이용자 태그를 활용한 비디오 스피치 요약의 자동 생성 연구*

Investigating an Automatic Method in Summarizing a Video Speech Using User-Assigned Tags

김 현 희(Hyun-Hee Kim)**

목 차

초 록

본 연구는 스피치 요약의 알고리즘을 구성하기 위해서 방대한 스피치 본문의 복잡한 분석 없이 적용될 수 있는 이용자 태그 기법, 문장 위치 및 문장 중복도 제거 기법의 효율성을 분석해 보았다. 그런 다음, 이러한 분석 결과를 기초로 하여 스피치 요약 방법을 구성, 평가하여 효율적인 스피치 요약 방안을 제안하는 것을 연구 목적으로 하고 있다. 제안된 스피치 요약 방법은 태그 및 표제 키워드 정보를 활용하고 중복도를 최소화하면서 문장 위치에 대한 가중치를 적용할 수 있는 수정된 Maximum Marginal Relevance 모형을 사용하여 구성하였다. 제안된 요약 방법의 성능은 스피치 본문의 단어 빈도 및 단어 위치 정보를 적용하여 상대적으로 복잡한 어휘 처리를 한 Extractor 시스템의 성능과 비교되었다. 비교 결과, 제안된 요약 방법을 사용한 경우가 Extractor 시스템의 경우 보다 평균 정확률은 통계적으로 유의미한 차이를 보이며 더 높았고, 평균 재현율은 더 높았지만 통계적으로 유의미한 차이를 보이지는 못했다.

ABSTRACT

We investigated how useful video tags were in summarizing video speech and how valuable positional information was for speech summarization. Furthermore, we examined the similarity among sentences selected for a speech summary to reduce its redundancy. Based on such analysis results, we then designed and evaluated a method for automatically summarizing speech transcripts using a modified Maximum Marginal Relevance model. This model did not only reduce redundancy but it also enabled the use of social tags, title words, and sentence positional information. Finally, we compared the proposed method to the Extractor system in which key sentences of a video speech were chosen using the frequency and location information of speech content words. Results showed that the precision and recall rates of the proposed method were higher than those of the Extractor system, although there was no significant difference in the recall rates.

키워드: 스피치 요약, 태그, 비디오, 표제, 코싸인 유사계수, 내재적 평가, 적합 문장
MMR Model, Social Summarization, Redundancy, Acoustic features, Prosodic Features, Extractor, Transcripts

---

* 이 논문은 2010년도 정부재원(교육과학기술부 인문사회연구역량강화사업비)으로 한국연구재단의 지원을 받아 연구되었음(NRF-2010-327-H00017).

** 명지대학교 문헌정보학과 교수(kimhh@mju.ac.kr)

논문접수일자: 2012년 1월 16일 최초심사일자: 2012년 1월 20일 게재확정일자: 2012년 2월 8일
한국문헌정보학회지, 46(1): 163-181, 2012. [http://dx.doi.org/10.4275/KSLIS.2012.46.1.163]

1. 서 론

1.1 연구 배경과 목적

대학이나 개인에 의해서 제공되는 강의 및 발표와 같은 스피치 자료들은 인터넷 사이트들을 통해서 이용가능하지만 이러한 교육 정보원이 효율적으로 활용되지 못하는 경우도 많다(Liu & Hakkani-Tur 2011). 스피치 요약은 이러한 정보원에 대한 효율적인 접근을 가능하게 해 줄 수 있을 것이며 이는 장기적으로 볼 때 교육에 긍정적인 영향을 미칠 것으로 생각된다. 또한 스피치 요약은 현재 이용자들이 많이 사용하고 있는 작은 사이버 공간을 갖는 태블릿 PC 또는 스마트폰에 효율적으로 적용될 수 있는 이점이 있다. 이외에 스피치 요약은 비디오 내용을 파악하는데 있어서 이미지 및 스피치 요약이 결합된 멀티모달(multi-modal) 요약 만큼 성능이 좋다는 연구 결과도 있다(Song & Marchionini 2007). 이외에 이미지와 스피치 요약이 결합된 멀티모달 요약을 시청한 후에 멀티모달 요약내의 스피치 요약이 이미지 요약 보다 비디오 의미 파악을 하는데 더 유용하게 사용되었다는 연구 보고도 있다(김현희 2011).

스피치 요약은 크게 두 종류로 나눌 수 있다. 먼저 일반화된 법칙에 기초한 비지도(unsupervised)(규칙 기반) 방법과 변수간의 연관성이나 유사성에 기초한 지도(supervised)(특징 기반) 방식이 있다. 비지도 방식에는 중복은 최소화하면서 적합성과 다양성을 최대화하는 MMR (Maximum Marginal Relevance) 모형 기반 방식, 단어와 단어, 단어와 문장 및 문장과 문장 사이의 의미 유사성 추정치를 제공하는 분석 기법인 잠재의미 분석(Latent Semantic Analysis) 등이 있다. 지도 방식에는 문장 위치와 같은 구조 정보를 이용하는 방식, 단서어 기법 또는 스피치의 피치(pitch)나 길이와 같은 음향학/운율적인 특징을 이용하는 방식들이 있다. 스피치 요약의 최근 연구 경향을 살펴보면 어떤 하나의 방법을 적용하기 보다는 위에서 언급한 방법들을 통합하여 사용하는 쪽을 지향하고 있다(Chen & Lin 2012).

많은 연구자들(Zhu, Penn, & Rudzicz 2009; Maskey & Hirschberg 2006; Xie et al. 2009)은 스피치의 음향학/운율적인 특징을 이용하는 방식들에 대하여 연구하고 있다. 그러나 이러한 음향 기반 연구는 동일한 화자의 문헌들만 이용할 수 있다는 점에서 모든 경우에 적용하는데 한계가 있다. 또한 현재로서는 음향학/운율적인 정보를 스피치 요약에 효율적이면서 신뢰할 수 있게 적용하기에 아직까지 기술적인 어려움이 있다(Zhu, Penn, & Rudzicz 2009; Chen & Lin 2012). 더 나아가, Zhang et al.(2007)은 강의 자료에서는 어휘적 특징이 음향학 특징 보다 더 중요하게 나타났다고 보고하고 있다. 따라서 이러한 음향학/운율적인 정보를 활용하는 방안을 계속 모색하면서 텍스트 요약에 성공적으로 적용된 기법들을 텍스트와 본질적으로 다른 스피치 요약에 얼마나 효율적으로 사용할 수 있는지 조사하는 것이 필요한 시점이 되었다.

본 연구는 먼저 방대한 비디오 스피치 자막에 적용하기 어려운 심층적인 어휘 처리가 요구되지 않은 태그 기법, MMR 기법, 문장 위치 기법 등이 스피치 요약에 얼마나 효율적으로 사용될 수 있는지 분석해 보고자 한다. 그런 다음 이러한 분석 결과를 기초로 하여 비디오의 스피치

요약을 자동으로 추출하기 위한 알고리즘을 제안하고, 제안된 알고리즘의 성능을 스피치 본문의 단어 출현 및 위치 정보를 적용하여 상대적으로 복잡한 어휘 처리를 한 Extractor 시스템의 성능과 비교하여 효율적인 스피치 요약 방안을 제안하는 것을 연구 목적으로 하고 있다.

