Extracting Character Information from Movie Script

20160413 SoyoungYoon **20140407 SangwonLee** **20160655 MinyeopChoi**

20170357 ShinDongHwan **20130240 KyuminPark**

1 Introduction

Vast amount of movies are newly made every year, and there's a clear limitation for us to see all the movies. Also, there is hardly any work done on analysis on movies, in contrast to literature analysis. Therefore, we propose a novel, quantitative analysis on each film. Detailed analysis will help us better understand, analyze, and even compare each movie. Based on the script *frozen* and *beauty and the beast*, we successfully completed 1. Anaphora resolution, and 2. Characteristic extraction based on big-5 personality traits, with reasonable f-score of above 0.7. Then, we 3. extracted relation between characters by hostility, romance and top-down relationship. We also showed changing relation through time. Lastly, we 4. ranked main characters. Detailed results are shown in the appendix and the code is open at https://github.com/okas832/playNLP.

2 Approach

2.1 Preprocessing

This includes the process of extracting text from raw script, and structuring information. There was various types of speech, including conversation, narrator talk, and cutaway information. For handling various type of conversation, we introduce a 3-state finite-state machine to know where the lines start and end, and parse normal talk, narration and sing differently. Our finite state machine has an additional state to handle special cases that two individual speech in one line.

2.2 Listener Resolution

Listener resolution is needed to correctly understand the data from the conversation. The relationship between characters can be seen through the distance between conversation and the time_index variable, which indicates the time & space of conversation. Here we define listener as who speaks in five conversations ahead and back within the same time_index. Listener resolution will use the definition and characters' presence in conversation text.

2.3 Anaphora Resolution

Anaphora resolution is the problem of resolving references to earlier or later items in the discourse. First, gender and number agreement filter is applied. After that, we use Mitkov's resolution system which is introduce in his paper(Mitkov, 1998). There are 10 rules for scoring candidate noun phrase. Since there are some rules unable for our system, we use 8 out of Mitkov's 10 rules. To improve our anaphora resolution system, we need to solve the way to matching unseen noun phrase to anaphora. Also, we need to identify "pleonastic-it", anaphoras used without an antecedent, more correctly. In this particular case *"Hup! Ho! Watch your step! Let it go!"* the word "it" does not indicate any antecedent.

2.4 Personality Extraction

Our goal here is to infer personality of each character with an automated procedure. Defining personality uses Big-5 personality traits(McCrae and Costa, 1997), commonly used personality feature in psychology. Five traits, extraversion, aggreeableness, neutroticism, conscientiousness, and openness, are scored at a numerical value. In this work, we infer gender, age, and personality of each character by accumulating conversation-wise scores. We propose Naive Bayes classifier for inferring personality feature.

2.5 Relationship Extraction

Relationship between characters may defined in several methods since is has no gold standard to

quantify. Here we define relationship as three dimensions, sentiment, romantic, and top-down relationship. Similar to the characteristic extraction scheme, such relationships can be obtained by accumulating per-conversation scores. Since we know speaker and listener by prior procedures, we can collect speaker-listener relationship, assuming conversation score is relationship from speaker to listener.

2.6 Main Character Decision

Using the result of relationship extraction and count of the conversation, main characters can be determined. Two considerations exists. They are the number of conversations for each characters, and the number of relations from the result of relationship extraction. Combining two components can provide the numerical reference of character importance, result in main character decision. Following is the weighting equation.

$$weight = 5N_(relations) + N_(conversations) \quad (1)$$

If the proportion of the total exceeds 7.5% we think it is the main character.

3 Experiment & Results

3.1 Listener & Anaphora Resolution

For listener & anaphora resolution's evaluation, We generate true data from heuristics. Accuracy of listener resolution on first 100 conversation recorded 73%, sufficient to confirm the algorithm works. We also calculated accuracy of anaphora resolution on 40 anaphora and received fairly high score on F1-metric.

Precision	Recall	F score
0.689	0.815	0.647

Table 1: Anaphora resolution evaluation

3.2 Personality Extraction

Here we used PAN-15 dataset, personality-annotated dataset normalized into -0.5-0.5 range, for training classifier. We then accumulated result from each sentence by character. Since personality inference quality is ambiguous, we evaluated gender prediction as an indirect result. Comparing entire character's estimated gender and real gender, we obtained 73% F-score, which can assure sufficient performance of our model. The sample result of our model is shown in the appendix.

