템플릿 기반의 자동 소셜 매거진 및 영상 합성 서비스

*이재원 *장달원 *김미지 *김지수 *김서율 *이종설
전자부품연구원
*{jwlee0121, dalwon, miji0425, rlawltn9408, ssyykk, leejs}@keti.re.kr

Template-based Auto Social Magazine and Video Creation Service

*Lee, Jae-Won *Jang, Dal-Won *Kim, Mi-Ji *Kim, Ji-Su *Kim, Seo-Yul *Lee, Jong-Seol
Korea Electronics Technology Institute

요약

최근 자연어 처리 기술에 대한 중요도가 높아지고, 발전 속도가 빨라지면서, 산업 전반에 걸쳐 챗봇에 대한 수요가 증가하고 있다. 본 논문은 챗봇을 이용한 소셜 매거진 생성 및 배포, 그리고 이를 활용하여 사용자에게 텍스트를 음성으로 변환하여 동영상의 형태로 전달해 주는 시스템을 다루고 있다. 챗봇이 사용자 대화를 수집, 분석하여 상황에 맞는 키워드를 추출하고, 중복 콘텐츠 제거, 텍스트 요약 등 일련의 과정을 거쳐 소셜 매거진을 생성 및 배포하는 서비스와, 매거진의 각 콘텐츠를 구성하는 이미지, 텍스트 정보를 가지고 음성 합성, 자막 생성, 영상 효과 등을 이용하여 영상을 합성하는 서비스에 관한 것이다. 본 논문에서 제안한 시스템에 대한 성능은 실험을 통하여 검증하였다.

1. 서론

최근 1인 미디어 생산 및 소비의 증가와 자연어 처리 기술의 발전과 진보로 인해, 텍스트를 사람의 목소리로 자연스럽게 읽어주는 음성합성(Text to Speech, TTS) 기술이나 텍스트, 이미지를 통한 영상 합성, 그리고 질의 응답형, 정보 제공형 등의 서비스를 위한 챗봇의 수요가 증가하고 있다[1-4]. 특히, 여러 메신저 플랫폼에서의 챗봇 서비스와, 미디어 분야에서의 영상 합성 서비스가 증가하고 있다.

본 논문에서는 챗봇을 이용한 소셜 컨텐츠를 매거진 형식으로 제공하는 서비스와, 텍스트 및 이미지를 기반으로 음성 합성을 이용하여 동영상 형식의 미디어를 제공하는 플랫폼을 소개한다.

이 논문의 구성은 다음과 같다. 제안하는 챗봇 및 TTS 기반의 영상 합성 서비스에 대한 시스템 프레임워크는 2장에서 설명한다. 3장에서는 소셜 매거진 생성 및 영상 합성 모듈에 대해 설명하고, 이 플랫폼에 대한 실험적 검증은 4장에서 설명한다. 마지막으로 5장에서는 정리 및 결론으로 마무리한다.

2. 시스템 개요

2.1 챗봇 서비스

본 논문에서 제안하는 챗봇 서비스는 사용자의 관심사 분석 과정을 거쳐, 대화 주제와 연관성이 높은 소셜 미디어로부터 생성되는 소셜 매거진 서비스를 제공하는 것이다. 챗봇은 소셜 플랫폼에 사전 등록되어 사용자 대화방에 입장할 수 있으며, 사용자간 대화를 수집, 분석 등을 통해 사용자 대화와 관련성이 높은 소셜 매거진을 생성 및 배포할 수 있다.

2.2 영상 합성 서비스

생성된 소셜 매거진을 구성하는 JSON 형식 메타데이터에 포함된 텍스트 정보를 통해 음성합성(Text to Speech, TTS)과정을 거친 후, 텍스트 정보와 관련된 이미지를 가지고 영상을 합성하게 된다. 제공되는 영상에는 영상 효과 및 배경 음악 뿐만 아니라, 음성 속도에 맞게 텍스트가 자막으로 제공된다.

