Journal of The Korea Society of Computer and Information
Vol. 24 No. 1, pp. 49-55, January 2019
www.ksci.re.kr
https://doi.org/10.9708/jksci.2019.24.01.049

Classification of Characters in Movie by Correlation Analysis of Genre and Linguistic Style

Eun-Soon You*, Jae-Won Song**, Seung-Bo Park ***

Abstract

The character dialogue created by AI is unnatural when compared with human-made dialogue, and it can not reveal the character's personality properly in spite of remarkable development of AI. The purpose of this paper is to classify characters through the linguistic style and to investigate the relation of the specific linguistic style with the personality. We analyzed the dialogues of 92 characters selected from total 60 movies categorized four movie genres, such as romantic comedy, action, comedy and horror/thriller, using Linguistic Inquiry and Word Count (LIWC), a text analysis software. As a result, we confirmed that there is a unique language style according to genre. Especially, we could find that the emotional tone than analytical thinking are two important features to classify. They were analyzed as very important features for classification as the precision and recall is over 78% for romantic comedy and action. However, the precision and recall were 66% and 50% for comedy and horror/thriller. Their impact on classification was less than romantic comedy and action genre. The characters of romantic comedy deal with the affection between men and women using a very high value of emotional tone than analytical thinking. The characters of action genre who need rational judgment to perform mission have much greater analytical thinking than emotional tone. Additionally, in the case of comedy and horror/thriller, we analyzed that they have many kinds of characters and that characters often change their personalities in the story.

▸Keyword: Artificial Intelligence, Dialogue, Personality, Linguistic style, Movie genre

I. Introduction

기계학습의 눈부신 발전으로 딥러닝(deep learning)을 이용한 스토리 자동 생성기(story generator) 연구가 활발하게 이루어지고 있다. 인터랙티브 소설을 창작하는 세헤라자데(Scheherazade) 시스템부터 트위터(Twitter)를 이용하여 사용자와 함께 호러(horror) 소설을 쓰는 인공지능 '쉘리(Shelly)', 그리고 영화 시나리오를 쓰는 인공지능 '벤자민(Benjamin)'까지 그 종류도 다양하다[1,2,3]. 세헤라자데(Scheherazade)는 여러 토픽에 대해 사용자들이 작성해 놓은 문장들을 학습하고 쉘리는 레딧(Reddit)에 올려진 14만 개 이상의 공포 스토리를 학습했다. 벤자민은 장단기 기억 신경망(Long Short-Term Memory recurrent neural network)을 이용하여 20편의 SF 영화 시나리오를 문장 단위로 학습하였다[3]. 벤자민이 쓴 공상 과학 시나리오 2편은 2016년과 2017년에 각각 <선스프링(sunspring>과 <It's no game>이라는 영화로 제작되면서 큰 화제가 되었다. 벤자민은 SF 영화에 자주 나오는 단어와 표현들을 사용하는 등 SF 영화의 전형성을 모방하려 했다는 점에서

• First Author: Eun-Soon You, Corresponding Author: Seung-Bo Park
*Eun-Soon You (tesniere@naver.com), AI Content Creation Research Center, Inha University
**Jae-Won Song (jwsong03@gmail.com), Value Finders
***Seung-Bo Park (molaal@inha.ac.kr), Dept. of Software Convergence Engineering, Inha University
• Received: 2018. 11. 29, Revised: 2018. 12. 12, Accepted: 2018. 12. 13.
• This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (No. NRF-2017R1D1A1B03036046).

주목을 받았다. 그러나 벤자민이 학습을 통해 생성한 인물들의 대사는 인간이 쓴 영화 대사와 비교할 때 질적인 측면에서 여러 문제점을 노출했다. 첫째, 벤자민의 대사는 인물에 대한 기본적인 정보는 물론 무엇을 말하고 있는지를 분명하게 전달하지 못했다. 영화 대사는 인물의 신분과 직업, 처한 상황, 사건과 같은 정보를 알려줌으로써 스토리에 대한 이해를 돕는다. 예를 들어 영화 <마이 페어 레이디(My fair lady)>의 여주인공 일라이자가 사용하는 투박한 코크니 사투리와 거친 은어는 주인공이 교육을 제대로 받지 못한 런던 하층민 출신이라는 것을 드러내는 언어적 특징이다. 둘째, 인물들이 주고받는 대화가 맥락에 맞지 않거나 동문서답으로 이루어지는 것과 같이 대화의 내용이 전체적으로 어색하고 부자연스럽다. 셋째, 벤자민의 대사에는 인물의 성격을 드러내는 언어적 특징이 부족하다. 대사의 목적은 스토리의 내용을 전달하는 것뿐만 아니라 인물의 정체성을 표출하기도 한다. 인물의 정체성은 성격 묘사를 통해 구체화 되는데, 성격은 인물의 행동과 말하는 방식을 통해 발현된다. “인물은 말을 통해 자신의 면모를 드러낸다”라는 영화 비평가 Vassé의 언급처럼 영화 속 인물들은 각자의 언어 스타일을 갖고 있다[4]. 그리고 대사를 통해 표출되는 감정과 심리 상태, 내적 욕망 등은 성격을 인지할 수 있는 유용한 단서가 된다. 이러한 측면에서 아직은 인공지능이 만든 대사와 작가에 의해 창작된 대사 사이에는 큰 간극이 존재한다.