Precision	Recall	F score
0.761	0.711	0.735

Table 2: Characteristic extraction evaluation

3.3 Relationship Extraction

As defined in prior section, we extracted relationship in three dimensions. For sentiment scoring, we applied vader sentiment analysis to each conversation. Romantic and top-down analysis is done by keyword searching. We granted a score for each presence of keyword such as 'love' for romantic and 'sir' for top-down. We divided script into three sections to show relationship change in time. The result is represented as graph for each segment, shown in appendix.

3.4 Main Character Decision

The program shows main characters are Anna, Kristoff, Elsa, Olaf, Hans. In official Frozen said the main characters are those 5 people and reindeer Sven. Sven was not detected, because reindeer does not have any conversations. We also apply this model to "Beauty and the Beast". It seems there are 5 main characters, but 5th character is not included in official. The most important thing of this result is main characters' weight of story is almost over half. In the Frozen, Only 5 out of 55 characters make up 56 percent of story. Therefore, we made the relationship graph around the main characters.

4 Discussion & Conclusion

We successfully analyzed anaphora resolution, listener & speaker resolution, characteristic extraction, relationship extraction, and main character decision. Still, we have room to improve our system. Overall, we should think about ways to analyze characters who have no dialogues. For anaphora resolution, more proper methodology to classify pleonastic-it is needed. For characteristic extraction, we could enhance the system with using more complex model like n-grams. For relationship extraction, we could extract more diverse relationships and do more delicate act division. Nevertheless, our system has opened the way for movie analysis. With publicly available code, we hope that this work could be extended further.

A Appendix

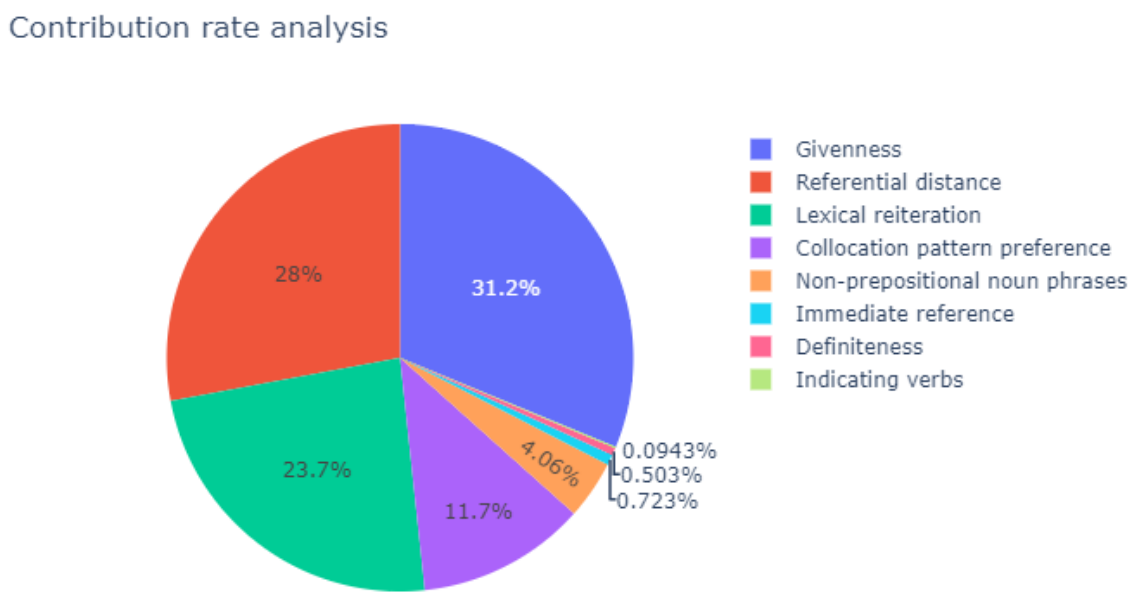


Figure 1: Contribution rate of each rule in anaphora resolution

((a)) *frozen*

Personality	Young Elsa	Elsa	Anna
Openness	0.125	0.147	0.142
Conscientiousness	0.125	0.121	0.118
Extraversion	0.119	0.115	0.117
Agreeableness	0.141	0.161	0.157
Neuroticism	0.05	0.112	0.107

((b)) *beauty and the beast*

Personality	Belle	Beast	Gaston
Openness	0.173	0.141	0.217
Conscientiousness	0.132	0.122	0.129
Extraversion	0.125	0.127	0.131
Agreeableness	0.160	0.166	0.132
Neuroticism	0.121	0.139	0.165