1.2 연구 문제와 방법

본 연구에서 조사하고자 하는 연구 문제는 다음과 같이 여섯 가지이다.

연구문제 1: 비디오 태그와 표제의 키워드간의 일치율은 어느 정도 인가이다. 이용자들이 태그를 할당할 때 비디오 표제와 같은 메타데이터를 참조하는 경향이 있기 때문에 이들간의 일치율이 매우 높으리라 가정한다.

연구문제 2: 비디오 태그와 텍스트 본문의 키워드간의 일치율은 어느 정도 인가이다. 비디오 태그와 텍스트 본문의 키워드 간의 일치율은 높을 것으로 가정하지만 태그와 표제 키워드간의 일치율 보다 더 낮을 것으로 예측한다.

연구문제 3: 비디오의 스피치 대본을 자동적으로 요약하기 위해서 이용자 태그가 얼마나 유용한지 조사해 본다. 이를 위해서 태그에 기반한 스피치 요약이 본문 키워드에 기반한 스피치 요약간에 요약문 품질면에서 어떤 차이를 나타내고 있는지 조사해 본다. 기본 가정은 유투브 비디오에 할당된 태그들이 각 비디오에 포함된 다양한 주제를 나타내는 좋은 지시자임을 가정한다. 이러한 가정을 뒷받침하는 연구에는 소셜 태그가 웹 페이지에서 멀티 주제를 요약하는데 있어서 문헌 내용에 대한 좋은 보완을 제공하고 있음을 보여주는 Zhu et al.(2009)의 연구가 있다. 또한 이용자들이 비디오에 태깅을 하는 이유는 개인 수집물을 관리하기 위해서 보다는 가능한 많은 사람들이 그들이 등록한 비디오들을 검색하고 브라우징할 수 있도록 하기 위해서라고 주장하는 Heckner, Neubauer 및 Wolff(2008)의 연구, 그리고 비디오 태그는 색인작성자가 할당한 용어와 유사하며 비디오의 객체 또는 인물을 기술하는데 유용하게 사용될 수 있다고 제안하는 Kim(2011)의 연구가 있다.

연구문제 4: 스피치 요약에서 문장 위치 정보의 중요도는 어느 정도인가이다. 위치 정보의 중요도는 비디오의 장르에 따라서 다를 것으로 가정해 볼 수 있는데 발표 및 강의 비디오에서는 위치 정보가 중요한 기준이 될 것으로 예측된다. 왜냐하면 일반적으로 이런 종류의 비디오에서는 비디오 개요 및 논점을 비디오 시작 부분에서 언급하는 경향이 있기 때문이다. Christensen et al.(2003)은 텍스트 및 스피치 요약에서 개별적인 특징들의 효과를 비교하였다. 비교 결과, 텍스트에서는 문장 위치가 가장 중요한 기준으로 나타났으나, 스피치에서는 절대적인 기준은 없었다. 이외에 Liu와 Hakkani-Tur(2011)는 텍스트에서, 일반적으로 전체적인 개요가 문헌의 앞부분에 나타나지만, 스피치에서는 정보가 분산되는 경향이 있다고 보고하고 있다. 이와는 달리, Hirohata et al.(2006)는 발표 스피치 자료를 10%로 요약하는 실험에서 사람들이 수작업으로 비디오의 마지막 부분에서 주로 문장을 선택하는 경향을 발견하고 문장 위치가 스피치 요

약의 결과를 개선시키는데 이용될 수 있다고 기술하고 있다. Zhang et al.(2007)는 구조적 특성 특히 위치 정보가 방송 뉴스의 발췌 요약을 위해서 가장 유용한 예측 요인임을 발견하였다. 따라서 위치 정보가 스피치 요약에 중요한 요인인지 또는 비디오의 장르에 따라서 그 중요도가 어떻게 달라지는지 좀 더 자세한 조사가 필요해 보인다.

연구문제 5: 스피치 요약에서 선택된 적합성이 높은 문장간의 유사도가 어느 정도인가이다. 스피치 비디오 자료는 텍스트 자료 보다 중복된 내용이 많을 수 있으며, 이에 따라서 태그 및 표제의 키워드에 기반하여 선택된 문장들은 서로 의미적으로 유사한 경우가 많을 것이다. 다시 말해서, 비디오 대본은 텍스트와는 달리 단순히 콘텐츠를 전달하기 보다는 비주얼 또는 오디오 수구반복(anaphora, 여러 개의 문장이나 절의 첫 부분에서 한 낱말이나 구를 반복하는 방법)이 비디오 채널을 위한 문맥과 동기를 제공하기 위해서 사용되는 경우가 있기 때문에 선택된 문장간의 유사도가 높을 것으로 예측된다.

연구문제 6: 최종적으로 이러한 연구 문제들의 조사 결과를 기반으로 효율적인 스피치 요약 방안을 제안하고 평가해 본다.

이러한 연구 문제들을 조사하기 위한 표본 비디오 자료는 음성으로 많은 정보를 표현하는 강의, 교육 및 연설 비디오로 정하고 유투브 사이트에서 24개의 영어로 된 비디오들을 선정하였다. 표본 비디오로 영어 비디오를 선정한 이유는 소셜 태그를 풍부하게 포함하고 있기 때문이며, 또한 본 연구에서 제안한 스피치 요약 알고리즘은 언어에 관계없이 모두 적용될 수 있기 때문이다. 스피치 요약의 효율성을 평가하기 위해서 요약 기법의 성능을 평가하는 내재적 평가를 한다(정영미 2005). 내재적 평가를 위해서 연구팀은 스피치 대본에서 비디오의 의미를 가장 잘 나타내는 문장들을 추출하여 표준 요약을 구성한다(자세한 설명은 3.1 표본 비디오 참조). 통계 분석을 위해서 SPSS 통계 패키지를 사용한다.

2. 선행 연구

선행 연구는 스피치 요약과 소셜 요약으로 구분하여 국내외 연구들을 살펴보았다. 스피치 요약에서는 비디오 스피치 대본을 대상으로 한 요약 방법들을 다룬 연구들을 조사하였고, 소셜 요약에서는 문헌이나 비디오를 요약하기 위해서 태그, 댓글, 북마크 및 트위터 메시지를 활용한 연구들을 살펴보았다.