2.3 시스템 구조

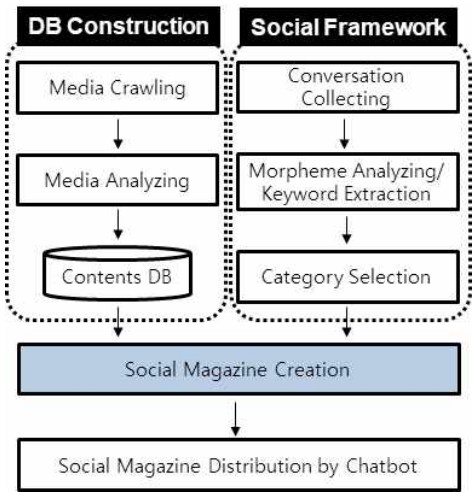

Fig. 1. 챗봇 서비스 구조

그림 1에는 소셜 매거진 생성을 위한 챗봇 서비스의 구조가 나타나 있다. 챗봇은 먼저 사용자 간 대화의 키워드를 찾기 위해, 대화 내용

을 수집 및 분석한다. 키워드 분석이 완료되면, 이 서비스에서 미리 설정된 미디어 콘텐츠 데이터베이스에 대해 만들어질 매거진의 카테고리가 선택된다. 데이터베이스는 뉴스 및 소셜 미디어의 웹 크롤링을 통해 구축하였으며, 크롤링 된 콘텐츠에 대한 형태소 분석을 통한 고유명사 추출을 통해 키워드 셋을 생성하였다. 소셜 매거진은 대화 키워드와 미리 클러스터링 된 카테고리로부터 얻어진 DB의 콘텐츠 집합을 통해 만들어지고, 챗봇에 의해 대화방에 배포된다.

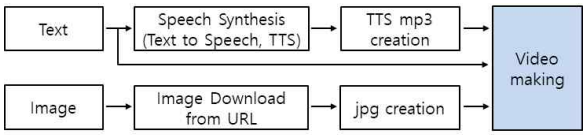

Fig. 2. 영상 합성 서비스 구조

그림 2에는 영상 합성 서비스의 구조가 나타나 있으며, 대화방에 배포되기 전 JSON 형식의 메타데이터로부터 콘텐츠의 텍스트 정보를 통해 음성 합성을 먼저 수행한다. 전체 텍스트를 문장 단위로 분리하고, 각 문장 단위와 합성된 음성 정보의 길이를 분석하여 음성과 자막정보의 싱크(sync)를 맞추게 된다. 영상 합성에 사용되는 이미지에 대해서는 크기 및 종횡비(aspect ratio)를 일정하게 리사이징하여 사용한다.

3. 소셜 매거진 생성 및 영상 합성 모듈

3.1 소셜 매거진 생성 모듈

그림 3은 소셜 매거진 생성 모듈의 구조를 나타내고 있다. DB에서 콘텐츠를 불러오면, 중복 결과 및 아웃라이어 검출 과정이 수행된다. 텍스트로 된 웹 콘텐츠는 내용의 복사, 등록이 쉽게 이루어지기 때문에, 다음 단계 모듈의 효율성 측면에서 중복 결과 감지 모듈이 필요하다. 또한, 아웃라이어 검출 프로세스를 통해 메인 미디어와 아웃라이어 미디어의 콘텐츠 분리가 이루어지고, 분리된 각 그룹에 대해 텍스트 요약이 수행된다.

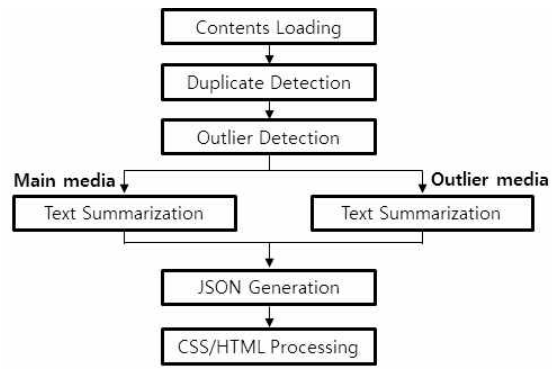

Fig. 3. 소셜 매거진 생성 모듈

아웃라이어 검출은 추출된 키워드에 대해 비슷한 내용을 가지는 콘텐츠만을 가지고 매거진을 만드는 것을 방지하기 위함이다. 아웃라이어 검출 과정이 없는 결과는 정보 제공 측면에서 정확도는 높지만 다양성은 떨어질 수 있다. 본 논문에서는 MeanDIST 알고리즘[5]을 사용하였으며, [6]에서 제안한 알고리즘의 변형된 형태이다.

텍스트 요약 과정은 TextRank[7] 알고리즘과 유사하지만 다양한 문서 또는 주제에 적용할 수 있는 LexRank 알고리즘에 기반하여 수행된다[8]. LexRank는 형태소 분석, TF-IDF, 문장 간 유사도 계산, 그래프 클러스터링 등의 과정을 거치며, 이를 통해 가장 관련성이 높은 문장을 추출 및 콘텐츠 목록을 얻게 된다.

모든 과정이 완료되면 텍스트, 이미지, 템플릿 정보 등이 JSON 형태의 메타데이터로 생성되고, 웹 페이지 형식의 소셜 매거진이 사용자에게 배포된다. 템플릿은 연(年), 월(月), 일(日)의 시간적 흐름에 따른 것과, 나열형, 두 가지 이상의 템플릿의 혼합형 등으로 다양화할 수 있다.