본 연구의 장기적인 목표는 성격 차원에 따라 언어 표현을 변주하여 자연스러운 대사를 구성하는 캐릭터 대화 생성기(character dialogue generator)를 개발하는 것이다. 이에 대한 선행 연구로 본 논문은 영화 속 성격에 따른 인물들의 대사를 분석하여 그들의 언어 스타일을 장르에 따라 분류(classification)하고 그들의 언어 사용과 장르 간의 상관성을 살펴보는 것을 목적으로 한다. 그동안 영화 연구를 통해 영화 장르에는 해당 장르에 부합하는 전형적인 인물이 등장하고[5] 장르에 따라 고유한 대사 유형이 존재한다는 사실이 강조된 바 있다[4]. 본 연구는 기존의 주관적이고 정성적 분석이 아닌 객관적인 실험을 통해 그러한 사실을 증명한다는 점에서 의미가 있다.

영화 장르와 성격에 따른 언어 스타일 간의 관계를 분석하기 위해 본 논문은 현재 텍스트 분석에 널리 사용되고 있는 Linguistic Inquiry and Word Count (LIWC)를 이용하여 인물들의 대사를 분석한다. LIWC는 텍스트에 포함된 단어들이 개인의 심리적, 성격적 특성을 반영하여 텍스트를 분석하는 소프트웨어이다. LIWC를 이용하여 텍스트로부터 피처를 추출한 후에 KNN 기반의 캐릭터 분류를 하고 분류된 캐릭터들의 언어 스타일을 분석한다.

논문의 구성은 다음과 같다. 먼저 2장에서는 관련 연구들을 검토하고, 3장에서는 LIWC를 활용한 영화 캐릭터 대사 분석 방법과 분석 결과와 의미에 대해 상세히 논의한다. 마지막으로 4장에서는 결론 및 향후 연구 방향을 제시한다.

II. Related Works

단어 사용에 대한 객관적 연구 방법으로 ‘어휘접근법(lexical approach)’이 있다. ‘어휘가설(lexical hypothesis)’이라고도 불리는 이 방법은 개인의 성격이 그가 사용하는 어휘와 일치한다는 것을 전제로 한다. 즉 사람들이 사용하는 단어에서 성격을 발견할 수 있다는 것이다. 이 이론에 근거하여 Sir Francis Galton은 1884년에 자신의 논문 ‘성격의 측정(Measurement of Character)’에서 어휘를 이용한 성격 측정을 통해 처음으로 성격에 대한 과학적 접근을 시도했다[6]. 그는 사람들 간의 성격적 차이는 그들이 사용하는 단어에서부터 발생한다고 주장하고 영어 사전에서 성격을 묘사하는 1000개 이상의 단어들을 선별했다. 골턴의 연구에 이어 Allport 역시 인간이 저마다 갖는 특유의 성격 특성은 언어를 통해 측정 가능하다고 간주하고 성격을 설명하는 단어 17,953개를 수집하였다[7]. 골턴과 올포트 모두 성격을 기술하는 단어들을 분류하고 단어들의 사용을 정량화하여 개인별 성격적 차이를 설명할 수 있다고 보았다. 하지만 성격을 반영하는 단어가 부족하거나 부재할 경우 특정 성격이 포착되지 않을 수도 있다는 문제점이 있다.

인간의 성격을 외향성, 신경성, 성실성, 친화성, 개방성의 다섯 개로 분류한 ‘빅 파이브(Big Five)’가 성격 모델의 표준으로 널리 사용되면서 심리언어학 분야에서는 단어와 ‘빅 파이브’간의 연관성을 계량화하는 연구들이 활발하게 이루어졌다. 대량의 텍스트를 분석하기 위해 많은 학자들이 심리 측정(pschometric) 소프트웨어인 LIWC를 활용하여 성격이 단어 사용에 어떻게 영향을 미치는지를 관찰하였다. 외향성의 언어 특징에 관한 연구를 살펴보면 사교적이고 낙천적인 외향적 성격은 조용하고 내성적인 내향적 성격보다 더 많은 단어와 더 높은 긍정적인 감정어(emotion word)를 사용하고 있으며[8,9,10], 사회적 관계와 인간과 관련된 단어들의 사용이 높은 것으로 나타났다[10]. Gill, A., & Oberlander, J.과 Mehl et al.은 각각 이메일 텍스트와 학생들이 작성한 설문지 답변을 분석한 결과 외향적 성격의 글이 더 많은 단어 수를 포함하고 있다고 하였다[8,9].

