The Hidden Shape of Stories

Reveals Positivity Bias and Gender Bias

Huimin Xu[1], Da Xiao[2], Lingfei Wu[3], Cheng-Jun Wang[1*]

[1] Computational Communication Collaboratory, School of Journalism and Communication, Nanjing University, Nanjing 210093, P.R. China.
[2] National Engineering Laboratory for Disaster Backup and Recovery, Computer Science and Technology, Beijing University of Post and Telecommunications, Beijing 100876, China.
[3] Department of Sociology, University of Chicago, 1126 E 59th St, Chicago, IL 60637.
* Corresponding author. E-mail: wangchj04@126.com (C.W)

Abstract: To capture the shape of stories is crucial for understanding the mind of human beings. In this research, we use word emdeddings methods, a widely used tool in natural language processing and machine learning, in order to quantify and compare emotional arcs of stories over time. Based on trained Google News word2vec vectors and film scripts corpora (N =1109), we form the fundamental building blocks of story emotional trajectories. The results demonstrate that there exists only one universal pattern of story shapes in movies. Furthermore, there exists a positivity and gender bias in story narratives. More interestingly, the audience reveals a completely different preference from content producers.
Keywords: Story shape, story bias, story success, word emdedding, emotional arc

Introduction

The communication power of stories to transfer information has been shown over time and thus seeking to better understand story narratives is very essential (Campbell, 2008; McKee, 1997). To be specific, In Kurt Vonnegut's rejected master's thesis, he defined the emotional arc of a story on the "beginning-end" and "ill fortune-great fortune" axes (1981, 1995). Inspired by this idea, Peter and his colleagues (2011, 2015, 2016, 2016) identified the emotions of the stories over time using sentiment analysis methods, and they found six core emotional arcs which form the essential building blocks of complex story trajectories. They also confirm that there is a positivity bias in human languages (Peter et al., 2014).

In this study, we employ word emdeddings method to analyze the film scripts. First, we use two methods to validate our findings that there is only one main story trajectory, rather than six kinds of different story shapes. That is to say, there is a universal shape of stories which dominates the narrative of movies. Second, we extend our findings from the perspective of story bias in terms of

gender and emotion. Third, we offer a new insight to look upon the success of story in films by analyzing the ratings of the movies.

One of the most common approaches to computational social science is to develop predictive models. However, computational methods can not only be used for prediction, but also for explanation and understanding. In this research, we want to develop a narrative that helps us to understand qualitatively how cultural production reflects and shapes our social world.

Methods

Data

We have collected around more than one thousand (N = 1109) film Scripts from imsdb.com and relevant information, such as film genre (shown in figure 1) and user rating score. All of our codes are available publicly online and the data can be crawled from the website of imsdb (https://www.imsdb.com). Therefore, it is very convenient to reproduce our findings.

Insert Figure 1 here

Measures

We apply the word embeddings as a quantitative lens through which to study emotional arcs. Word embedding is a powerful machine-learning framework that represents each word by a vector. The geometric relationship between these vectors can capture meaningful semantic relationships between corresponding words. Vectors being closer have been shown to correspond to more similar words (Collobert et al., 2011).

In this paper, we use Garg's method (Garg et al., 2018) to demonstrate how temporal dynamics of the embeddings helps to quantify changes in emotional narratives of stories. The emotional arc construction includes three steps.

1. First, we choose several word lists to represent emotional arcs, one side fortune ('success', 'succeed', 'lucky', 'fortunate', 'smile', 'happiness') and the other side tragedy ('failure', 'fail', 'unlucky', 'unfortunate', 'tear', 'sad').
2. Second, we measure the average emdedding distance that represents fortune and tragedy separately, using Google News word2vec vectors trained on the Google News dataset (Mikolov et al., 2013; Mikolov et al., 2013). If the average distance for fortunate side minus the average distance for unfortunate one is greater than 0, then the emotion is more positive.
3. Third, we average the stories into 10 (or 100, 200) parts based on word counts and repeat the step above on every part separately in order to get the temporal patterns of emotion.

After identifying the trajectory for each story, we regularize them with Z-score, namely, z = (s' – s(mean))/ s(sd), where s' is each story's emotional time series, s(mean) and s(sd) are the mean and standard deviation of emotional trajectories for all considered stories. Figure 2 is an emotional arc visualization