2.1 스피치 요약

Zechner(2002)는 MMR 모형에 기반한 대화 스피치를 위한 요약 시스템을 구성하고, 이 요약 시스템이 여러 다른 데이터 집합에서 좋은 성능을 보여 주었다고 보고하고 있다. Murray, Renals 및 Carletta(2005)는 회의 비디오 요약에서 MMR 기법이 중복을 줄이는 기능 때문에 잠재의미 분석 및 특징 기반(운율) 방법과 비교하여 유사한 성능을 보여 주었다고 보고하였다. Zhang et al.(2007)는 방송 뉴스와 강의 자료

라는 두 개의 장르를 비교하여 스피치 요약을 위한 음향학/운율적인, 언어학적 및 구조적 특징에 대한 비교 연구를 수행하였다. 비교 결과, 방송 뉴스의 경우 앵커와 리포터의 이야기하는 유형과 전형적인 뉴스 스토리의 흐름 때문에 음향학 및 구조적 특징이 어휘적 특성 보다 중요한 반면, 강의 자료에는 어휘적 특징이 다른 두 개의 특징 보다 더 중요하게 나타났다.

Xie와 Liu(2008)는 회의 자료 요약에서 MMR 모형에서 다양한 유사도 측정 방안들을 평가해 보았다. 평가 결과, 동시 빈도를 이용하여 측정하는 코퍼스 기반 유사도 측정 방법이 코싸인 유사도 측정 방법 보다 성능면에서 더 우수한 것으로 나타났다. Marchionini, Song 및 Farrell (2009)는 MAGIC 시스템을 이용하여 자동으로 생성한 오디오 요약을 구성하였다. MAGIC 시스템은 비디오 스피치 자막에서 문장을 추출한 후 등급을 매기기 위해서 담화 세분화(discourse segmentation)와 주제 변환 탐지(topic shift detection)를 포함한 다양한 요약 기법들을 활용하였다. Marchionini et al.은 자동으로 생성된 오디오 요약이 비디오 검색과 비디오 내용 파악의 목적으로 사용될 때 비록 수작업으로 생성한 오디오 요약 보다는 성능면에서 떨어지기는 하지만 비디오의 의미 파악을 적절하게 지원하고 있다고 기술하고 있다.

이한성 외(2010)는 뉴스 비디오 클립과 스크립트를 동시에 이용하는 멀티모달 방법론과 텍스트 마이닝 기반의 뉴스 비디오 마이닝 시스템을 제안하였다. 제안된 시스템은 텍스트 마이닝의 군집분석을 통해 뉴스 기사들을 자동 분류하고, 분류 결과에 대해 기간별 군집 추이그래프, 군집성장도 분석 및 네트워크 분석을 수행함으로써, 뉴스 비디오의 기사별 주제와 관련한 다각적 분석을 수행할 수 있었다. Chen과 Lin(2012)는 스피치 요약을 리스크 최소화 문제로 공식화하고 다양한 요약 방식들의 장점은 살리고 제한점은 극복하기 위해서 지도 및 비지도 요약 방법들을 결합하는 통합된 확률적 프레임워크를 제안하였다. 제안된 프레임워크에 기초하여 구성한 요약 방법들을 단순히 문헌의 처음 부분에 나타나는 몇 개의 문장들을 선택하는 LEAD 방식, 벡터간의 유사도에 기초한 벡터공간모형(VSM) 등과 같은 기존 요약 방법들과 비교한 결과, 제안된 방법들의 성능이 더 우수한 것으로 나타났다.

2.2 소셜 요약

Zhu et al.(2009)은 태그의 희박성 문제를 해결하기 위해서 연관 마이닝 기술을 이용하여 태그를 확장하고 태그에서의 잡음을 줄이는 태그 랭킹 알고리즘을 이용한 태그 기반 웹 문서 요약 방법을 제안하였다. Zhu et al.의 실험 결과는 태그 기반 요약이 태그를 사용하지 않은 다른 기법들에 비해서 상당한 성능 개선을 가져다 주었다고 기술하고 있다. 김현희(2009)는 텍스트 요약에 적용된 이론과 방법이 오디오 요약에도 적용될 수 있을 것이라는 가정하에 오디오 내용을 압축하기 위해서 위치 정보, 표제 이외에 소셜 메타데이터인 태그와 댓글 정보를 이용하였다. 이 연구는 좀 더 효율적인 오디오 요약을 구현하기 위해서는 오디오 정보의 특성에 맞춘 요약 기법에 대한 연구가 요망된다고 기술하고 있다.

Boydell과 Smyth(2010)는 문헌 요약을 위

해서 소셜 북마크 태그 및 탐색 질의 정보를 사용하는 방안을 제안하였다. Boydell et al.는 제안된 방안을 표층적 자연언어처리기법을 통계적 단어 빈도 방법과 결합한 방식을 채택한 두 개의 텍스트 요약 시스템 즉, 오픈 텍스트 요약기(Open Text Summarizer)와 MEAD 시스템과 비교하였다. 비교 결과, 제안 방안이 두 개의 시스템들 보다 더 높은 품질의 요약문을 생성하는 것으로 나타났다.

Chung, Wang 및 Sheu(2011)는 이용자에 의해서 북마크된 비디오 프레임들은 비디오 내용을 잘 표현할 수 있을 것이라는 가정하에 비디오 스토리보드(이미지 요약)를 구성하기 위해서 통계적으로 비디오 북마크들을 분석한 후에 의미 있는 키프레임들을 추출하였다. Chung et al.의 제안 방법이 낮은 수준의 오디오-비주얼 특성을 활용한 기존 방법들 보다 의미적으로 더 중요한 요약문들을 생성하는 것으로 나타났다. 끝으로, Hannon et al.(2011)는 시간 표시 트위터 메시지의 빈도와 내용에 기초하여 월드컵 비디오의 하이라이트들을 생성하는 두 가지 방안을 제안하고 평가해 보았다. 평가 결과, 이용자들은 두 가지 방안에 의해서 생성된 요약에 대해서 광범위한 만족감을 나타냈다.

선행연구들을 살펴본 결과, 강의 및 교육 비디오 자료의 스피치 요약을 구성하는데 어휘 및 구조 정보 측면에서 효율성이 높은 요인들이 어떤 것들이 있는지 체계적인 연구가 부족한 편이다. 또한 자동 스피치 요약의 연구 경향이 어떤 하나의 요인을 적용하기 보다는 효과적인 요인들을 결합하여 사용하는 쪽을 지향하고 있다. 따라서 본 연구에서는 세밀한 분석을 통해서 어휘 및 구조 정보 중 어떤 요인들이 효율적인지 조사한 후에 이를 기초로 하여 스피치 요약을 생성하기 위한 알고리즘을 제안해 보고자 한다.

3. 비디오 태그 및 스피치 내용 분석

비디오 태그 분석에서는 앞에서 언급한 연구 문제 1(비디오 태그와 표제의 키워드간의 일치율 정도), 연구 문제 2(비디오 태그와 텍스트 본문의 키워드간의 일치율 정도) 및 연구 문제 3(태그에 기반한 요약문과 본문 키워드에 기반한 요약문간의 품질 비교)를 조사하였다. 한편 비디오 스피치 내용 분석에서는 연구 문제 4(문장 위치 정보의 중요도)와 연구 문제 5(선정된 문장간의 유사도 정도)를 조사하였다.