Pennebaker & King은 중독 치료 센터에 입원한 환자들의 일기를 통해 외향적 성격이 더 많은 긍정적 감정어를 사용하는 반면 부정적인 감정어는 거의 사용하지 않는다는 사실을 발견했다[10]. Pennebaker는 특히 1인칭대명사 ‘I’의 사용에 주목했는데 신경증적 성격의 환자들의 글에서 ’I’의 사용 빈도가 다른 성격에 비해 높게 나타난 것에 주목하고 그 이유를 자신의 내면의 고통이나 슬픔에 집중하는 행위라고 설명하였다. 그래서 명사나 형용사와 같은 내용어(content word)보다는 대명사, 전치사, 관사와 같은 기능어(function words)들이 말하는 사람에 대해 더 많은 것을 드러낸다고 강조하였다. Pennebaker는 기능어는 명사나 동사와 같이 실질적 의미를 가진 단어들보다 그 수가 매우 적으나 실제 우리가 사용하는 단어의 60%를 차지하고 있으며, 우리가 무의식적으로 사용하는 사소한 기능어들이 오히려 개인의 지위와 권력, 성별, 나이, 사고방식 등을 훨씬 더 많이 보여준다고 하였다[11].

Beukeboom et al.은 LIWC를 활용하여 사진 5장에 대해 40명의 피험자들이 묘사한 글을 분석한 후 설문지를 이용하여 피험자들의 외향성과 내향성을 측정하였다[12]. 그리고 내향적인 사람들이 외향적인 사람들보다 관사와 숫자, 수량(quantification)과 같은 기능어를 더욱 많이 사용하여 구체적이고 정확한 표현을 선호하며 but과 except와 같은 배타적이고 부정적인 단어들의 사용도 높다고 밝혔다.

III. Dialog Analysis and Character Classification

1. Data for Experiment

본 연구는 캘리포니아대학교 샌타크르주(UCSC)의 NLDS (Natural Language And Dialogue Systems) 랩(lab)에서 구축한 영화 코퍼스 1.0(Film Corpus 1.0)의 일부를 실험데이터로 활용하였다[13]. 해당 코퍼스에는 인터넷 영화 대본 데이터베이스(IMSDB)로부터 수집된 862편의 정제된 영화 스크립트와 영화 속 인물 7,400명의 대사 등이 포함되어 있다. 그리고 각 인물의 대사는 '영화 제목_인물 이름' 형식으로 되어 있다. 예를 들어 영화 <애니 홀>의 주인공 '애니'의 모든 대사는 'annie-hall-annie.txt'로 저장된다.

본 논문은 862편 중에서 로맨틱 코미디(Romantic Comedy), 코미디(Comedy), 호러/스릴러(Horror/Thriller), 액션(Action) 장르로 구성된 60편의 영화 스크립트를 선택하였다. 그리고 각 영화에서 대화 차례(turns of dialogue)가 100회 이상이고 대사량이 가장 많은 주인공만을 고려하여 총 92명을 최종 선정하고 그들의 대사를 추출하였다. 대사는 주인공의 화면 밖 목소리인 보이스 오버(voice over), 독백 그리고 대화(dialogue)로 구성되었으며, 92명으로부터 수집된 대사의 문장 수는 총 47931개이다. Table 1은 선정된 영화와 각 영화 속 주인공들을 기술하고 있으며 Figure 1은 영화 스크립트와 주인공 대사의 일부를 보여준다.

Table 1. Selected movie characters

Genre	Movies	Characters
Romantic Comedy	Punch drunk love, He's just not that into you, Annie Hall 외 12편	Barry, Lena, Gigi, Alex, Annie, Alvy 외 24명
Comedy	Dumb and dumber, Human nature, Ace ventura 외 12편	Harry, Lloyd, Nathan, Puff, Ace ventura 외 15명
Horror/ Thriller	The Shining, Misery, Birds 외 12편	Jack, Paul, Annie, Melanie, Mitch 외 15명
Action	The Dark Knight, Sherlock Holmes, Iron Man 외 12편	Batman, Joker, Gordon, Holmes, Iron Man 외 17명

2. Dialog Analysis using LIWC

LIWC는 현재 널리 사용되고 있는 단어 빈도 기반의 텍스트 자동 분석 도구이다. 1993년에 개발된 이후 지속적으로 업그레이드되고 있다. 본 연구는 가장 최근에 소개된 2015년 LIWC 버전을 분석에 이용하였다.

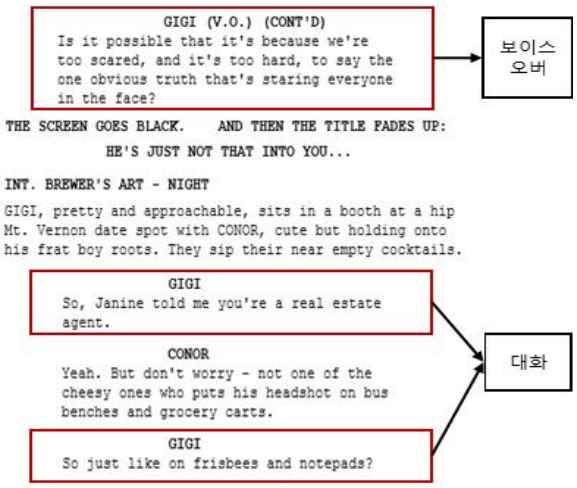


Fig. 1. The dialogue of the character 'Gigi' in the script of Movie <He's just not that into you>

