Motivation

- Literary grace (文采) is regarded as one of the elements constituting the style and quality of articles in China.
- Literary grace also has attracted scholars in the natural language processing (NLP) field. Many studies have referred to several aspects of it and been applied in real-world applications, including Automated Essay Evaluation (AEE), quality assessment, machine translation, opinion mining, dialogue modeling, and modeling argumentative discourse.
- There have been few publicly available corpora annotated with literary grace.
- Constructing a corresponding larger-scale, multi-level, high-quality, and adaptable corpus is a matter of concern.

Contributions

- We provide a relatively large-scale Chinese corpus with literary grace annotation, containing multi-level labels. The corpus can be used to study the variations of literary grace in literature, linguistics, and natural language processing.
- We analyze the corpus and explore linguistic features that affect readers’ aesthetic judgments from three aspects: word, sentence, and paragraph levels.
- We propose a new task, named Literary Grace Evaluation, which aims to make a comprehensive assessment of the literary grace level. Then we classify the corpus with machine learning models. The result shows that the features we chose are very effective for the LGE task.
- We further analyze the corpus from three aspects to find out linguistic features that affect Chinese readers’ aesthetic judgment.

Experiments

- We validate the effectiveness of the features mentioned above on the Literary Grace Evaluation task, using SVMs as the classifier to classify the corpus automatically. Our experiment is based on 8,000 training data, 1,000 test data, and 1,000 validation data. The results are presented in Table 4.
- The results prove that all the features are effective for the LGE task. They can improve the model performance to varying degrees. Word-level features have the most significant impact among them. Sentence and paragraph-level features have relatively little influence. However, they also play an essential role in literary grace assessment, further illustrating the necessity of evaluating literary grace in language units above the sentence level.

The Corpus

- The texts in CLGC were divided into three classes according to their literary grace level. The language of segments labeled as “1” is plain and less rhetorical. Higher-level labels represent higher literary grace. Table 1 shows sentences selected from level 3 texts. Each text contains 200 to 300 tokens, and we provide multi-level annotations for every sentence in the text.
- This corpus consists of 10,000 texts from novels and prose, including 64% level 1 texts (6,448 texts, 1,192,144 tokens), 19% level 2 texts (1,959 texts, 396,846 tokens) and 15% level 3 texts (1,593 texts, 294,748 tokens). Compared with the existing corpora, our corpus has the following features:
  1. Focus on the real sense of literary grace, not just limited to the specific features;
  2. Provide multi-level annotations that include literary grace level, sentence category, and figure-of-speech type tags;
  3. Use a simple majority voting scheme to select the most probably literary grace label as the ground truth;
  4. Content and paragraph-level materials varied in the domain.

Analysis

- We further analyze the corpus from three aspects to find out linguistic features that affect Chinese readers’ aesthetic judgment.
- Lexical semantic
  - The words related to machinery, meteorology, natural objects, plant, color, and condition are easier to bring readers high aesthetic experience, e.g., 向日葵 (sunflower), 岁月 (years), 风 (wind). In comparison, the words related to the general term, relatives, society, social contact, address, and interjection are harder to bring this experience, e.g., 他们 (they), 警察局 (police station), 妈妈 (mom).
- Part-of-speech
  - Texts with high literary grace intend to use more descriptive words. Descriptive words are used to describe colors, states, and modalities of people or objects, e.g., 白色 (white), 热气腾腾 (hot), 毛茸茸 (fluffy). They are related to the appearance and shape of people and objects, which can give readers a vivid feeling.
- Type-token ratio
  - The lexical diversity of high literary grace text is richer than low literary grace text.
- Sentence category
  - The use of exclamatory and interrogative sentences can lead to lower literary grace. Table 2 shows the basic statistics of sentence category.
- Figure-of-speech type
  - The figure-of-speech number is highly correlated with literary grace level (r=0.535**, p<0.01). As for the different types of figure-of-speech, BN (Personification and Skeumorphism) is the most correlated with the literary grace level (r=0.440**, p<0.01), then is BY (Simile and Metaphor) (r=0.379**, p<0.01), PB (Parallelism) (r=0.238**, p<0.01) and DO (Contrast) (r=0.132**, p<0.01). The last is TG (Transferred Epithet) (r=0.175**, p<0.01). Table 3 shows the basic statistics of the figure-of-speech types.

Acknowledgements

This work is funded by the Humanity and Social Science Youth foundation of Ministry of Education (19YJCZH230) and the Fundamental Research Funds for the Central Universities in BLCU (No. 21PT04).