3.1 표본 비디오

비디오 태그 및 스피치 내용 분석을 위해서 24개의 영어로 된 표본 비디오를 유투브에서 선정하였다(〈표 1〉 참조). 표본 비디오의 선정 기준은 음성으로 많은 정보를 표현하는 강의, 교육 및 연설 비디오로 4개 이상의 태그를 갖고 있으면서 재생시간이 4분~25분 사이에 있는 것들을 선택하였다. 이로써 표본 비디오의 수는 24개로 하였는데 비디오 태그 및 스피치 내용 분석을 위해서 최소한 20개 이상의 표본 자료가 필요하다고 생각되었기 때문이다.

또한 태그 기반 요약 방식의 효율성을 평가하기 위해서 각 비디오의 표준 요약을 두 명의 연구자들이 공동으로 작성하였다. 작성 방법은 연구자들이 비디오를 시청한 다음 유투브 사이트에 있는 텍스트 요약 및 메타데이터 내용을

〈표 1〉 표본 비디오 목록

No.	표 제(재생시간(분:초))	태그수		No.	표 제(재생시간(분:초))	태그수	
		단일어	복합어			단일어	복합어
1	Meet the Mentor(05:09)	4	0	13	Sunni Brown: Doodlers, unite!(05:51)	7	2
2	Disability Services at ASU Libraries(10:05)	9	0	14	Amber Case: We are all cyborgs now(08:24)	6	1
3	Kate Lundy: What I do for Open Government(05:01)	10	6	15	Tim Berners-Lee: The next Web of open, linked data16:51	27	13
4	Learning English-Lesson Forty Three(Superstition)(10:41)	9	0	16	Tim Berners-Lee: The year open data went worldwide(06:04)	17	2
5	How-to Make Mosaic Art: How to Grout a Mosaic Work of Art(05:42)	8	1	17	Malcolm Gladwell: What we can learn from spaghetti sauce(18:16)	8	3
6	Thrive Food Storage: Spaghetti and Meatball(10:48)	1	6	18	Bill Gates: How state budgets are breaking US schools(10:47)	36	12
7	Application and Preparation of Limewash(06:27)	7	3	19	Al Gore: 15 ways to avert a climate crisis(16:58)	9	2
8	Experience The Learning Connexion(09:46)	2	2	20	Clay Shirky: How cellphones, Twitter, Facebook can make history17:03	12	1
9	the Department of Dance at California State University, Long Beach(CSULB)(10:00)	68	1	21	Clay Shirky: How cognitive surplus will change the world13:39	11	2
10	Washtenaw Community College School of Culinary Arts(04:41)	14	0	22	David Cameron: The next age of government14:34	14	1
11	Steve Jobs' 2005 Stanford Commencement Address(15:04)	14	0	23	Dan Ariely asks, Are we in control of our decisions?(17:27)	28	8
12	President Clinton 1997 Inaugural Address(22:57)	7	0	24	Graham Hill: Less stuff, more happiness(05:50)	7	2

세밀히 분석한 후 비디오 자막을 마침표를 기준으로 문장 단위로 구분한 후 비디오의 내용을 가장 잘 나타내는 문장들을 선정하였다. 선정된 문장 중에 두 명의 연구자가 똑같이 선정한 문장은 그대로 사용하고 서로 다른 문장을 선정한 경우는 서로 상의하여 최종적으로 적합한 문장을 선택하여 표준 요약을 구성하였다.

24개의 표본 비디오를 분석한 결과, 태그 평균수가 16.8개로 나타났다. 구체적으로, 단일어 태그의 평균수는 14.0개, 복합어 태그의 평균수는 2.8개이다. 비디오 5번은 가장 많은 69개의 태그를 갖고 있으며, 비디오 1번과 8번은 가장 적은 4개의 태그를 각각 갖고 있다.

3.2 비디오 태그 분석

3.2.1 태그와 표제 및 본문 키워드간의 비교 분석

키워드를 자동 또는 반자동으로 생성하는 방법에는 크게 두 가지가 있다. 첫째는 표제, 초록 및 본문에서 단어 빈도와 단어간의 동시 출현 빈도를 기초로 하여 단일어와 복합어를 키워드로 추출하는 방법이 있다. 둘째로 이용자들에 의해서 생성되는 소셜 태그를 활용하는 방안이다.

본 연구에서는 소셜 태그와 표제의 키워드간의 의미적 관계 그리고 태그와 텍스트 본문에서 빈도에 기초하여 생성한 키워드간의 의미적 관계를 분석하여 소셜 태그의 주제어로서의 가치를 평가해 보고자 한다. 이를 위해서 표본 비디오 24개의 각각의 비디오에서 태그와 표제의 키워드간의 일치율 그리고 태그와 본문 텍스트의 키워드간의 일치율을 분석해 보았다. 비디오 대본 텍스트에서 키워드를 추출하기 위해서 텍스트 요약 시스템인 Extractor(http://www.extractor.com)을 이용하였다(3.2.2에서 자세히 설명).

분석 결과, 태그와 표제의 키워드간의 일치율이 0.70, 태그와 본문 텍스트 키워드간의 일치율이 0.22로 나타났다. 또한 태그/표제 키워드와 본문 텍스트의 일치율은 0.25로 나타나 태그만을 사용한 경우(0.22) 보다 조금 높게 나타났다. 이러한 결과로 비추어 볼 때 태그가 표제의 키워드를 모두 포함하지 못하고 있음을 확인할 수 있었다. 또한 태그와 본문 텍스트의 키워드간의 일치율(0.22)은 처음 기대한 것 보다 낮게 나타났다. 따라서 태그 기반 스피치 요약과 본문 키워드 기반 스피치 요약은 서로 다르게 나타날 수 있음을 예측해 볼 수 있다.

3.2.2 태그 기반 요약과 본문 키워드 기반 요약간의 질적 비교 분석

다음 단계로 24개의 각 비디오의 표준 요약을 기준으로 태그 기반 요약과 본문 키워드 기반 요약간의 품질이 어떻게 다른지 재현율과 정확률 측면에서 서로 비교해 보기로 하였다. 요약문을 구성할 때 태그만을 사용할 경우 태그의 희박성 때문에 주요 문장을 추출하는데 어려운 경우가 있었다. 따라서 만약 표제의 키워드가 태그에 포함되어 있지 않으면 이를 태그 집합에 포함시켜 분석하였다.

구체적인 분석 방법은 다음과 같다. 먼저 각 비디오 대본을 마침표를 기준으로 하여 여러 개의 문장들로 구분한다. 그런 다음, 태그/표제 키워드 기반 요약 방식에서는 표제 단일어(표제의 키워드는 모두 단일어로 처리함)와 태그 집합(단일어와 복합어)을 기준으로 전체 용어 집합을 구성한 후 이를 용어 벡터로 표시한다. 다음은 이를 이용하여 각 문장과 태그/표제 키워드(용어 벡터)간의 유사도를 코싸인 유사계수를 이용하여 계산한다. 이때 용어의 가중치는 표제 단일어에 1.5를, 태그 단일어, 태그 복합어에 각각 1.0, 2.0을 할당하였다. 용어의 빈도에 대한 가중치는 빈도가 두 번 이상인 경우에는 일률적으로 원래 가중치에 2를 곱하였다. 최종적으로 코싸인 유사계수가 높은 문장 5~7개를 선정하여 요약문을 구성하였다.