Table 3: Characteristic extraction result

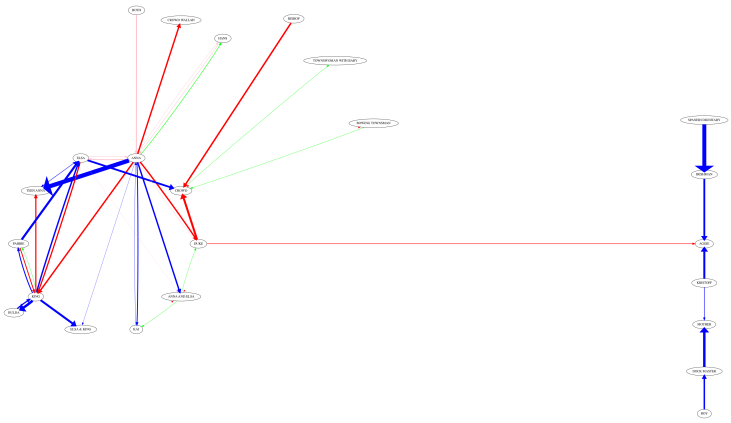
((a)) *frozen*

Name	Weight	Importance
ANNA	436	MAIN
KRISTOFF	244	MAIN
ELSA	167	MAIN
HANS	154	MAIN
OLAF	134	MAIN
DUKE	79	SUB
KAI	42	SUB
KING	42	SUB
TROLLS	29	SUB
OAKEN	28	SUB
PABBIE	27	SUB
BULDA	27	SUB
GERDA	19	SUB
SPANISH DIGNITARY	19	SUB
MARSHMALLOW	18	SUB
CROWD	18	SUB
BOTH	17	SUB
FRENCH DIGNITARY	12	SUB
TROLL PRIEST	12	SUB
'CITIZEN ONE	8	SUB
Total 55 Characters	1747	-

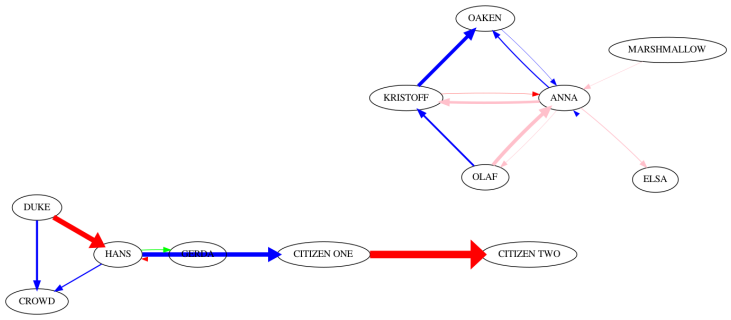
((b)) *beauty and the beast*

Name	Weight	Importance
BELLE	186	MAIN
BEAST	99	MAIN
GASTON	97	MAIN
LUMIERE	92	MAIN
MAURICE	91	MAIN
COGSWOTH	81	SUB
MRS.POTTS	63	SUB
LEFOU	56	SUB
CHIP	44	SUB
ALL	30	SUB
WOMAN	28	SUB
FEATHERDUSTER	27	SUB
MAN	23	SUB
BOOKSELLER	21	SUB
TOWNSFOLK	17	SUB
BOTH	13	SUB
BAKER	13	SUB
MOB	12	SUB
D'ARQUE	12	SUB
BIMBETTIES	12	SUB
Total 40 Characters	1160	-

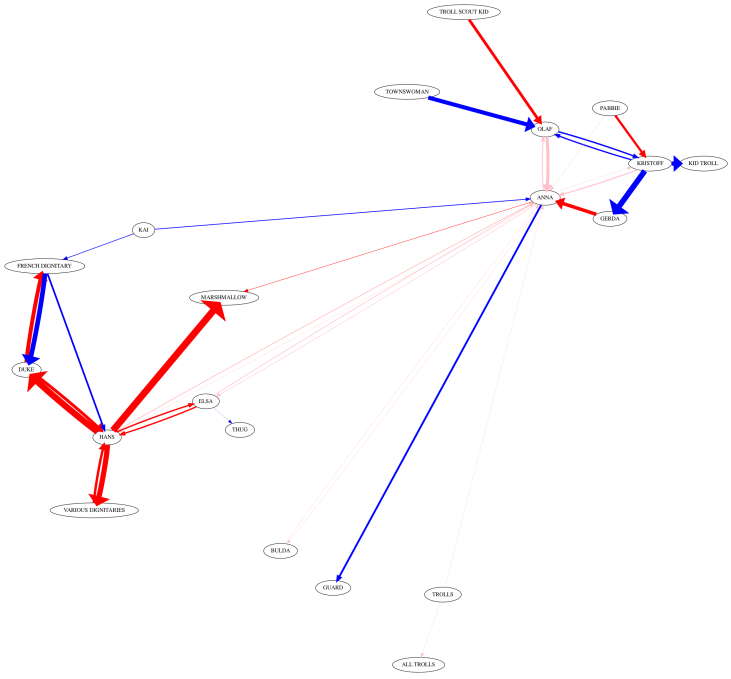
Table 4: Top 20 Characters of the result of Main Character Decision