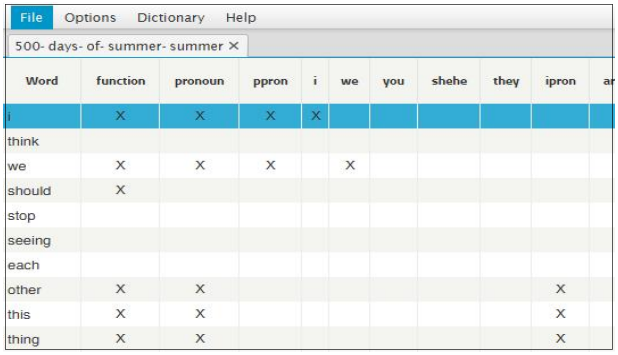
LIWC는 텍스트 분석을 위해 계층구조로 이루어진 총 93개의 피처(feature)들을 이용한다. 'Word Count(단어수)', 'Summary Language Variables(요약 언어 변수)', 'Linguistic Dimensions (언어 차원)', Other Grammar', 'Psychological Processes (심리 과정)'은 대분류에 속하는 피처들이며 Table 2에서 보는 것처럼 각각의 피처는 다시 중분류와 소분류로 나뉜다. 텍스트 분석의 핵심 역할을 하는 LIWC의 단어 사전은 각 피처들과 관련된 단어들로 구성되어 있다. LIWC 사전에서 한 단어는 여러 개의 피처에 속할 수 있다. 가령 단어 'cried'는 'Negative emotion', 'sadness', 'common verbs' 등 5개 피처에 포함된다[14].

Table 2. LIWC features and words in each feature

Features			Words in feature
Word count			
Summary Language Variables			
Analytic thinking			
Emotional tone			
...			
Linguistic Dimensions			
Total pronouns			I, them, itself
Article			a, an, the
Prepositions			to, with, above...
...			
Other Grammar			
Common verbs			eat, come, carry...
...			
Psychological Processes			
Affective processes	Positive emotion		love, nice, sweet...
	Negative emotion		hurt, ugly, nasty...
		Anxiety	worried, fearful...
		Anger	hate, kill, annoyed...
		...	
Social processes	Family		daughter, dad, aunt...
	Friends		buddy, neighbor...
	...		

Table 2에서 알 수 있듯이 LIWC의 가장 큰 특징은 동사, 명사, 형용사와 같은 내용 단어뿐만 아니라 그동안 분석에서 제외되었던 전치사와 관사, 대명사와 같은 기능어와 'er', 'hm'와 같은 비형식적 표현들까지 텍스트 분석을 위한 피처로 사용했다는 점이다. Pennebaker는 이러한 기능어들이 개인의 심리와 사회적 지위를 드러내는 중요한 "언어 지문"임을 강조한 바 있다[11].

LIWC를 이용한 분석은 크게 3개의 과정으로 이루어진다. 먼저 인물의 대사 텍스트를 입력하면 텍스트 속 단어 중 LIWC 사전에 포함된 단어들을 검출한다. 다음으로 표시된 단어들은 Figure 2에 보이는 것처럼 93개의 피처로 분류된다. 최종적으로 LIWC 사전을 이용하여 93개의 피처로 분류된 단어들의 빈도를 측정하여 백분율로 산출한다.


File Options Dictionary Help

500- days- of- summer- summer ×

Word	function	pronoun	ppron	i	we	you	shehe	they	ipron	ar
i	X	X	X	X						
think										
we	X	X	X		X					
should	X									
stop										
seeing										
each										
other	X	X							X	
this	X	X							X	
thing	X	X							X	

Fig. 2. The dialogue of the main character 'Summer' of the movie <500 Days of Summer>

Figure 3은 92명의 대사에 대해 피처별 단어 빈도값을 보여주고 있다. 이중에 영화 <애니 홀>의 남자 주인공 알비(Alvy)의 대사(annie-hall-alvy.txt)에 대한 피처별 분석 결과를 살펴보면 알비 대사에서 차지하는 '단어 수(WC)'는 7795개로 이는 92명 중에서 가장 높은 값이다. 이것은 로맨틱 코미디가 코미디 장르와 더불어 다른 장르에 비해 대사 의존도가 높다는 장르적 속성을 보여준다. 그리고 'Analytic thinking(분석적 사고, Analytic)'과 'Emotional tone(감정적 어조, Tone)'은 각각 11.02와 60.30을 나타내고 있다. 'Analytic'은 형식적이고 논리적 단어들의 사용과 관련되며 'Tone'은 감정어 사용에 영향을 받는다. 따라서 알비는 이성적이고 분석적인 단어들보다 감정적 단어들을 더 많이 사용했음을 유추할 수 있다.