다음으로, 본문 키워드 방식에서는 Extractor 시스템을 이용하였다. Extractor 시스템은 단일어를 어미를 제거한 어간으로 변환한 후, 각 어간의 출현빈도와 처음 어간이 출현한 위치 정보를 고려하여 핵심 어간을 추출하였다. 즉 높은 빈도를 갖고 있으면서 첫 번째 어간이 전체 텍스트의 앞부분에 출현하는 어간에 더 높은 가중치를 부여하는 방법을 사용하였다(Turkey 2000). 최대 3개까지 단일어가 결합되도록 한 복합어도 어간으로 변환한 후 단일어와 유사하게 처리하였다. 이와 같이 각 비디오 대본의 핵심 단일어와 복합어를 추출한 후 이를 용어 벡터로 표시한 후 이를 기준으로 각 문장과 본문 키워드(용어 벡터)간의 유사도를 코싸인 유사계수를 이

용하여 계산하였다. 이때 용어의 가중치는 단일어, 복합어에 1.0, 2.0을 각각 할당하였다. 다음 작업은 앞의 경우와 동일하게 적용하여 최종적으로 코싸인 유사계수가 높은 문장 5~7개를 선정하여 요약문을 구성하였다.

다음은 두 가지 방식에 의해서 생성된 각 요약문과 표준 요약문과 비교하였다. 이들간의 비교를 위해서 요약 재현율(측정하고자 하는 요약문 내 적합문장 수 / 표준 요약문의 요약 문장 총 수)과 요약 정확률(측정하고자 하는 요약문 내 적합문장 수 / 측정하고자 하는 요약문의 요약 문장 총 수)을 측정하였다. 측정된 내재적 평가 결과는 〈표 2〉에 기술하였다. 태그/표제 기반 요약문이 본문 키워드 기반 요약문과 비교하여 재현율은 더 높은 것으로 나타났으나 맨-위트니(Man-Whitney) 검증 결과 통계적으로 유의미한 차이를 보이지는 못했다(0.35 vs. 0.25, $p(=0.099) > 0.05$). 또한 정확률도 태그/표제 기

〈표 2〉 태그/표제 키워드와 본문 키워드 요약간의 비교

비디오 번호	태그 및 표제 키워드 기반		본문 키워드 기반	
	재현율	정확률	재현율	정확률
1	0.60	0.50	0.20	0.17
2	0.33	0.33	0.17	0.17
3	0.75	0.50	0.50	0.33
4	0.14	0.20	0.17	0.17
5	0.20	0.17	0.20	0.14
6	0.67	0.33	0.33	0.14
7	0.25	0.17	0.0	0.0
8	0.50	0.33	0.50	0.33
9	0.25	0.14	0.50	0.29
10	0.60	0.50	0.40	0.33
11	0.38	0.50	0.0	0.0
12	0.14	0.14	0.29	0.29
13	0.25	0.14	0.75	0.43
14	0.40	0.29	0.40	0.29
15	0.50	0.33	0.25	0.17
16	1.0	0.43	0.0	0.0
17	0.20	0.14	0.20	0.14
18	0.0	0.0	0.0	0.0
19	0.25	0.17	0.0	0.0
20	0.0	0.0	0.25	0.17
21	0.40	0.29	0.20	0.17
22	0.33	0.17	0.0	0.0
23	0.17	0.14	0.17	0.14
24	0.0	0.0	0.0	0.0
평균 (표준편차)	0.35 (0.25)	0.23 (0.20)	0.25 (0.16)	0.16 (0.13)

반 요약문이 더 높은 것으로 나타났으나 맨-위트니 검증 결과 통계적으로 유의미한 차이를 보이지는 못했다(0.23 vs. 0.16, $p(=0.082) < 0.05$).

3.3 스피치 내용 분석

3.3.1 문장 위치의 중요도

문장의 위치 정보 중요도를 파악하기 위해서 24개의 표준 요약을 분석하였다. 먼저 표준 요약에서 선택된 문장들의 위치 정보를 분석해 보았다. 먼저, 비디오 스피치의 서두 및 마지막 부분의 5% 그리고 중간 부분 90%에 속한 표준 요약내의 문장의 할당 비율은 표준 요약의 25.9%가 스피치의 처음(5%)에 기술되고 있었고, 표준 요약의 6.8%가 마지막(5%)에 기술되고 있었다. 이는 비디오의 처음 부분(5%)에 표준 요약내의 문장들이 원래 비율 보다 5배 이상으로 많게 분포되어 있음을 의미한다. 두 번째는 비율을 처음(10%), 중간(80%), 끝부분(10%)로 구분하여 표준 요약의 분포를 살펴보았다. 평균적으로 전체 비디오에서 표준 요약의 30.5%가 스피치의 처음(10%)에 기술되고 있었고, 표준 요약의 9.5%가 스피치의 마지막(10%)에 기술되고 있었다. 이와 같이 비디오 처음 부분의 비율을 5%에서 10%로 상향 조정하니 표준 요약내의 문장의 할당 비율이 5배에서 3배로 줄어들었다. 본 연구 결과는 Hirohata et al.(2006)의 연구 결과와는 다르게 나타났다. Hirohata et al.은 즉흥 스피치 자료를 10%로 요약하는 실험에서 사람들은 비디오의 마지막 부분(10%)에서 주요 문장의 20% 이상을 선택하고, 그 다음으로 처음 부분(10%)에서 대략 14% 정보를 선택하는 경향을 발견하고 문장 위치가 스피치 요약의 결과를 개선시키는데 이용될 수 있다고 기술하고 있다. 이와 같이 스피치 자료의 문장 위치 정보가 요약문을 구성하기 위한 중요한 기준이 될 수 있음을 확인하였다(〈표 3〉 참조).

3.3.2 중복도

태그/표제 기반 방법에 의해서 구성된 24개 비디오의 각 요약문에서 문장간의 유사도가 0.80 이상인 문장을 한 쌍 이상 가지고 있는 요약문을 조사해 보았다. 문장간의 유사도는 문장을 태그/표제 키워드 벡터로 표현한 후 코싸인 유사계수를 이용하여 측정하였다. 전체 24개 요약문 중 83%(20개)의 요약문이 이러한 기준에 적용되었다. 특히 비디오 3의 경우 선택된 7개의 문장 중에 유사도가 1.0인 경우가 세 쌍이나 되었다. 심지어 비디오 13에서는 동일한 문장이 두 번 선택되기로 하였는데 이는 비디오 스피치에서 동일한 문장을 두 번 반복해서 사용했기 때문이다. 이러한 분석 결과를 볼 때 중복된 문장을 제거하는 작업이 필요해 보인다.

