RNN-LSTM 기반 장면 자막 메타데이터 생성 방법

*곽창욱 **김선중
한국전자통신연구원
*cukwak@etri.re.kr **kimsj@etri.re.kr

A method for creating the Scene closed-caption metadata based on RNN-LSTM

*Kwak, Chang-Uk **Kim, Sun-Joong
Electronic Telecommunications Research Institute

요약

정확한 영상 검색을 지원하기 위해 다양한 데이터와 방법들을 통한 메타데이터 생성 연구들이 이루어지고 있다. 자막 데이터를 기존의 키워드 기반의 메타데이터 생성 방법을 이용했을 경우, 구어체, 불완전 문장의 특징을 가진 특징을 반영하는데 어려움이 있었다. 또한, 단순히 키워드 매칭에 의존하기 때문에 문장에 중의적 단어가 포함되어 있을 경우에 검색 정확도가 떨어진다는 한계점이 있다. 따라서, 본 논문에서는 이러한 문제를 해결하기 위해 문장 전체를 특정 단위로 표현한 메타데이터를 생성한다. 이를 위해 비지도 학습인 RNN-LSTM 기반 네트워크를 이용하여 자막을 인코딩하고 장면 지식으로 생성하는 방법을 제안한다. 실험에서는 본 시스템을 통해 임의의 자막을 입력하고 유사도 기반의 결과 비교를 통해 자막 메타데이터의 정성적 평가를 수행하였다.

1. 서론

최근 검색 엔진에서의 영상 검색 비중이 증가하고 있으며, 영상을 대상으로 한 검색 엔진의 산업적, 사회적 영향력이 확대되고 있다. 보통 미디어 영상들은 시청자들을 통해 소비되기도 하지만, 2차 저작물과 같이 기존 영상들을 짜깁기하여 새로운 컨텐츠로 생산하는 등 컨텐츠 재활용 측면의 산업적 접근도 시도되고 있다.

일반적인 영상 검색 엔진에서는 해시태그와 같은 키워드 기반으로 메타데이터를 생성하여 검색을 지원하고 있지만, 이러한 검색 엔진에서는 장면을 묘사한 지문, 등장 인물의 발화 내용과 같은 자연어 기반 질의 검색에는 취약하다는 단점이 있다. 함경준 외[1]는 시나리오 형태의 질의어를 기반으로 한 영상 검색 시스템을 제안하였다. Han et al.[2]은 사용자가 입력한 시나리오를 기반으로 기존 영상을 활용하여 새로운 영상을 창작하는 시스템을 제안하였다. 이처럼 정확하고 복합적인 영상 검색을 위해 입력 질의의 단위가 문장 단위로 확장되는 연구가 이루어지고 있으며, 이러한 자연어 질의에 대한 검색을 위해 문장 단위의 메타데이터 생성이 필요하다.

본 논문에서는 영상의 자막을 이용하여 자막 메타데이터를 생성하는 방법을 제안한다. 본 논문에서 제안한 시스템에서는 먼저 영상을 장면 단위로 분할하고, 분할한 영상과 자막을 동기화한다. 이후, 영상의 자막을 비지도 학습인 RNN-LSTM 기반 네트워크를 통해 특정 벡터로 변환한다. 실험에서는 실제 입력된 임의의 문장 질의와 생성된 자막 메타데이터의 유사도 비교를 통해 정성적인 평가를 수행하였다.

본 논문의 구성은 다음과 같다. 2절에서는 시스템 구조 및 구현 내용에 대해 설명한다. 3절에서는 본 시스템을 통해 실험한 결과를 비교분석한다. 마지막으로 4절에서는 결론을 맺는다.

2. 시스템 구조 및 구현

본 논문에서 제안하고 있는 자막 메타데이터 생성 시스템은 영상과 자막을 입력으로 장면 영상 단위의 자막 메타데이터를 생성한다. 시스템 구조도는 아래 그림 1과 같다.

