실시간 채팅 데이터를 이용한 하이라이트 추출

이주헌, 염익준*
성균관대학교, *성균관대학교

wngjs0216@naver.com, *ikjun@skku.edu

Extract highlights using real-time chat data

Lee Ju Heon, Yeom Ik Jun*
Sungkyunkwan Univ., * Sungkyunkwan Univ.

요 약

실시간 스트리밍 시장이 최근 몇 년 사이에 급격한 성장하는 추세에서 코로나 정국을 맞이하여 더욱 더 가파른 성장을 하고 있다. TV 프로그램과는 다른 상당히 긴 영상길이를 특징으로 하는 스트리밍 동영상의 하이라이트를 추출한다. 하이라이트 추출에는 KoNLPy 의 Okt 분석기를 사용하여 해당 채팅의 특징들을 학습한 뒤 이를 이용하여 하이라이트 구간을 추출해 낸다. 그 결과, 채팅 수만을 지표로 했을 때보다 12% 적은 구간을 추출해 내면서 동일한 성능을 보인다.

Ⅰ. 서 론

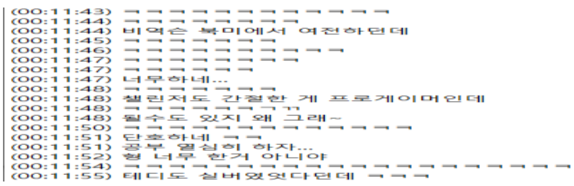
실시간 스트리밍 시장은 해가 갈수록 더욱더 성장하는 추세이며, 특히 코로나 정국을 맞이한 지금 그 성장속도가 더욱 더 가팔라지고 있다. Twitch tv[1]의 경우 주당 시청자 수는 120 만명이 넘고, 월간 시청 시간이 10 억 시간에 이르는 등 수많은 사람들이 실시간 스트리밍을 이용하고 있다. 하지만 실시간 생방송이라는 특징으로 편집이 가미된 TV 프로그램과는 다르게 총 방송시간이 매우 길며, 다시 보기 동영상의 경우 총 영상시간이 10 시간이 넘는 영상도 흔하게 볼 수 있다. 영상의 전체 길이가 상당한 점은 동영상을 시청하는 시청자들에게 큰 부담을 준다.

[그림 1] Twitch tv 동영상 화면

이러한 단점을 해결하기 위해 본 논문은 Twitch tv 의 다시 보기 동영상에서 실시간 채팅 데이터를 추출하여, 그 데이터를 이용하여 하이라이트 구간을 추출해 내는 것을 목표로 한다.

Ⅱ. 본론

1. 데이터 수집



(00:11:43) ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ
(00:11:44) ㅋㅋㅋㅋㅋㅋㅋㅋ
(00:11:44) 비역슨 북미에서 여전하던데
(00:11:45) ㅋㅋㅋㅋㅋㅋㅋ
(00:11:46) ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ
(00:11:47) ㅋㅋㅋㅋㅋㅋㅋㅋ
(00:11:47) ㅋㅋㅋㅋㅋㅋ
(00:11:47) 너무하네...
(00:11:48) ㅋㅋㅋㅋㅋㅋㅋ
(00:11:48) 챌린저도 간절한 게 프로게이머인데
(00:11:48) ㅋㅋㅋㅋㅋㅋㄱㄲ
(00:11:48) 될수도 있지 왜 그래~
(00:11:50) ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ
(00:11:51) 단호하네 ㅋㅋ
(00:11:51) 공부 열심히 하자...
(00:11:52) 형 너무 한거 아니야
(00:11:54) ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ
(00:11:55) 테디도 실버였었다던데 ㅋㅋㅋ

[그림 2] 채팅데이터

데이터수집 단계에서는 Twitch tv API 를 이용하여 해당 영상으로부터 채팅정보들을 수집한다. 수집한 채팅들은 (시간:분:초) 채팅내용을 1 줄로하여 텍스트파일로 저장한다.

2. 데이터 분석

(00:49:05) 이번 시즌 아프리카 최약체네 ㅋㅋ
(00:49:07) 라바가 인간상성인가 ㄷㄷ
(00:49:09) 세체미 라바임?
(00:49:09) 포탑골드 항상 1천 넘게 찍다가 오늘은 960이네
(00:49:10) 저런 쇼메를 이긴 라바를이긴 페이커는 얼마나 잘하는거냐 ?
(00:49:10) 아니 티원 이겨놓고 이젠 아프리카랑도 비비네...
(00:49:11) 내가 누누이 말하지만 쇼메이커가 플라이보단 잘함
(00:49:11) 칸은 계속 내려오는데 cs 차이 안나는거 뭐임? 기인 거품 확빠졋누
(00:49:11) 와
(00:49:12) 주식그래픈
(00:49:15) 잘못업는게 어딨음 팀게임에 ㅋㅋ
(00:49:16) 돌아와라 김기인...
(00:49:17) 자연이 웰케 화났누
(00:49:17) 데굴데굴
(00:49:17) 탑은 없어진거 체감안드는데 갑코가 이렇게 크구나...양파 돌아와
(00:49:20) 라업쇼왕 누구냐 개웃기네 ㅋㅋ
(00:49:22) 드레드 우리가왔다알리네 ㅋㅋ
(00:49:24) 쓰레쉬 개사기네 아펠이랑같이하니까
(00:49:27) 서리불꽃 오른
(00:49:27) 이래도 080이 안나오네 ㅋ