〈표 3〉 표준 요약의 문장 위치 정보

비디오 스피치 비율	처음 (5%)	중간 (90%)	끝부분 (5%)	처음 (10%)	중간 (80%)	끝부분 (10%)
표준 요약 분포 비율	25.9%	67.3%	6.8%	30.5%	60.0%	9.5%

3.4 논의

비디오 태그분석에서는 태그와 표제의 키워드간의 일치율이 0.70, 태그와 본문 텍스트의 키워드간의 일치율이 0.22로 나타났다. 태그가 표제 키워드의 30%를 포함하지 못하고 있음을 확인할 수 있었다. 태그/표제 키워드와 본문 키워드간의 일치율(0.25)은 처음 기대한 것 보다 훨씬 낮게 나타나 태그/표제 기반 스피치 요약과 본문 키워드 기반 스피치 요약은 서로 다르게 구성될 수 있다는 것을 예측할 수 있었다.

다음 단계로 두 방법에 의한 스피치 요약의 품질을 비교한 결과, 태그/표제 기반 요약문이 본문 키워드 기반 요약문과 비교하여 통계적으로 유의미한 차이를 보이지는 못했으나 재현율과 정확률이 각각 더 높은 것으로 나타났다. 이러한 분석 결과는 상당한 의미를 갖고 있다. 첫째, 태그는 비디오의 주제를 나타내는 키워드(색인어)로 사용될 수 있다는 점이다. 두 번째, 태그/표제 기반 요약은 단어의 출현 빈도와 위치 정보를 적용한 본문 키워드 기반 요약에 비해서 품질면에서 거의 같거나 조금 나은 것으로 나타났다. 이는 방대한 비디오 스피치 대본을 대상으로 한 세밀한 어휘적 분석 없이 이용자 태그/표제 키워드만으로 효율적인 스피치 요약을 할 수 있다는 것을 의미하는 것이다. 다만 태그/표제 기반 요약 방식을 적용하기 위해서는 태그 데이터가 있어야 하고 이외에 태그의 희박성 때문에 문제가 될 수 있다는 점이 있다. 태그 희박성 문제는 본 연구에서 사용한 방법과 같이 표제의 키워드를 함께 사용하거나 또는 태그를 시소러스와 같은 어휘 도구를 사용하여 동의어, 관련어로 확장하여 사용하는 방법을 통해서 완화시킬 수 있을 것이다.

스피치 내용 분석에서는 위치 정보의 중요성을 분석한 결과, 평균적으로 표준 요약의 30.5%가 스피치의 처음 부분(10%)에서 기술되고 있었고, 표준 요약의 9.5%가 스피치의 마지막 부분(10%)에서 기술되고 있었다. 이와 같이 교육 및 연설 비디오 자료의 앞부분에 출현한 문장들의 중요도를 확인할 수 있었다. 선택된 문장간의 유사도를 측정한 결과, 예상한대로 태그 기반 방법에 의해서 구성된 24개 요약 중 20개(83%)의 요약에서 문장간의 유사도가 0.80 이상인 문장을 한 쌍 이상 가지고 있는 것으로 나타났다. 이러한 분석 결과를 볼 때 중복된 문장을 제거하는 작업이 좀 더 다양한 요약문 구성을 위해서 필요한 것으로 보인다.

4. 비디오 스피치 요약을 위한 모형 설계와 평가

4.1 개요

본 장에서는 연구 문제 6(요약 모형 설계와 평가)을 위해서, 앞의 비디오 태그 및 스피치 분석 결과 즉, 태그가 주제어로서의 자질이 충분하고, 문장 위치 정보가 중요한 기준이 될 수 있으며, 한 요약문 내의 선택된 문장간의 중복도가 매우 높다는 것을 활용하기로 하였다. 이를 위해서 Goldstein et al.(2000)가 제안한 중복성을 최소화하면서 주요 문장을 선택하고 문장 위치 정보에 대한 가중치를 할당하는 항목을 추가한 수정된 MMR 모형을 활용하여 스피치 요약 기법을 제안하고 평가해 본다.

4.2 설계

다음은 수정된 MMR 모형과 이 모형을 기초로 하여 구성한 스피치 알고리즘에 대해서 상세히 기술한다.

4.2.1 수정된 MMR 모형

수정된 MMR 모형은 태그 및 표제 키워드 집합과 가장 관련 있는 문장이면서 동시에 앞부분에 출현한 문장에 더 높은 가중치를 부여하며, 이미 선정된 문장간의 유사도를 최소화하는 방식으로 주요 문장을 선택하도록 구성하였다(자세한 설명은 4.2.2 스피치 요약 알고리즘 설계 참조).

$$MMR = \underset{S_i \in R \setminus S}{Argmax}(\lambda\ Sim_1(S_i, K)\ w(S_i) - (1-\lambda) \max_{S_j \in S} Sim_2(S_i, S_j))$$

위에서 기술한 MMR 공식의 K는 태그 및 표제 키워드 집합, R는 문헌 집합에 있는 순위화된 문장 리스트, S는 요약을 위해서 이미 선정된 문장들의 리스트이다. 또한 R＼S는 S를 제외한 R 리스트 내에 있는 아직 선정되지 않은 문장들의 리스트, Sim_1는 문장 S_i와 K(태그 및 표제 키워드 집합)간의 유사도이다. 이외에 $w(S_i)$는 본 연구에서 새로 추가한 항목으로 문장 S_i에 대한 가중치로 만약 해당 문장이 전체 비디오의 처음 5%에 속하면 가중치 1.2를 할당하고 나머지 95%에 속한 문장에는 가중치 1을 할당한다. Sim_2는 문장 S_i와 문장 S_j 간의 유사도이며, λ는 0에서 1사이의 값을 갖는 파라미터로 적합성을 강조하거나 중복을 피하기 위해서 결합된 점수를 조정하기 위해서 사용한다. λ=1인 경우에는 문장의 순위화는 태그 및 표제 키워드 집합과의 적합성 및 문장 위치에만 기반하게 되며, 반대로 λ=0인 경우에는 문장간의 최대 다양성을 반영한 순위화가 이루어진다(정영미 2007). 본 연구는 테스트 결과, 0.7이 λ값으로 가장 적절하게 생각되어 이를 채택하였다.

4.2.2 스피치 요약 알고리즘 설계

스피치 요약 생성 절차는 다음과 같이 크게 세 단계로 구분하였다. 요약 생성 과정의 중요 부분들은 위해서 프로그램들을 작성하여 자동으로 수행하였다.

•**제 1단계**: 각 표본 비디오의 스피치는 스크립트를 이용하여 영문 텍스트로 변환하고, 변환된 텍스트는 마침표를 기준으로 하여 문장 단위로 구분한다. 그런 다음 각 문장에 순서대로 번호를 매긴다. 제 2단계에 들어가기 전에 세 단어 이하로 구성된 짧은 문장은 분석에서 제외시켰다.