3.2 영상 합성 모듈

그림 4는 영상 합성 모듈의 구조를 나타내고 있다. 먼저 메타데이터에 포함된 텍스트를 가지고 음성 합성을 수행한다. 음성 합성은 네이버에서 제공하는 Clova Speech Synthesis(CSS) 클라우드 서비스를 사용하였다.[9]

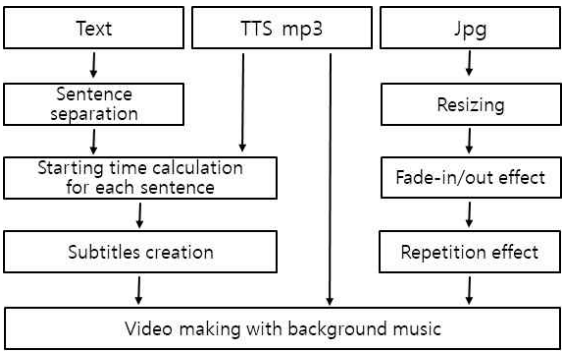

Fig. 4. 영상 합성 모듈

음성 합성 이후, 텍스트를 문장 단위로 분리하고, 각 문장에 해당되는 음성 정보의 시작점을 계산하여 자막 파일을 생성한다. 해당 콘텐츠에 대한 이미지 정보에 대해서는, 여러 이미지가 존재하는 경우 각 이미지를 일정한 크기로 리사이징하여 사용한다. 각 이미지는 영상 내에서 5초 간 재생되며, 이미지 재생시간이 영상 전체의 길이보다 짧은 경우에는 자동으로 이미지의 반복재생이 되도록 설정하였다. 영상에 오디오 합성 및 이미지 반복재생, 페이드인, 페이드아웃 등의 영상 효과와 배경음악 합성은 FFMPEG을 이용하여 수행하였다.

4. 실험 결과

본 논문에서 제안한 챗봇 서비스에 대한 테스트를 텔레그램 메신저에서 수행했다. 사용자가 있는 대화방에서 챗봇을 생성, 초대하고, 사용자 간 대화를 하게 되면, 각 문장마다 챗봇은 사용자들의 대화를 수집 및 분석한다. 3개의 문장이 입력되면 DB로부터 해당 키워드에 해당되는 콘텐츠를 불러와서 소셜 매거진 생성 모듈을 거치면 소셜 매거진이 생성된다.

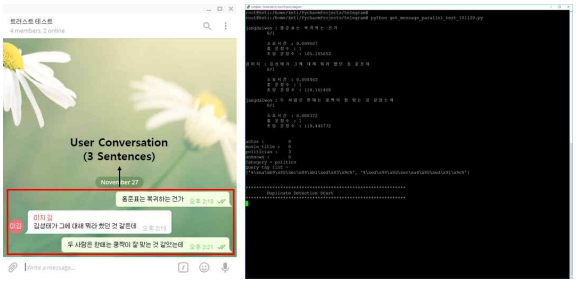

Fig. 5. 사용자 대화 및 데이터베이스로부터 콘텐츠를 불러오는 실험 및 결과

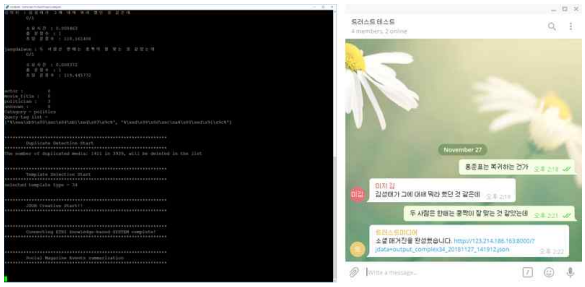

Fig. 6. 텍스트 요약 및 소셜 매거진 생성 모듈 실험 및 결과

그림 5에는 3개의 문장이 입력되었을 때, 대화 내용의 키워드를 분석하고 DB로부터 적절한 콘텐츠를 불러오는 예시가 나타나 있다. 입력된 3개의 문장에 대해 데이터베이스에서 총 3939개의 콘텐츠가 로드되었고, 그 중 중복 콘텐츠는 1411개로 분류되었다. 복합형 템플릿이 선택되어 아웃라이어 검출과정은 생략되었다. 그림 6은 텍스트 요약 과정 및 매거진 생성 이후 챗봇에 의한 매거진 링크 배포 결과를 나타낸다.

Fig. 7. 소셜 매거진 생성 결과

소셜 매거진 생성 및 배포의 결과로, 일별, 월별 형식의 복합 템플릿 결과물이 그림 7과 같이 생성된 것을 확인할 수 있었으며, 시스템이 잘 작동하는지 확인할 수 있었다.