Filename	Segment	WC	Analytic	Clout	Authentic	Tone	WPS	Sixltr	Dic	function
500-days-of-summer-summer.txt	1	1590	4.58	69.93	72.87	95.30	4.44	9.56	94.09	59.37
annie-hall-alvy.txt	1	7795	11.02	66.63	72.53	60.30	6.78	10.37	90.53	56.92
dark-knight,-the-the joker.txt	1	2847	61.91	81.57	31.81	66.96	6.93	10.85	85.95	54.55
he's-just-not-that-into-you-gigi.txt	1	2598	11.10	77.00	67.88	80.79	7.47	10.24	93.46	61.82
misery-annie.txt	1	4372	10.86	71.78	61.39	64.36	8.87	10.02	90.53	59.58

Fig. 3. Value of features for character dialogue

3. Feature Selection using T-test

LIWC의 93개 피처들 중에서 장르 구분에 가장 유효한 피처를 선별하기 위해 t검정(t-test)을 이용하였다. Table 3과 같이 93개의 피처를 6개 가설에 적용하여 하나의 피처가 각각의 가설을 만족하면 '0' 값을 갖게 하고, 만족하지 않으면 '1' 값을 갖도록 하였다. 그 결과 6개의 모든 가설에서 '1' 값을 갖는 피처는 'Analytic'과 'Tone'으로 나타났다.

Table 3. Six hypotheses for four genres

Genre pair	Hypotheses	Value
(A,H)	(A=H) Action genre is same to Horror/Thriller.	0(accept), 1(reject)
(C,A)	(C=A) Comedy genre is same to Action.	
(C,H)	(C=H) Comedy genre is same to Horror/Thriller.	
(R,A)	(R=A) Romantic comedy genre is same to Action.	
(R,C)	(R=C) Romantic comedy genre is same to Comedy.	
(R,H)	(R=H) Romantic comedy genre is same to Horror/Thriller.	

이 두 개의 피처는 Figure 4처럼 액션과 호러/스릴러(A,H), 코미디와 액션(C,A), 코미디와 호러/스릴러(C,H). 로맨틱 코미디와 액션(R,A), 로맨틱 코미디와 코미디(R,C), 그리고 로맨틱 코미디와 호러/스릴러(R,H)의 장르적 차이를 모두 구분하는 중요한 피처로 선정되었다.

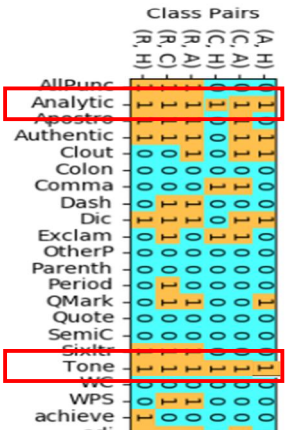


Fig. 4. T-test results for 93 features

4. Experiment and Discussion for Character Classification

영화 60편으로부터 선택된 92명 인물에 대해 'Analytic'과 'Tone'의 피처를 이용하여 최근접 이웃(K-Nearest Neighbors, K-NN) 알고리즘을 이용하여 4개의 그룹으로 분류하였다. 4개의 장르로 분류를 시도하여 검증을 진행하였다.

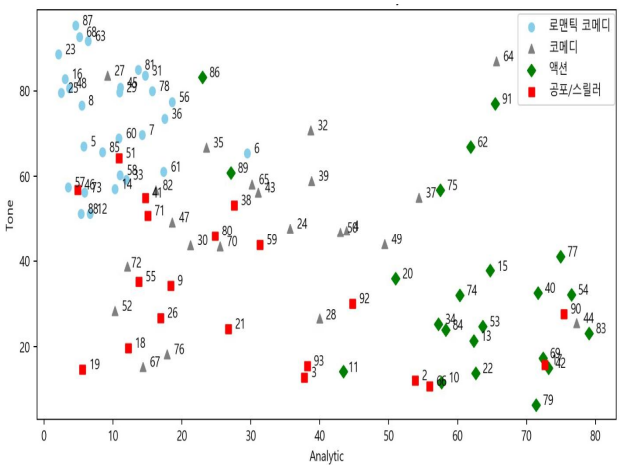


Fig. 5. Scatter plot of 'Tone' and 'Analytic'

선별된 피처 'Analytic'과 'Tone'의 값을 X축과 Y축으로 하여 92명의 인물들을 Figure 5와 같이 산점도(scatter plot)로 표시하였다. 기호는 인물을 나타내며 기호의 종류는 장르를 나타낸다. 숫자는 인물의 번호를 가리킨다.

그리고 K-NN 알고리즘을 이용하여 장르에 따라 인물들을 Figure 6과 같이 4개의 그룹(G1, G2, G3, G4)으로 분류하였다. 인물들의 분류 결과를 분석하면 Table 4와 같이 나타난다. 전체적으로 Precision이 72.0%이고 Recall이 70.6%로 나타났다. 로맨틱 코미디와 액션에 해당하는 인물들은 장르의 분류 결과가 70% 이상의 성능을 나타냈으나 코미디와 호러/스릴러의 인물들은 원하는 분류 성능을 보이지 못했다. 코미디와 호러/스릴러의 인물들은 Figure 6의 G2와 G3에 표시된 인물들로 상당히 넓게 분포하는 것을 알수 있다. 이것은 코미디와 호러/스릴러에 등장하는 인물들이 다양한 성격을 갖는 인물들이라는 것을 의미한다.

