhou. "Lost in Translation: What Linguistic Measurements Best Measure Text Quality of Online Listings" *Procedia Computer Science* (2021): 1474-1477. https://doi.org/10.1016/j.procs.2022.01.187

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 199 (2022) 1474–1477

Procedia Computer Science

www.elsevier.com/locate/procedia

The 8th International Conference on Information Technology and Quantitative Management (ITQM 2020 & 2021)

Lost in Translation: What Linguistic Measurements Best Measure Text Quality of Online Listings

Yang Sun[a], Shaonan Tian[b] and Ming Zhou[b]*

[a]California Northstate University, 9700 W Taron Dr, Elk Grove, CA 95757
[b]San Jose State University, One Washington Square, San Jose, CA 95192

Abstract

Ecommerce websites are filled with international sellers. Product descriptions on these sites are often written in English by non-native speakers. Linguistic imperfections in these descriptions confuse consumers, which may further attenuate their purchase intentions. How descriptive quality/efficacy can be defined and then improved shall be of great interest to all sellers and their consumers. In this research, we attempt to evaluate online product description quality using lexical measurements from linguistics studies. Linguistics measurements of writing quality were mostly developed in pure academic settings. We test and analyze these measurements' applicability in defining and contrasting business description quality using Amazon.com data. Modern classification techniques in the artificial intelligence and machine learning field are deployed in identifying measurement applicability and assessing computational efficiency. Our findings enable automatic identification of descriptive efficacy through artificial intelligence methods on real ecommerce text data.

Keywords: Lexical measurement; online listing; data mining; text mining; listing description

1. Introduction

Characters played by Bill Murray and Scarlett Johansson in the movie "Lost in Translation" enjoyed an atypical yet romantic relationship, which formed fond memories for many of its viewers. However, real life lost-in-translation is about nothing but frustration and confusion. Many e-commerce consumers had to read between

* Corresponding author. Tel: 1-408-924-3572. *E-mail address: ming.zhou@sjsu.edu*. Authors are listed in alphabetic order.

10.1016/j.procs.2022.01.187

the lines of listing descriptions and decipher the true meaning of these texts. Lack of trying is not an accurate description of sellers' efforts. Sellers tried very hard to describe their products in a market using a different language, but actual efficacy bluntly leaves more to be desired. One would easily point to so called translation services when penetrating a foreign market. The problem is that promoting/marketing a product demands much more than straightly replacing a Chinese word with an English word. There are a lot more at work, such as cultural, wording, structure, context and so on. For instance, a good example [9] quoted on Quora.com was about "One of the most famous Christian saints, St. Augustine of Hippo, after undergoing Google Translate from English to Russian, turns into a hippopotamus named Augustine." This is exactly the reason why basic translation services like Google translation are at the best references for business use. People using basic translation services often made their marketing/promotion contents unclear and sometimes laughable. For example, a speaker seller on Amazon.com described its product as "The speakers can create … a spiritual setting that allows you to be more playful, high on spirit or thoughtful." Apparently, the seller didn't want to insinuate speakers can get people drunk, which is exactly what "high on spirit" means to an English speaker.