•**제 2단계**: 각 표본 비디오의 스피치 텍스트의 개별적인 문장에 다음의 같은 네 가지 작업(표제 키워드 추출, 태그 필터링과 그룹화, 키워드 점수 계산 및 MMR 점수 계산)을 통해서 가중치를 할당하였다.

(1) 표제 키워드 추출: 각 비디오의 표제에서 키워드를 추출한다. 예를 들어서, 표본 비디오 3번의 표제(Kate Lundy: What I do for Open Government)에서 4개의 단일어를 추출하고, 각 표제 단일어에 가중치 1.5를 부여한다(〈표 4〉 참조).

〈표 4〉 표제 키워드(단일어) 리스트

번호	표제 키워드	가중치
1	Kate	1.5
2	Lundy	1.5
3	open	1.5
4	government	1.5

(2) 태그 필터링과 그룹화: 각 비디오에 할당된 태그들 중 스팸 태그는 삭제하고 단복수, 같은 의미의 다른 철자 사용 단어 등을 그룹화 시킨다. 예를 들어서, 태그 필터링 및 그룹화를 수행한 결과, 표본 비디오 3번에 16개의 태그(6개의 복합어와 10개의 단일어)가 추출되었는데 표제의 키워드와 중복되는 두 개의 복합어 태그 "Kate Lundy"와 "open government"를 제외한 14개의 태그 리스트를 최종적으로 만들었다(〈표 5〉 참조). 이때 비디오 대본에 출현하지 않은 태그들도 분석할 때 모두 사용한다. 단일어 태그와 복합어 태그에 가중치 1과 2를 각각 부여한다.

(3) 키워드 점수 계산: 키워드 점수(수정 MMR 공식의 Sim_1)는 태그/표제 키워드와 문장의 단어와의 매칭을 통해서 측정한다. 비디오 3의 경우, 표제 키워드 및 태그 리스트(〈표 4〉와 〈표 5〉)의 데이터를 이용하여 키워드 집단(K)과 문장 16(S_{16})간의 유사도를 다음과 같은 두 개의 용어 벡터를 이용하여 코싸인 유사계수로 구하면 0.49가 산출된다.

$K = (1.5,1.5,1.5,1.5,1.5,1.1,1.1,1.1,1.1,1.1,1.1,1.1,2,2,2,2)$

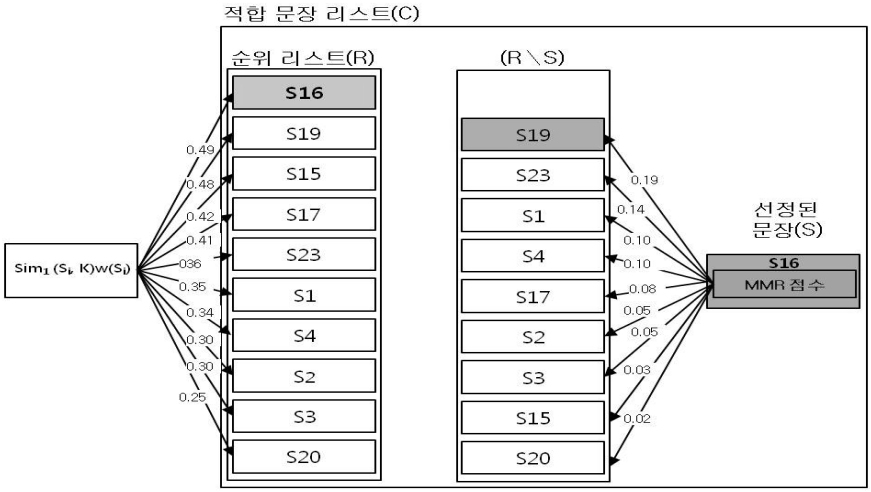
$S_{16} = (0,0,0,1.5,0,0,0,0,0,1,1,0,0,0,0,0,2,0)$

그 다음 단계로, 비디오 3의 총 26개의 문장에서 키워드 점수가 0 보다 큰 10개의 후보 문장들을 선정하였다(〈그림 1〉 참조). 키워드 점수를 문장의 길이에 따라서 정규화하지 않았는데 이는 코싸인 유사계수를 사용하여 문장 길이를 정규화한 효과가 어느 정도 있고 또한 긴 문장이 짧은 문장에 비해서 적합 문장이 될 가능성이 높다는 가정하에 길이도 하나의 중요한 기준이 될 수 있다고 생각했기 때문이다.

(4) MMR 점수 계산: MMR 점수를 얻기 위해서 문장간의 유사도(Sim_2)와 이전 단계에서 얻은 키워드 점수($Sim_1(S_i, K)$)를 문장 위치 가중치($w(S_i)$)로 결합한 "$Sim_1(S_i, K)w(S_i)$" 항목의 값이 필요하다. Sim_2를 계산하기 위해서 코싸인 유사 계수 공식을 사용한다. MMR 점수를 얻기 위해서, 〈그림 1〉에서처럼 가장 높은 "$Sim_1(S_{16}, K)w(S_{16})$"($0.49 \times 1 = 0.49$) 점수를 갖는 S_{16}을 제일 먼저 선정하였다. 〈그림 1〉의 순위리스트(R)에 있는 "$Sim_1(S_i, K)w(S_i)$"

〈표 5〉 태그 리스트

번호	태그	가중치	번호	태그	가중치
1	Australia	1	8	streaming	1
2	blog(ging)	1	9	transparency	1
3	Canberra	1	10	twitter	1
4	government 2.0	2	11	video	1
5	politics	1	12	video streaming	2
6	politician	1	13	Web 2.0	1
7	public sphere(s)	2	14	wiki	1



〈그림 1〉 비디오 3의 MMR 점수

값 중 "$w(S_i)$"값은 하나의 문장(S_1)(전체 26개 문장의 5%에 해당하는 1.3개를 1개로 봄)(예, $Sim_1(S_1, K)w(S_1)=0.29\times1.2=0.35$)에서 1.2이고 나머지에서는 모두 1이다. 다음 단계로 선정된 S_{16}과 나머지 선정되지 않은 9개의 문장간의 유사도(Sim_2)를 측정한다. 예를 들어서, S_{16}과 S_{19}간의 유사도 계수(0.49)는 다음의 두 개의 문장 벡터를 이용하여 측정하였다.

S_{16}=(0,0,0,1.5,0,0,0,0,0,1,1,0,0,0,0,0,2,0)

S_{19}=(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,2,0)

그런 다음, "$Sim_1(S_{19}, K)w(S_{19})$"(0.49 x 1=0.49)과 "$Sim_2(S_{16}, S_{19})$"(0.49)값을 MMR 공식에 대입하여 가장 높은 MMR 점수(0.19)를 얻은 S_{19}를 두 번째로 선정한다. 그 다음 단계는 선정된 S_{16}과 S_{19}의 센트로이드벡터(0,0,0,0.75,0,0,0,0,0,0.5,0.5,0,0,0,0,0,2,0)를 구한 다음 이 센트로이드벡터를 기준으로 하여 나머지 8개 문장간의 유사도를 측정하여 가장 높은 MMR 점수(0.18)를 얻은 S_{23}을 세 번째로 선정한다. 이러한 절차는 원하는 문장을 얻을 때까지 반복된다. 비디오 3의 경우는 6개의 문장이 선택될 때까지 이 과정을 반복하였다(〈표 6〉 참조).