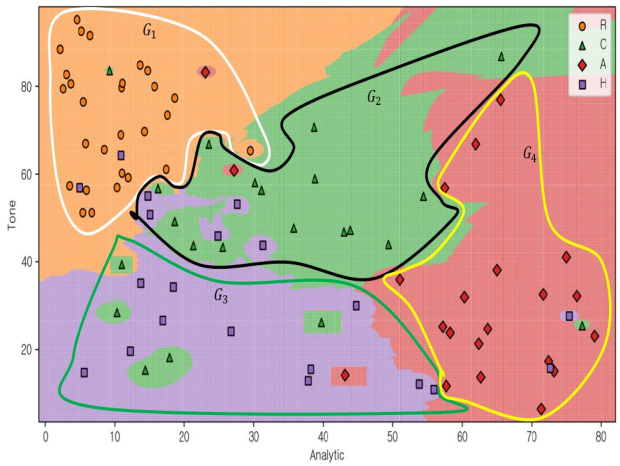


Fig. 6. KNN(k=9) classification for 'Analytic' and 'Tone'

'Tone' 값이 가장 높은 87번 인물은 영화 <500일의 썸머>의 여주인공 '썸머(summer)'이다. K를 9로 설정하고 살펴봤을 때 '썸머(summer)'와 가장 근접한 거리에 있는 9명은 68번 <펀치 드렁크 러브>의 레나(Lena), 63번 <애니홀>의 애니(Annie), 23번 <시애틀의 잠못 이루는 밤>의 애니(Annie), 16번 <시애틀의 잠못 이루는 밤>의 샘(Sam), 25번 <한나와 그의 자매들>의 한나(Hannah), 48번<펀치 드렁크 러브>의 베리(Barry), 27번 <휴먼 네이처>의 나단(Nathan), 81번 <노팅힐>의 안나(Anna), 8번 <500일의 썸머>의 톰(Tom)이다.

G1으로 분류된 장르를 살펴보면 로맨틱 코미디 영화 인물 28명이 모두 포함되었다. 남녀 간의 애정과 갈등, 행복한 결합으로 끝나는 로맨틱 코미디의 장르적 특성상 인물의 대사에 다양한 감정어들이 포함되어 있다. 로맨틱 코미디인 G1의 인물들은 'Tone' 높은 반면에 'Analytic'이 낮은 전형성이 강하게 드러난다. 로맨틱 코미디는 대부분 상업영화의 특성을 가지기 때문에 대부분 유사한 성격의 인물들이 주연으로 등장하게 되어 장르의 특성과 강하게 결합하게 된다. 이로 인해 로맨틱 코미디의 경우 Recall이 100%로 나타난 것으로 판단된다. 인물들은 예를 들어 영화 <500일의 썸머>의 여주인공 '썸머'(87번), 영화 <펀치 드렁크 러브>의 레나(68번) 그리고 영화 <애니 홀>의 애니(63번)는 자유분방하고 충동적이며 사랑을 적극적으로 표현하는 로맨틱 코미디 장르의 전형적인 인물들이다. 예외적으로 호러/스릴러 장르의 인물 4명과 코미디 장르의 인물 3명, 그리고 액션 장르의 인물 1명이 로맨틱 코미디의 인물들과 함께 분류되었는데, 이 영화들은 공통적으로 로맨스 장르가 결합된 복합장르의 특징을 갖고 있다. 예를 들어 IMDB(Internet Movie Database)에서 메인 장르가 호러/스릴러로 분류된 영화 <미저리>는 주인공 애니의 광기가 드러나기 전까지 폴에 대한 애니의 애정을 보여주며 스토리가 진행된다. 마찬가지로 호러/스릴러 영화 <새>에서도 멜라니와 밋치의 로맨스가 극 전체에서 높은 비중을 차지한다.

Table 4. Experimental result of classification

Class	Correct	Mismatch	Precision	Recall
G1 (Romantic Comedy)	28	8	77.8%	100%
G2 (Comedy)	10	3	76.9%	45.5%
G3 (Action)	18	5	78.3%	81.9%
G4 (Horror/ Thriller)	11	9	55.0%	55.0%
Total	67	25	72.0%	70.6%

G2는 코미디 장르의 인물들을 포함하고 있다. 기본적으로 해피엔딩(happy ending)의 결말과 웃음을 유발하는 과장된 인물 또는 상황 설정이 코미디 장르의 전형적인 서사(narrative)이지만, 코미디의 가장 큰 특징은 그 형식이 매우 다양하고 코미디가 포괄할 수 있는 서사의 범위 또한 방대하다는 것이다[15]. 따라서 코미디 장르에 등장하는 인물들은 저마다 각양각색이다. G2로 분류된 13명의 코미디 영화의 인물은 'Analytic'