(01:02:27) ?
(01:02:27) 칸 ㅋㅋㅋㅋㅋ
(01:02:27) ???
(01:02:27) 왜이러지 애들
(01:02:27) 오른 ㅋㅋ
(01:02:27) ?
(01:02:27) ㅋㅋㅋㅋㅋ
(01:02:27) 패기 ㅋㅋㅋ
(01:02:28) 느그칸 수준ㅋㅋㅋㅋㅋㅋㅋㅋ
(01:02:28) ?
(01:02:28) 지나갑니다 ㅋㅋㅋㅋㅋ
(01:02:28) 지나가던ㄱ리 ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ
(01:02:28) ?
(01:02:28) ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ
(01:02:28) 잼 른
(01:02:28) ???
(01:02:29) 아니 진짜 뭐임?
(01:02:29) 캐니언님?
(01:02:29) ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ

[그림 3] 비하이라이트, 하이라이트 구간 채팅

그림 3 에서 볼 수 있듯이 하이라이트 구간의 경우 비 하이라이트 구간보다 채팅이 짧으며 “ㅋㅋㅋㅋ”와 같은 한글 낱자가 포함된 채팅이 많아지는 것을 확인 할 수 있다. 이에 기반하여 KoNLPy[2]의 Okt 형태소분석기를 이용하여 동영상의 채팅 내용들을 분석하였다. Okt 를 활용하여 “지나갑니다 ㅋㅋㅋㅋㅋ”의 채팅을 분석할 경우 [('지나갑니다', 'Verb'), ('ㅋㅋㅋㅋㅋ', 'KoreanParticle')]의 결과를 얻을 수 있다.

[표 1] 전체 구간의 채팅내용 분석 결과

영상	총 길이 (초)	총 채팅 수	초당 평균 채팅	한글 낱자 포함 비율	명사 포함 비율	채팅 평균 길이	평균 형태소 개수
1	19393	37624	1.940	0.374	0.768	11.223	4.346
2	32749	32749	1.450	0.329	0.797	12.097	4.810
3	21832	31548	1.445	0.355	0.792	10.802	4.281
4	12815	32692	2.551	0.376	0.737	11.335	4.656
5	19101	39753	2.081	0.362	0.789	10.670	4.122
6	18284	23180	1.268	0.342	0.799	11.908	4.686
7	26223	72499	2.765	0.357	0.780	11.571	4.456
8	20000	34234	1.711	0.331	0.795	12.285	4.618
9	18274	26731	1.462	0.374	0.788	12.012	4.457

10	13860	44131	3.184	0.332	0.735	13.918	4.768

표 1 과 표 2 를 살펴보면 공통적으로 초당 평균채팅 수는 하이라이트 구간에서 약 2 배 증가하며, 한글 낱자 포함 비율은 소폭 증가하는 경향을 볼 수 있다. 명사 포함 비율, 채팅 평균 길이, 평균 형태소 개수는 10~20%가량 감소하는 것을 확인할 수 있다.

[표 2] 하이라이트 구간의 채팅내용 분석 결과

영상	총 길이(초)	총 채팅 수	초당 평균 채팅	한글 낱자 포함 비율	명사 포함 비율	채팅 평균 길이	평균 형태소 개수	구간 개수
1	1976	7959	4.028	0.408	0.708	8.598	3.343	51
2	3227	9151	2.836	0.337	0.753	10.196	3.962	56
3	3261	8710	2.670	0.383	0.725	9.136	3.581	60
4	1374	6937	5.049	0.412	0.697	9.463	3.830	30
5	1984	8469	4.269	0.415	0.701	8.805	3.225	49
6	1944	6071	3.123	0.391	0.717	9.545	3.650	47
7	2917	14081	4.827	0.381	0.703	9.168	3.555	64
8	1957	5935	3.033	0.378	0.759	10.297	3.868	43
9	2007	6725	3.351	0.420	0.697	9.654	3.446	45

표 1 과 표 2 를 살펴보면 공통적으로 초당 평균채팅 수는 하이라이트 구간에서 약 2 배 증가하며, 한글 낱자 포함 비율은 소폭 증가하는 경향을 볼 수 있다. 명사 포함 비율, 채팅 평균 길이, 평균 형태소 개수는 10~20%가량 감소하는 것을 확인할 수 있다.

