Extracting Character Information on the Story

20160413_SoyoungYoon **20140407_SangwonLee** **20160655_MinyeopChoi**

20170357_ShinDongHwan **20130240_KyuminPark**

1 Problem Description

We plan to analyze literature and extract information based on the story. First, we plan to figure out relationships between fictional characters, such as determining hostility between characters, or even family relations. Second, we also plan to conduct a detail analysis of characters, including personality traits, age, gender, characteristics, and personal backgrounds.

2 Target Data

To evaluate the performance of the proposed below method, we try to pick a famous novel with many protagonists and complicated relationships. The title of the book is <Adrift in the Pacific>. In this story, there are 15 adrift boys and the number of relations between characters is ${}_{15}C_2$. In addition, there are many sub characters and more relations. The book's PDF file that we have, is a total of 177 pages and about 52860 words and 248446 characters. To transform the PDF files to text files in python, PyPDF2 library is used.

3 Methodology

Our work separate into three steps; *Substitution*, *Direction Configuration*, and *Relationship Calculation*. Substitution step converts all the proverbs into proper noun so that they can be used as anchor fur further steps. Direction configuration step figures out the provider of action and receiver of action in each of the sentence. If there is no provider or receiver of the action in the sentence, we may ignore it. Relationship calculation step evaluates the emotion or action that provider gives to the receiver, using score-annotated corpus such as `SentiWordnet`. With three steps for entire character in whole book, we can get relationship graph among the characters.

4 Related Work

The field of literary analysis was a long-studied field by many scholars. Compared to (Kim and Klinger, 2019) and (Labatut and Bost, 2019) which only classify the emotional relationships between characters, we plan to also classify family relations. By extending work by (Bamman et al., 2014) and (Pizzolli and Strapparava, 2019), we plan to extract various information about each literary characters.

5 Evaluation Plan

To evaluate our work, we must determine 1) the main character and 2) the relationships between main characters to extract. We decide to give the characters a score based on their number of appearances and defined the relationship score as the average of the scores of two or more characters who get involved in the relationship. Evaluations will consist of the following metrics for the top 50 relationships based on its score.

Let P is

$$\frac{\texttt{(\# correctly extracted relations)}}{\texttt{(Total\# extracted relations)}}$$

Let R is

$$\frac{\texttt{(\# correctly extracted relations)}}{\texttt{(Actual\# extracted entity relations)}}$$

F-Measure $F1$ can be represented

$$\frac{2PR}{P+R}$$

Users who want to get the point of a book with using our model, it is easy to think that precision is just important. However, in order to grasp the whole story in the book, it must also be considered important how much of the actual relationships our model can extract. Since F1 score is the weighted average of Precision and Recall, it takes both false positive and false negative into account, so it is a suitable metric to reflect our intension.

References

David Bamman, Ted Underwood, and Noah A. Smith. 2014. A Bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 370–379, Baltimore, Maryland. Association for Computational Linguistics.

Evgeny Kim and Roman Klinger. 2019. Frowning Frodo, wincing Leia, and a seriously great friendship: Learning to classify emotional relationships of fictional characters. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 647–653, Minneapolis, Minnesota. Association for Computational Linguistics.

Vincent Labatut and Xavier Bost. 2019. Extraction and analysis of fictional character networks: A survey. *ACM Computing Surveys*, 52:89.

Daniele Pizzolli and Carlo Strapparava. 2019. Personality traits recognition in literary texts. In *Proceedings of the Second Workshop on Storytelling*, pages 107–111, Florence, Italy. Association for Computational Linguistics.