과 'Tone'의 값이 모두 높은 <브루스 올마이티>의 '브루스'(64번)부터 'Tone'의 값이 높은 <베리 배드 씽>의 '피셔'(47), 그리고 <금발이 너무해>의 엘르(4번), <라이어 라이어>의 플레처(50)와 같이 'Analytic'과 'Tone'의 값이 균형을 이루는 인물들에 이르기까지 다양하다. 한편 G2에는 액션 영화 <아이언맨>의 주인공 아이언맨(89번) 외에도 호러/스릴러 영화의 인물 3명(38번, 59번, 80번)이 함께 포함되어 있다. 이 영화들은 복합 장르가 아닌 단일 장르임에도 불구하고 코미디 영화 인물들과 근접 거리에 위치하였다. 아이언맨의 경우 이성적이고 냉철한 사업가였던 토니 스타크가 사고 이후에 유쾌하고 유머러스한 인물로 성격의 변화를 겪으면서 단어 사용에서도 변화가 일어난 것으로 보인다. 그러나 3편의 호러/스릴러 영화 속 인물들은 코믹한 인물들과는 거리가 멀다. 해당 장르의 전형적인 인물임에도 불구하고 코미디 인물들과 함께 분류된 이유에 대해서는 G3의 특징을 먼저 살펴보면서 설명할 것이다. G3은 호러/스릴러 영화 인물 11명과 코미디 영화 인물 5명 그리고 액션 영화 인물 1명을 포함하고 있다. G2와 유사하게 호러/스릴러의 인물과 코미디 장르의 인물이 동일한 그룹에 분포되어 있는데, 그 이유를 호러/스릴러 장르와 코미디 장르의 기본적인 서사 전략이 모두 '긴장과 놀람'에 있다는 사실에서 유추해볼 수 있다[5,15]. 긴장과 놀람은 정보의 분배와 관련된 것으로 일찍이 알프레드 히치콕 감독이 그 개념을 설명한 바 있다. 긴장은 영화 속 중요한 정보를 관객은 알고 있지만 주인공은 모르고 있을 때 생겨나며 놀람은 그 반대로 관객이 정보를 전혀 모르고 있을 때 일어난다. 호러/스릴러가 공포를 자아내기 위해 긴장과 놀람의 서사 전략을 취한다면 코미디는 웃음을 유도하기 위해 긴장과 놀람의 전략을 이용한다. 또한, 공포/스릴러는 코미디 장르와 마찬가지로 그 형식이 매우 다양한데, 호러/스릴러 인물들의 특징을 살펴보면 'Analytic'의 값이 높은 <콘스탄틴>의 안젤라(2번)와 <파이널 데스티네이션>의 알렉스(66번)부터 'Analytic'의 값이 낮은 <메멘토>의 레너드까지 다양하다. 단기 기억상실증 환자인 레너드의 대사는 단순하고 건조한 정보 전달형에 가깝고 그의 심적 상태나 감정은 주로 그의 얼굴 표정과 행위를 통해 드러난다. 따라서 레너드는 'Analytic'뿐만 아니라 'Tone'의 값이 91명의 인물 중에서 가장 낮다. 초자연적인 대상이나 혹은 절대적 악과 대립하는 호러/스릴러 장르의 인물들은 레너드처럼 감정 전달을 대사가 아닌 얼굴을 통해, 그리고 대사보다는 행위에 더 집중된다. 91명의 대사를 LIWC로 분석했을 때 호러/스릴러 장르의 인물들의 WC(단어수)가 다른 장르의 인물들에 비해 적은 것으로 나타났다.

G3에는 액션 영화 <왓치맨>의 로렐(11번)이 포함되어 있는데, 로렐은 액션 장르의 다른 인물들처럼 미션을 수행하거나 문제 해결에 직접적으로 참여하지 않는다. 그녀는 대화를 통해 혼란에 빠진 동료를 이끌어주는 인간적이고 이성적인 인물로 액션 영화의 전형적인 인물과는 거리가 있다.

G4에는 액션 장르의 인물 19명, 코미디 장르의 인물 1명 그리고 호러/스릴러 장르의 인물 2명이 포함되어 있다. G4의 인물들은 'Analytic'의 값이 'Tone'의 값보다 높은 것이 특징으로 언어적 측면에서 G1의 로맨틱 코미디의 인물들과 대조를 이룬다.

액션 장르는 주인공이 자신에게 주어진 임무를 완수해나가는 서사를 담고 있기 때문에 '감정'의 문제보다는 자신에게 주어진 임무를 해결해 가는 과정에서 형식적이고 논리적인 단어들의 사용이 높은 것으로 분석된다. 뛰어난 추리 능력으로 사건의 진실을 밝히는 영화 <셜록 홈즈>의 셜록 홈즈(75번), 킬러로서 위험한 미션을 수행하는 영화<원티드>의 웨슬리(42번), 살인사건에 감춰진 진실을 밝히려는 영화 <왓치맨>의 로어셰크(79번) 등이 대표적이다. 예외적으로 코미디 영화 <에이스 벤추라>의 주인공 에이스 벤츄라(44번)와 호러/스릴러 장르의 영화 <천사와 악마>의 랭돈, <콘스탄틴>의 존이 포함됐다. <천사와 악마>, <콘스탄틴>은 내면 심리에 초점을 둔 심리 스릴러라기보다는 액션에 중점을 둔 복합 장르에 속하며 <에이스 벤츄라>는 사건을 해결하는 사립탐정이라는 인물 설정에 맞게 이성적인 단어들의 사용이 높은 것으로 보인다.