Linguistics research on writing quality and writings by second language learners has been abundant [10]. Linguists have been motivated to study writing quality by native and non-native speakers within the context of academic writings [2]. For instance, vocabulary has been a central issue for linguistic research [4]. Such research mostly used writing samples from academic writing assignments, where vocabulary, grammar and structure were believed to be building elements of composition quality [5]. In business settings, research on linguistic traits mostly focused on marketing features. [8] studied language characteristics and marketing branding emphasis. They concluded that Chinese-speaking consumers tended to rely on visual cues where English-speaking consumers attended more to phonological cues. Brand naming shall practice sharply different conventions in order to attract Chinese and English speaking consumers. [11] worked on linguistic structural features as well as how such features invoked responses to marketing stimuli such as brand names and advertisements. They studied a feature called "classifier" and related the existence of classifier with branding and advertisement responses. Obviously, linguistic characteristics of a language and lexical differences among languages can affect consumer responses to marketing efforts. Our communication with a marketing company in the Bay Area of California inspired us to wonder if descriptive quality of online listings may also make a difference. This question is increasingly important as many online platforms are filled with international sellers. Tech Crunch [6] estimated that 75% of new sellers on Amazon.com were from China. Online listings start to display wild variations in descriptive quality. Similar to academic writings that strive to convey theoretical points, online listings as compositions also serve a purpose, which is to persuade consumers to take purchase actions [7]. Quality listings shall better communicate with a potential buyer and convince them to buy. A listing written by a non-native seller may miss many linguistic cues that are important to consumer purchase decisions. Before we can explore the connection between listing quality and business outcomes, this research attempts to study how quality of online listings can be measured, as a first step. Given the plethora of linguistic measurements developed to capture lexical characteristics within academic settings, we specifically test what lexical measurements may work better for ecommerce listings. This is a critical question for further research as product descriptions differ vastly from academic writings in purpose and audiences. If and to what extend existing linguistics measurement can be used for assessing writing quality in a business marketplace are fundamental to future model development and implementation.

2. Lexical Measurements and Data Collection

An extensive literature review of linguistics research returned us many options of lexical measurements related to writing quality [3]. Lexical diversity measures different types of vocabulary in comparison to tokens in a text; lexical sophistication measures the usage of less frequently used words; cohesiveness refers to consistency of a text from the perspective of readers' mental process and coherence indicating elements; syntactic complexity is

about ease of processing for a reader, such as shorter but not long sentences. There are many more measurements we can use to capture all subtle aspects of a writing sample. In this work, we attempt to provide empirical evidence and insights on if one, several or a combination of multiple measurements can best differentiate the higher quality from the lower quality for online product listings. We defined a high quality description as a text that is written by a native speaker where a lower quality description tends to deviate from native writing habits and display linguistic imperfections.

We have already started collecting listing samples on Amazon.com. We have developed a procedure to dissect an Amazon listing into several elements, where textual description of a product is of the greatest relevance to this research. Our sample shall be close to three or four hundred listings varying across three different product categories, namely oximeters, vanity mirrors and massage guns. Markets for these products are highly competitive, not dominated by a handful of brands, and full of variations in brands and/or product specifications. These products are also complex enough so that a consumer will need to read descriptions before making purchases. Simultaneously, we have recruited a native speaker team to review a training sample of listing descriptions. Descriptions are evaluated using a rubric agreed upon by all members of our research team. Our team specifically discuss vague cases and refine our rubric criteria at weekly meetings. This training classification will enable us to establish a 1-3 scale for listing nativeness, where 1 is the least native and 3 is the most native.

3. Analysis Methodology

To identify whether and how a lexical measure or a combination of measures can measure linguistic quality of online descriptions, we propose to apply an ordinal logistic regression model with an ordinal categorical response labelled as 1, 2 and 3 on our sample data to classify the listing quality. This supervised machine learning technique uses an interactive-reweighted least-squares algorithm to obtain maximum likelihood estimates of the parameters in the logistic response function $\left(y \le y_j\right) = \frac{1}{1+e^{-(\beta_{0j}+\mathbf{x}'\boldsymbol{\beta})}}$, where y is the response (dependent) variable, and y_j is the j^{th} possible value of y, j = 1,2,3. Thus, $P\left(y \le y_j\right)$ is simply the probability that a listing is perceived to be of category y_j or lower for the nativeness measure. **x** is a vector of lexical measures as independent variables selected in the model. β is a vector of parameters (coefficients) in correspondence with the selected variables; $\beta_{0,j}$ is a constant for each j. We test the significance of the impact of a certain lexical measure or a combination of measures on nativeness and once a model is established, we can estimate $P\left(y = y_j\right) = P\left(y \le y_j\right), j = 1$ and $P\left(y = y_j\right) = P\left(y \le y_j\right) - P\left(y \le y_{j-1}\right), j = 2,3$ based on the given lexical measures.

