Extracting Character Information from Movie Script

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

20170357 ShinDongHwan **20130240 KyuminPark**

1 Introduction

We analyzed movie scripts and extract information based on the script *frozen*. This includes preprocessing, listener resolution, anaphora resolution, and personality extraction. We successfully identified the listener for each script and applied some of the rules from Mitkov's system for listener and anaphora resolution. For personality extraction, we successfully extracted gender, age, and big-5 personality traits. Based on what we conducted, we plan to determine hostility between characters, conduct protagonist and antagonist detection, and determine family relationship for the next step. Our works are open public at https://github.com/soyoung97/playNLP.

2 Problem statement

Although there was many works done to analyze literature, there was less work done on analyzing movie scripts. Thus, we decided to take a novel approach on analyzing play or movie scripts while keeping the directions and methods same as what we proposed at earlier project proposal.

3 Technical approach and models

3.1 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. Since some conversation are split into many lines, we parse and collect to one line. For this, we made a finite-state machine to know where the lines start and end. There are two types of speech: normal talk and sing. For these, we divided into 3 states in finite-state machine: parsing normal talk, paring song, parsing narrator lines, and expecting sing. There is some special cases that two individual speech in one line. So we made one more special state for that. We only extracted information about scene time and place for movie script elements.

3.2 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 classifies the conversations by time and space. Same time_index value indicates conversation of same moment and place. Here we define listener as who speaks in five conversations ahead and back within the same time_index. Then, add the listeners who is included at neighboring conversation's listeners, regarding the absence of a person who exists with a particular two person in conversation but does not have a conversation.

3.3 anaphora resolution

Anaphora resolution is the problem of resolving what a pronoun, or a noun phrase refers to. After research, we decided to follow Mitkov's anaphora resolution system(Mitkov, 1998). At first step of our work, we use pos_tag function from nltk to find proper pos tags for the data. Word which is noun and is in the same paragraph with anaphora become a candidate. For pronoun anaphora, gender and number agreement filter is applied. After that, the antecedent indicators are applied. Out of 13 rules in the Mitkov's anaphora resolution system. We implemented the following 3 rules.

- *First noun phrase*: A score of +1 is assigned to the first NP in a sentence.
- *Lexical reiteration*: A score of +2 is assigned to those NPs repeated twice or more in the paragraph in which the pronoun appears and a score of +1 is assigned to those NPs repeated once in the paragraph.

- *Indicating verbs:* A score of +1 is assigned to those NPs immediately following a verb which is a member of a predefined set.

Unfortunately, there are some rules can not be applied to our system. For example, Section heading preference rule can not be applied because play script doesn't have heading sections. Rather than using the given rules as they are, we will modify some rules and excluding improper rules. Also, there is some issue about finding "pleonastic-it". Pleonastic means anaphors used without an antecedent. In this particular case *"Hup! Ho! Watch your step! Let it go!"* the word "it" does not indicate any antecedent. To get a good quality result, we will have to handle pleonastic-it well.

3.4 personality extraction

Using preprocessed script, our goal is to infer personality of each character with an automated procedure. Big-5 personality traits(McCrae and Costa, 1997), are commonly used in the field of personality extraction. Five traits of extraversion, aggreeableness, neutroticism, consciousness, and openness to experience is scored at a numerical value. In this work, we will infer gender, age, and personality of each character by scoring each of five personality feature from the script. We added scores of each conversation by each character and averaged it when inferring each personality for each character. We used Naive Bayes classifier and trained the personality feature from external dataset. Computing personality feature from text using trained classifier, we can deduce character's personality from collection of result.

4 preliminary experiments & result

4.1 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 percent, sufficient to confirm the algorithm works. We also calculated accuracy of anaphora resolution on first 100 apperance, which recorded 38 precent accuracy. Following shows sample of each good result and one bad result.