IV. Conclusions

본 논문은 스토리 자동 생성에 필요한 성격에 따른 인물의 대사 생성기 개발에 앞서 영화 속 인물들의 대사를 분석하여 장르에 따른 언어 스타일을 분류하였다. 인물을 장르에 따라 분류하기 위한 주요 피처는 'Analytic'과 'Tone'의 2개로 파악되었다. 두 개의 피처는 로맨틱 코미디와 액션에 대해서는 Precision과 Recall이 모두 78% 이상 나올 정도로 분류에 매우 중요한 피처로 분석되었다. 하지만 코미디와 호러/스릴러의 경우 Precision이 66%, Recall이 50% 정도로 로맨틱 코미디와 액션 장르에 비해 분류에 대한 영향력이 좀더 적게 계산되었다. 이것은 두 개의 피처를 이용하여 장르와 인물 간의 관계를 분석하여 로맨틱 코미디와 액션의 인물들은 주로 전형적인 성격을 가지지만 코미디와 호러/스릴러의 경우 상당히 다양한 성격을 갖는다는 것과 인물들이 스토리 내에서 성격이 변화하는 경우가 많다는 것을 분석할 수 있었다.

스토리 창작에서 인물의 성격 설정은 매우 중요한 문제이다. 인물의 성별과 나이, 직업, 취미 등 인물의 성격화에 기반이 되는 것들을 결정해야 한다. 그중에서도 대사는 인물의 성격을 보여주는 중요한 단서로서 장르별 전형적인 인물들의 성격에 따른 언어 스타일 학습은 대사 생성기 개발을 위해 반드시 선행되어야 하는 문제이다.

본 연구는 언어적 측면에서 인물의 성격을 분석하여 실제 영화 속 인물의 성격과 언어 스타일이 얼마나 일치하는지에 대한 문제까지는 다루지 못했다. 또한, 장르를 구별하는 피처를 2개로 선별한 후에 장르와 인물 간의 관계성을 분석하였으나 93개의 피처들을 모두 이용하고 딥러닝 기술들을 적용할 경우 경계

에 있는 인물들에 대한 추가적인 정보들을 분석할 수 있을 것으로 예상한다. 이는 후속 연구로 남기면서 아직은 연구 초기 단계에 있는 인물의 대사 생성 연구를 활성화하는 계기가 되기를 기대한다.

REFERENCES

[1] Li, B. and Riedl, M. O., "Scheherazade: Crowd-Powered Interactive Narrative Generation," Proc. of the 29th AAAI Conference on Artificial Intelligence, pp. 4305-4306, 2015.

[2] https://twitter.com/shelley_ai

[3] http://benjamin-ai.tumblr.com/

[4] Vassé, C., "The Dialogue," Ewha Womans University Press, p. 19, 2010.

[5] Chung, Young Kwon, "Understanding Movie Genres," Amormundi, p. 27, 2017.

[6] Galton, F., "Measurement of character," Fortnightly Review, 36, pp. 179-185, 1884.

[7] Allport, G. W. and Odbert, H. S., "Trait-names: A psycho-lexical study," Psychological monographs, Vol. 47, No.1, pp. i-171, 1936.

[8] Gill, A. and Oberlander, J., "Taking care of the linguistic features of extroversion," Proc. of the 24th Annual Conference of the Cognitive Science Society, pp. 363-368, 2002.

[9] Mehl, M, R., Gosling, S. D., and Pennebaker, J. W., "Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life," Journal of Personality and Social Psychology, Vol. 90, pp. 862-877, 2006.

[10] Pennebaker, James W., and Laura A. King., "Linguistic styles: Language use as an individual difference," Journal of personality and social psychology, Vol. 77, pp. 1296-1312, 1999.

[11] Pennebaker, James W., "The secret life of pronouns: What our words say about us," Sa-i Book Publishing, p. 9, 2016.

[12] Beukeboom, C. J., Tanis, M., and Vermeulen, I. E., "The language of extraversion: Extraverted people talk more abstractly, introverts are more concrete," Journal of Language and Social Psychology, Vol. 32, pp. 191-201, 2013.

[13] https://nlds.soe.ucsc.edu/node/23/done?sid=4863&token=6c19f764f110e75e129fab4e4dd683d9

[14] Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K., "The development and psychometric properties of LIWC2015," LIWC2015 Development Manual, pp. 1-25, 2015.

[15] Krutnik, F., and Neale, S., "Popular film and television comedy," Communication Books, p.60, 2002.

Authors

Eun-Soon You received Ph.D. in Computational Linguistics from Franche-Comté University, France, in 2007. She is currently a researcher in the Artificial Intelligence Content Creation Research Center at INHA University. Her interests include computational narrative, deep learning, natural language processing.

Jae-Won Song received the M.S. and Ph.D. degrees in Computer and Information Engineering from Inha University, Korea, in 2007 and 2013, respectively. Dr. Song joined Value Finders Co., Ltd, Korea, in 2017. He is the CEO of Value Finders Co., Ltd. He is interested in financial data mining.

Seung-Bo Park received the BS, M.S. in Electrical Engineering and Ph.D. degrees in Information Engineering from Inha University, Korea, in 1995, 1997 and 2011, respectively. His research interests include video story analyzing, semantic contents, video knowledge representation, social network analysis, and A.I. He worked at Daewoo Electronics as an engineering researcher.