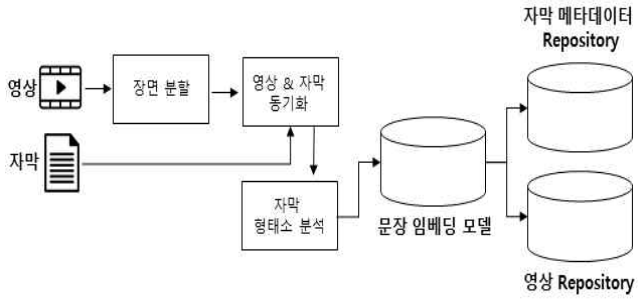

그림 1 자막 메타데이터 생성 시스템 구조도

현재 대부분의 포탈에서는 장면 단위로 스트리밍 서비스가 이루어지고 있다. 장면은 2~3분 내외 길이의 영상으로써, 동일한 장소나 유사한 주제를 기반으로 구분할 수 있다. 시스템에서 영상과 자막을 입력하면, Son et al.[3]이 제안한 딥 러닝 기반의 장면 분할 방법을 통해 영상을 장면 단위로 분할한다. 분할한 영상과 자막은 타임정보 비교를 통해 동기화한다. 자막은 트위터 형태소 분석기를 통해 전처리하고, RNN-LSTM 기반으로 학습된 문장 임베딩 모델을 통해 특정 벡터로

변환한다. 자막을 특정 벡터로 변환하는 문장 임베딩 모델은 아래 그림 2와 같으며, 인코더의 마지막 시퀀스의 Hidden State 값인 h_{an}을 인코딩된 벡터로 사용한다. 생성된 벡터들은 자막 메타데이터 Repository에 저장되고, 장면 영상은 영상 Repository에 저장한다.

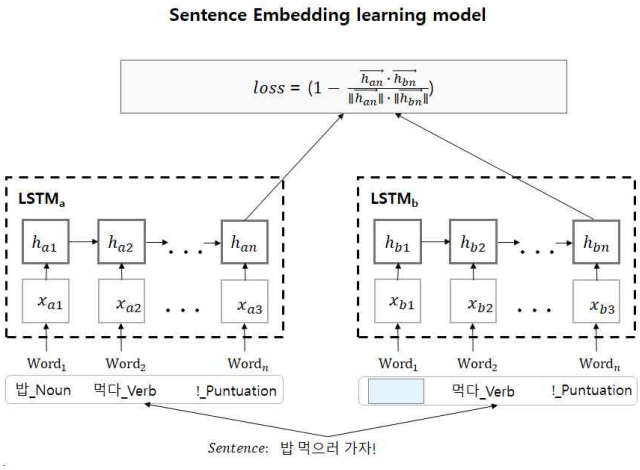

그림 2 문장 임베딩 모델 구조도

본 시스템을 통해 한국영화 75편을 대상으로 자막 메타데이터를 생성하였으며, 장면별로 생성된 자막 메타데이터 조회 화면은 아래 그림 3과 같다.

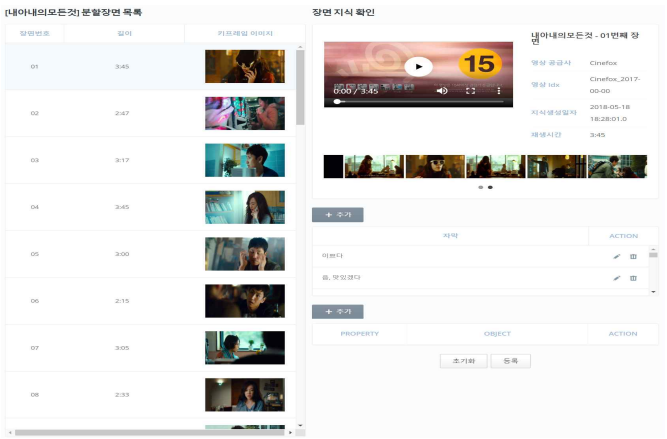

그림 3 장면별 자막 메타데이터 조회 화면

3. 실험 결과

자막 메타데이터 생성의 성능을 확인하기 위해 임의의 질의를 입력하고 정성적으로 평가하였다. 이를 위해 입력된 문장 질의와 생성된 자막 메타데이터 사이의 코사인 유사도를 비교하였다.