영상 합성 서비스에 대한 테스트에는, 정치 카테고리에서 소셜 매거진을 생성하기 위한 JSON 메타데이터 콘텐츠 중 하나를 사용하였다. 해당 콘텐츠는 총 47문장의 텍스트와 3개의 이미지로 구성되었다. 이미지 URL 정보와 텍스트 정보, 그리고 미리 텍스트 정보에 의해 만들어진 TTS 정보를 입력으로 넣은 결과는 mp4 형식의 영상으로 나오게 된다.

Fig. 8. 영상 합성 생성 결과

위 그림 8은 총 6분 20초의 합성 영상 결과에서 22~26초 사이 페이드아웃/페이드인 순간의 4개 프레임을 추출한 것이다. 그림 8을 보면, 영상 아랫부분에 하나의 문장이 자막으로 입력된 것을 볼 수 있는데, TTS에서 해당 텍스트가 시작되는 시간은 결과 영상에서 약 23.2초였으며, 계산으로 구한 자막이 시작되는 시간은 약 22.3초였다. 따라서 약 0.9초의 차이가 나는 것을 알 수 있었다. 영상이 진행되면서 음성이 시작되는 시작점과 자막이 표시되는 시점이 가끔씩 일치하지 않는 경우가 있었지만, 약 1~1.5초 정도의 차이로, 크게 기준점에서 벗어나지 않는 범위에서 발생하였고, 영상 반복 및 배경 음악의 합성 또한 잘 작동하는 것을 확인할 수 있었다.

5. 결론

본 논문에서는 사용자 간 대화중 대화 주제에 적합한 콘텐츠를 매거진 형태로 제공하는 소셜 매거진 서비스와, 텍스트 및 이미지 정보를 가지고 음성 변환 및 영상 합성 서비스를 제안하였다. 소셜 매거진 서비스에서는 아웃라이어 검출을 위한 MeanDIST, 텍스트 요약을 위한 LexRank 알고리즘을 사용하여 키워드 추출, 중복 문서 및 아웃라이어 탐지 모듈 등의 일련의 프로세스를 통해 제안 플랫폼의 타당성을 검증할 수 있었다. 또한 영상 합성 서비스에서는 텍스트를 음성으로 변환해주는 Text to Speech(TTS), 영상 효과 및 음성 처리에 대한 FFMPEG을 사용하여 제안 서비스의 동작을 확인할 수 있었다.

텔레그램 메신저 챗봇을 이용한 소셜 매거진 생성과 영상 합성 서비스의 경우, 제안하고자 하는 시스템적 측면에서는 잘 동작하는 것이 검증되었지만, 처리 시간 등의 측면에서는 개선의 필요성이 있었다.

감사의 말

이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.2017-0-00318, 건전한 미디어 소비환경 제공을 위한 소셜 IoM 기반 트러스트 미디어 생성·제어 프레임워크 기술 개발)

참고문헌

[1] A. Xu, Z. Liu, Y. Guo, V. Sinha, and R. Akkiraju,

"A new chatbot for customer service on social media," in *Proc. of the ACM conference on Human Factors in Computing system*, 2017

[2] P. B. Brandzaeg and A Folstad, "Why people use chatbots" in *Proc. of the International conference on Internet Science*, 2017

[3] P. M. ee, S. Santra, S. Bhowmick, A. Paul, P. Chatterjee and A. Deyasi, "Development of GUI for Text-to-Speech Recognition using Natural Language Processing," *2018 2nd International Conference on Electronics, Materials Engineering &Nano-Technology (IEMENTech)*, Kolkata, 2018, pp. 1-4.

[4] A. F. Jalin and J. Jayakumari, "Text to speech synthesis system for tamil using HMM," *2017 IEEE International Conference on Circuits and Systems (ICCS)*, Thiruvananthapuram, 2017, pp. 447-451.

[5] V. Hautamaki, I. Karkkainen, P. Franti, "Outlier detection using k-nearest neighbour graph, " in *Proc. of the 17th International Conference on Pattern Recognition*, Los Alamitos, CA, USA, 2004, pp. 430-433

[6] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," In *Proc. of the 2000 ACM SIGMOD Int. Conf. on Management of Data*, pages 427-438, Dallas, Texas, May 2000

[7] R. Mihalcea, P. Tarau, "Textrank: Bringing order into text" in *Proc. of the 2004 conference on empirical methods in natural language processing,* 2004

[8] G. Erkan, D.R. Radev, "Lexrank: Graph-based lexical centrality as salience in text summarization," *Journal of artificial intelligence research*, 2004

[9] [online] https://www.ncloud.com/product/aiService/css