[표 3] 전체 구간 대비 하이라이트 구간의 비율

영상	초당 평균 채팅	한글 낱자 포함 비율	명사 포함 비율	채팅 평균 길이	평균 형태소 개수	하이라이트 시작시간(초)
1	2.076	1.091	0.922	0.766	0.769	2383
2	1.956	1.024	0.945	0.843	0.824	2762
3	1.848	1.079	0.915	0.846	0.836	2387
4	1.979	1.096	0.946	0.835	0.823	2695
5	2.051	1.146	0.888	0.825	0.782	2123
6	2.463	1.143	0.897	0.802	0.779	2256
7	1.746	1.067	0.901	0.792	0.798	2568
8	1.773	1.142	0.955	0.838	0.838	2330
9	2.292	1.123	0.885	0.804	0.773	2430
10	1.400	1.111	0.902	0.783	0.826	3283

표 3 은 전체 구간 대비 하이라이트 구간의 비율과 각 동영상의 하이라이트 시작 시간을 보여준다. 본 논문에서는 이 정보를 이용하여 하이라이트 구간 추출을 구현한다.

표 3 을 기준으로 leave-one-out 방식을 이용하여 하이라이트를 추출 시 추출하려는 영상을 제외한 9 개의 영상을 훈련 집합으로 하여 평균 형태소 개수, 한글 낱자 포함 비율, 명사 포함 비율, 채팅 평균 길이, 초당 평균 채팅의 구간대비 하이라이트 구간의 비율과 하이라이트 시작시간의 평균으로 h, p, n, l, c, stim 을 정하고 입력 값으로 사용한다.

Ⅲ. 결론

1. 하이라이트 추출

하이라이트 구간을 추출해 내기 위해서 먼저 각 시간별로 채팅정보를 추출한다. 이 때, 영상을 보고서 채팅을 입력하는 시간, 사용자 별 네트워크 환경 등을 고려하여 10 초동안의 채팅기록을 해당 시간의 채팅으로 간주한다. 한 예로, 01:03:42 시각의 데이터는 01:03:42~01:03:51 의 10 초동안의 모든 채팅을 포함한다.

위의 데이터를 이용하여 전체 구간의 시간별 채팅 수, 평균 형태소 개수, 평균 명사 포함 비율, 평균 한글 낱자 포함 비율을 구한다.

하이라이트 구간은 아래의 규칙에 의해 선정한다.

1)평균 시작시간보다 큰 시간내에서 추출한다.

2-1) 평균 채팅 수가 앞에서 구한 비율만큼 전체구간의 평균 채팅 수보다 많은 경우 나머지 4 개의 지표 중에서 2 가지 이상을 만족하는 경우

2-2) 평균 채팅 수가 앞에서 구한 비율보다 15%이상 더 많은 경우

2. 결과 분석

하이라이트 추출의 성능을 분석하기위해 추출 정확도와 구간 길이의 두가지 지표를 이용하였으며, 채팅만을 지표로 하였을 경우와 위의 여러 지표들을 활용하였을 경우를 비교하였다. 실제 하이라이트 N 개의 구간 중에서 각 구간이 추출된 하이라이트와 겹치는 부분이 있을 경우 해당 구간을 탐지했다고 보고 N 개 중에서 n 개의 구간을 탐지한 경우 n/N 을 추출 정확도라고 정의한다. 구간 길이의 경우 위의 방식으로 추출한 하이라이트의 총 길이를 의미한다.

[표 4] 추출된 구간의 추출 정확도와 구간 길이

	평균추출 정확도	평균구간 길이(초)
(A) Onlychat	0.742	2072
(B) Chat + @	0.736	2000

표 5 에서 확인할 수 있듯이 (B)에서는 채팅 양 만을 기준으로 삼아서 하이라이트 구간을 선정했을 경우와 유사한 추출 정확도를 얻어내면서 추출 구간의 길이를 12%가량 감소시킬 수 있었다.

본 논문은 동영상의 채팅데이터만을 기준으로 하이라이트 구간을 추출했기에 이 데이터에 더불어 음성데이터와 영상데이터를 포함하여 하이라이트 구간을 선정한다면 더욱 높은 정확성을 가질 것으로 예상된다.

참 고 문 헌

[1] twitch tv , http://twitch.tv/.

[2] 박은정, 조성준, “KoNLPy: 쉽고 간결한 한국어 정보처리 파이썬 패키지”, 제 26 회 한글 및 한국어 정보처리 학술대회 논문집, 2014.