- *Good result*: "Young Kristoff struggles to get a block of ice out of the water. He(**Kristoff**) fails, ends up soaked. Sven licks his(**Kristoff's**) wet cheek. "
- *Bad result*: "A young Sami boy, KRISTOFF(8), and his(**KRISTOFF's**) reindeer calf, SVEN, share a carrot as **they** try to keep up with the men."

As you can see, our resolution system can resolve simple sentence like first sentence. However, our resolution system failed to translate "they" into "KRISTOFF and SVEN" in the second sentence. This is because our system never see the noun phrase "KRISTOFF and SVEN". We are going to add rules from Mitkov's, and if more is needed, we will challenge to make a new rule to get a better performance.

4.2 personality extraction

Personality features annotated in PAN-15 dataset was normalized into -0.5-0.5 range. We then accumulated result from each sentence by character. Regarding gender comparison as an indirect result, we compared 50 characters' estimated gender and real gender, not including characters who are ambiguous in gender. As a result, 11 out of 50 characters made an error on gender. Such prediction result can be interpreted as evidence on classifier, with 78 percent accuracy.

Young Elsa	Personality	Elsa
0.125	Openness	0.147
0.125	Consciousness	0.121
0.119	Extraversion	0.115
0.141	Agreeableness	0.161
0.05	Neuroticism	0.112

Table 1: Personality Comparison

In addition to the result on gender, we estimated the experiment result with personality feature extracted on same character. We could also compare single character in different time scene. Table 1, as an example, is a personality of Elsa and young Elsa. We can observe that neutroticism doubled as growing up. This aligns with the story of 'FROZEN' that Elsa had happy time in young age while having harsh time in aged. By this observation we can verified that classifier has ability to extract character personality. Analysis on important features on naive bayes classifier is listed on the appendix.

A Appendix

An analysis of most 10 important traits on Naive Bayes classifier.

Feature	More common on	Bias Weight
courtesy	M	46.9
francisco	M	28.4
sentiment	M	23.6
ny	M	20.2
np	F	17.8
semantic	M	17.5
processing	M	17.5
weekly	F	17.5
soundtrack	F	15.9
iemand	M	15.4

Table 2: Gender

Feature	More common on	Bias Weight
data	35-49	102.4
wrestling	50-XX	49.9
processing	35-49	49.2
natural	35-49	48.9
courtesy	25-34	45.4
social	35-49	41.3
abundance	35-49	38.9
interesante	35-49	37.2
web	50-XX	35.9
digital	50-XX	32.5

Table 3: Age_group

Feature	More common on	Bias Weight
mexican	negative	160.1
nowplaying	negative	160.1
wrestling	negative	95.0
ca	positive	85.6
mexico	negative	82.0
skills	negative	78.4
john	negative	75.4
acting	negative	75.4
confusion	negative	75.4
beb	negative	75.4

Table 4: extroverted

Feature	More common on	Bias Weight
bus	negative	95.4
ca	positive	82.4
?	positive	78.9
w	positive	72.3
nowplaying	positive	68.4
pic	positive	64.6
courtesy	positive	58.2
san	positive	50.5
fuck	negative	49.7
antwerp	positive	48.6

Table 5: stable

Feature	More common on	Bias Weight
recognition	negative	379.1
personality	negative	226.6
icwsm	negative	183.5
fotos	negative	168.4
computational	negative	149.5
lastfm	negative	136.0
workshop	negative	134.3
bus	negative	107.8
discover	negative	94.5
tht	positive	82.3

Table 6: agreeable

Feature	More common on	Bias Weight
personality	positive	93.5
afternoon	negative	87.9
politicians	negative	87.9
champ	negative	87.9
voy	negative	87.9
chase	negative	87.9
shadow	negative	87.9
ash	negative	87.9
strange	negative	87.9
empty	negative	87.9

Table 7: conscientious

Feature	More common on	Bias Weight
orange	positive	319.2
unlocked	positive	175.2
torino	positive	154.1
pic	positive	150.4
coffee	positive	135.5
york	positive	134.8
plant	negative	124.4
ousted	positive	104.4
attack	positive	101.9
turing	negative	96.0

Table 8: openness

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.