(a) phase 1

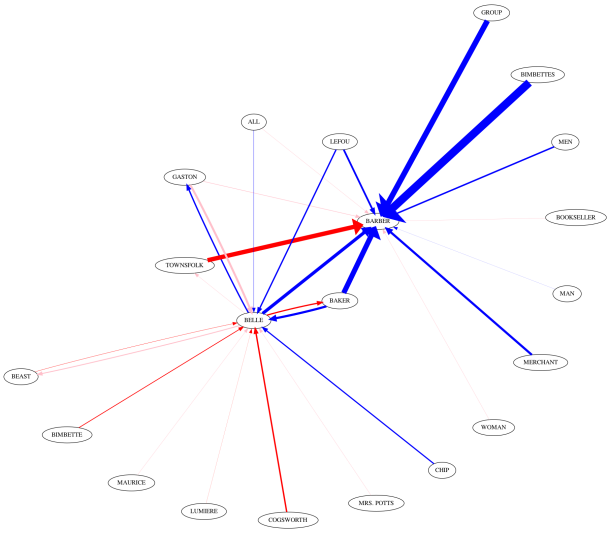


(b) phase 2

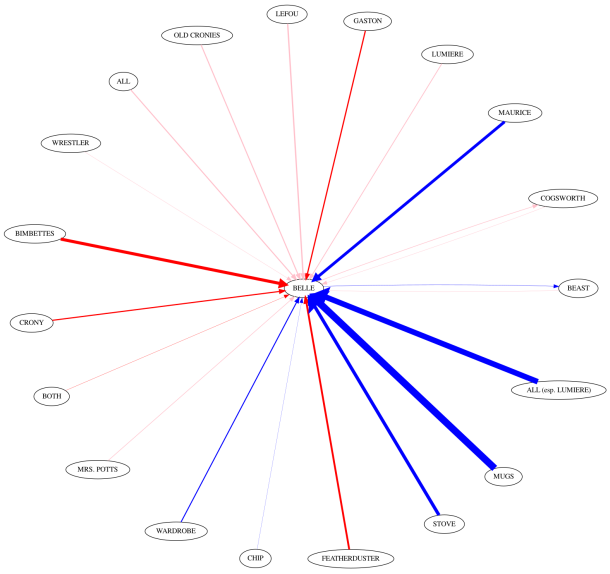


(c) phase 3

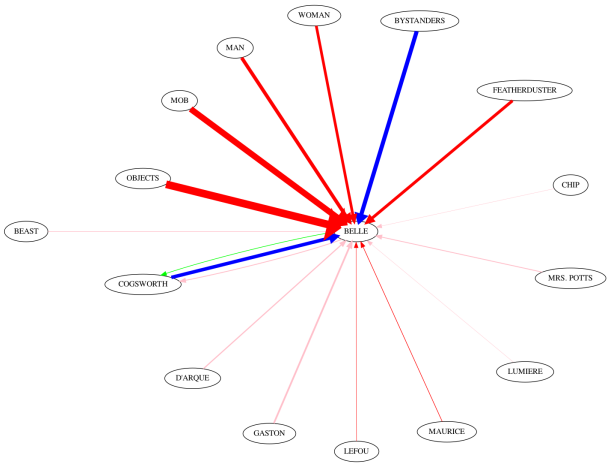
Figure 2: Relationship graph: *frozen*



(a) phase 1



(b) phase 2



(c) phase 3

Figure 3: Relationship graph: *beauty and the beast*

References

Robert R. McCrae and Paul T. Costa. 1997. Personality trait structure as a human universal. *American Psychologist*, 52(5):509–516.

Ruslan Mitkov. 1998. Robust pronoun resolution with limited knowledge. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 2*, ACL '98/COLING '98, page 869–875, USA. Association for Computational Linguistics.