Modern techniques in the field of machine learning have quickly gained popularity and they are now widely adopted in practice. One of the major advantages is that these techniques tend to make little or no assumption on the statistical properties of a dataset. In this research, we also propose to introduce a random forest classifier to evaluate the relationship between multiple lexical measures and the linguistic quality of online listing descriptions. Random forest classifier generates tree models using randomly selected features on different datasets from bootstrapping methods. Given the fact that many lexical measures are closely related with each other, we are optimistic that the random forest model may be able to bring improved classification performance.

4. Expected Findings

Studies on analyzing textual and lexical features have been voluminous in the field of natural language processing and artificial intelligence. [12] analyzed text data to identify cyber criminals by examining their message texts. [1] dealt with the authorship problem by outlining text features that were most relevant to author profiling, including identifying language nativeness. This profiling was helpful for police to identify suspects of crimes. A recent study [7] analyzed listing characteristics on a Japanese ecommerce website. Their research

deployed a limited amount of lexical measurements and established some traits of listings for more effective information.

Our findings will shed light on how lexical measurements can be applied in evaluating listing descriptions on ecommerce websites. To the best of our knowledge, this is the first research that endeavors to quantify descriptive efficacy using modern methods in the field of AI and potentially automate a process to differentiate native descriptions from non-native descriptions using online listing data. As a survey of more than three hundred consumers conducted by the Bay Area marketing company suggested, 80% of US consumers in their sample preferred to shop from listings that looked native and better described products. Our research shall greatly enhance our understanding of what native description may mean and parameterize defining characteristics of such descriptions. Different from [7], our findings provide more insights in English environments with a much more comprehensive set of lexical measurements.

Acknowledgements

The authors wish to extend our special thanks to Springfield Bay Inc. for their generous support of data access. Their professional insights in Amazon listing and platform dynamics greatly helped this research.

The authors would like to thank Ms Amanranta A. Quintero, Mr. Seth Tang and Ms. Keying Mao for their assistance in data collection and lexical measurement implementation.

References

[1] Argamon, S., M. Koppel, J. W. Pennebaker and J. Schler, Automatically profiling the author of an anonymous text. Communications of the ACM 2009, 52(2), 119 – 123.
[2] Coxhead, A. Academic vocabulary, wiring and English for academic purposes: Perspecives from second lanaugae leaners. RELC Journal 2012, 43(1), 13y7-145.
[3] Crossley, S. Linguistic features in wriing quality and development: An overview. Journal of Writing Research 2020, 11(3), 415-443.
[4] Gardner, D. Exploring vocabulary: Language in action. 2013. New York, NY: Routledge.
[5] Laufer, B. Vocabulary and writing. In C. A. Chapelle (Ed.), The encryclopedia of appllied linguistics 2013, Joh Wiley and Sons, Inc.
[6] Liao, R. Prime today, gone tomorrow: Chinse products get pulled from Amazon. TechCrunch May 11th, 2021.
[7] Pryzant, R., Y.J., Chung and D. Jurafsky. Predicting sales from the language of product descriptions. Proceedings of the SIGIR 2017 eCom Workshop, Tokyo, Japan, August 2017.
[8] Schmitt, B. H., Y. G., Pan and N. T. Tavassoli. Language and consumer memory: The impact of linguistic differences between Chinese and English. Journal of Consumer Research 1994. Vol 21, 419 – 431.
[9] Tereshchenko, Al. What are some of the best (and worst) Google translate translations?. Quora.com, May 13th, 2019.
[10] Xu, X. L., Analyses and comparison of three lexical feasures in native and nonnative academic english wring, Electronic Theoses and Dissertation 2018, University of Central Florida.
[11] Zhang, S. B. H., Schmitt and H. Haley. Lanugage and culture: Linguistic effects on consumer behavior in international marketing research. Handbook of Research in International Marketing 2003, 228 – 242.
[12] Zheng, Y. Q., Y. Qin, Z. Huang, and H. Chen. Authorship analysis in cybercirme investigation. International Conference on Intelligence and Security Information 2003, 59 – 73.