아래 표 1과 표 2는 임의의 2개의 질의에 대한 자막 유사도 비교 결과이다. 표 1에서 확인할 수 있듯이, 입력된 '밥 먹으러 가자'라는 질의에 대해 '밥 먹고 가', '그래, 만두 먹으러 가자' 등과 같은 유사한 자막들이 높은 유사도를 가지고 있음을 볼 수 있다. 또한 질의 '정말 미안합니다'라는 문장에 대해서는 '아이구 미안해요', '정말 죄송합니다'와 같이 특정 벡터로 표현했을 때, 유의어 사전 역할을 하고 있음을 확인할 수 있다. 본 방법을 통해 문장에서의 단어의 시퀀스 순서를 임베딩 벡터를 통해 반영할 수 있으며, 문장 단위의 맥락 비교를 통해 의미적 표현이 가능하다는 점에서 우수성을 찾을 수 있다.

표 1 자막 유사도 비교 결과

Q1: 밥 먹으러 가자	
순위	자막
1	밥 먹고 가
2	밥 먹으러 가자
3	뭐 먹을까? 그래 가자
4	그래, 만두 먹으러 가자
5	지검장님 연락 갔을 텐데
6	얼마전, 삼겹살을 먹으러 갔다가
7	살아있는 집에서 는어 떻게든 소리 가 나요
8	그래도 먹어 더 사올테니까
9	근데, 왜 갑자기 온천에들 가셨어요!
10	많이 먹어 뭐 더 줄까

표 2 자막 유사도 비교 결과

Q2: 정말 미안합니다	
순위	자막
1	아이구 미안해요
2	정말 죄송합니다
3	정말 죄송합니다
4	정말 죄송합니다
5	아유 아유 미안합니다
6	죄송합니다 아버님
7	근데 아빠
8	엄마 엄마
9	엄마 엄마
10	아니요 죄송해서요

4. 결론

본 논문에서는 딥러닝 기반의 자막 메타데이터 생성 시스템을 제안하였다. 이는 향후 자연어 기반 질의의 적용 가능성을 확인할 수 있었다는 점에서 우수성이 있다. 실험에서는 문장 전체를 하나의 특정 단위로 표현하였을 때, 문장 전체의 맥락을 포함하기 때문에 단어의 매칭이 아닐지라도 유사한 자막들이 높은 유사도를 가지고 있음을 확인할 수 있었다. 향후 연구로는 영상을 제작한 각본에서 지문 분석을 통한 장소, 인물 등의 정보를 추출하고 메타데이터를 생성하는 연구가 필요하다.

5. Acknowledgement

본 연구는 한국전자통신연구원 연구운영비지원사업의 일환으로 수행되었음. [18ZH1500, 오픈 시나리오 기반 프로그래머블 인터랙티브 미디어 창작 서비스 플랫폼 개발]

참고 문헌

[1] 함경준, 곽창욱, 한민호, & 김선중, "시나리오 형태 질의어 기반 영상 검색 시스템 개발." *한국정보과학회 학술발표논문집*, pp.344-346. 2018.
[2]Han, M., & Kim, S. J. (2018, January). User scenario based

video contents creation system. In 2018 International Conference on Information Networking (ICOIN) (pp. 61-63). IEEE.
[3] J. Son, et al., Deep "Ensemble Network for Movie Scene Boundary Detection,", *ISIS 2017*, pp.1157-1161, 2017.