•제 3단계: 가중치가 높은 순서대로 비디오의 재생시간에 따라서 5~7개의 문장을 선정한다. 최종적으로 선택된 문장들을 텍스트에서 출현 순서대로 요약문을 구성한 후 한글로 번역하고 이를 음성 합성기인 매직 잉글리쉬 플러스를 이용하여 오디오 파일로 변환하여 스피치 요약을 구성한다. 이때 구성된 스피치 요약의 재생시간은 30초~60초의 범주에 오도록 문장의 수를 재조정하였다. 예를 들어서 특정 비디오에 선정된 6개의 문장이 오디오로 변환한 후 만약 길이가 60초를 초과하는 경우 적합도가 가장 낮은 문장을 제외시켜 5개로 축약시키는 방법을 택할 수 있다. 재생시간을 60초 이내로 제한한 이유는 요약의 재생시간이 1분을 초과하면 이용자들이 집중하기가 어렵다고 판단했기 때문이다. 〈표 6〉는 최종적으로 선택된 비디오 3의 요약문이다.

〈표 6〉 비디오 3의 요약문

S_1. 호주에서의 정부는 연방정부, 주정부 그리고 지방정부라는 세 가지 형태가 있습니다.
S_4. 열린 정부란 국민이 정부로부터 알고자 하는 것을 얻을 수 있도록 하는 정부입니다.
S_{16}. 정부가 많은 사람들의 좋은 아이디어를 수용한다면 유익한 일이 될 것입니다. 이는 효율적이며 우리가 이러한 아이디어를 발표하는 것 외에 이러한 아이디어를 트위터 피드와 블로그를 통해 공론 영역에 제시하면 동료에 의해서 평가되고 활용될 수 있을 것입니다.
S_{17}. 우리의 공공 영역 이벤트는 주어진 주제의 범위에서 그들의 생각을 스스로 인식하는 사람들을 위해 계획되었습니다.
S_{19}. 우리가 계획하고 있는 다음 차례의 공공 영역은 정부 2.0에 관련된 것이 될 것입니다.
S_{23}. 그래서, 우리가 정부를 어떤 방식으로 개방할 것인지 그리고 디지털 혁명이 국민에게 의미 있도록 만드는 것이 제가 관심을 갖는 분야입니다.

4.3 평가

제안된 스피치 요약 알고리즘의 성능을 평가하기 위해서 연구문제 6을 다음과 같은 두 개의 연구 가설로 만들고 이를 검증하였다.

(1) 연구 가설 1: 제안된 알고리즘 기반 요약문의 재현율이 본문 키워드 기반 요약문의 재현율 보다 더 높을 것이다.

(2) 연구 가설 2: 제안된 알고리즘 기반 요약문의 정확률이 본문 키워드 기반 요약문의 정확률 보다 더 높을 것이다.

연구 가설을 검증하기 위해서 24개의 표본 비디오에서 9개를 무작위로 추출하여 제안된 알고리즘에 의해서 요약문을 구성하였다. 구성된 요약문을 표준 요약문과 비교하여 재현율과 정확률을 산출하였다. 제안된 알고리즘에 의해서 추출한 경우를 텍스트 키워드에 기반하여 추출한 경우와 비교한 결과(〈표 2〉 참조), 평균 재현율은 0.24에서 0.45로 증가하였으나 맨-위트니 검증 결과 통계적으로 유의미한 차이를 보이지는 못했다($p(=0.05) > = 0.05$). 평균 정확률은 0.18에서 0.36으로 증가하였고 맨-위트니 검증 결과 통계적으로 유의미한 차이를 보였다($p(=0.03) < 0.05$)(〈표 7〉 참조). 이와 같이 제안 방법은 본문 키워드만을 사용한 방법 보다

〈표 7〉 제안 알고리즘과 본문 키워드 방식의 비교

No.	제안 알고리즘		본문 키워드 기반	
	재현율	정확률	재현율	정확률
1	0.60	0.50	0.20	0.17
3	1.0	0.67	0.50	0.33
4	0.29	0.33	0.17	0.17
10	0.60	0.50	0.40	0.33
14	0.60	0.43	0.40	0.29
15	0.25	0.17	0.25	0.17
17	0.20	0.17	0.20	0.14
19	0.25	0.25	0.0	0.0
22	0.33	0.20	0.0	0.0
평균(표준편차)	0.45(0.26)	0.36(0.18)	0.24(0.17)	0.18(0.12)

정확률은 통계적으로 유의미한 차이를 보일 정도로 증가시켰으나, 재현율은 증가시키지 못한 것으로 나타났다. 이와 같이 연구 가설 2는 검증되었으나 연구 가설 1은 검증되지 못했다.

5. 결 론

본 연구에서는 스피치 요약에 텍스트 사용에 사용된 어떤 기준들이 효과적으로 사용될 수 있는지 조사해 보았다. 그런 다음, 이러한 분석 결과를 기초로 하여 효율적인 스피치 요약 방안을 구성하여 제안하였다.

첫째, 태그와 표제간의 일치율과 태그와 본문 키워드간의 일치율이 0.70, 0.22로 각각 나타났다. 태그/표제 기반 요약문과 본문 키워드 기반 요약문간에 요약 내용 품질을 비교한 결과, 태그/표제 기반 요약문의 재현율(0.35)과 정확률(0.23)이 본문 키워드 기반 요약문의 재현율(0.25)과 정확률(0.16) 보다 통계적으로 유의미한 차이를 보이지는 못했으나 더 높은 것으로 나타났다. 이러한 분석 결과는 이용자가 할당한 태그 및 표제 기법이 방대한 비디오 스피치 자막의 세밀한 처리 없이 효율적인 스피치 요약을 할 수 있다는 것을 확인시켜 주었다.

둘째, 문장 위치 정보의 중요성을 분석한 결과, 평균적으로 전체 비디오에서 표준 요약의 30.5%가 스피치의 처음(10%)에 기술되고 있었고, 표준 요약의 9.5%가 스피치의 마지막(10%)에 기술되고 있었다. 이와 같이 교육 및 연설 비디오 자료의 앞부분에 출현한 문장들의 중요도를 확인할 수 있었다. 이외에 선택된 문장간의 유사도를 측정한 결과, 태그/표제 기반 방법에 의해서 구성된 총 24개 요약 중 20개(83%)의 요약에서 문장간의 유사도가 0.80 이상인 문장을 한 쌍 이상 가지고 있는 것으로 나타났다.

앞의 분석 결과를 기초로 하여, 적합하면서 다양한 문장을 선택하고 문장 위치에 가중치를 부여할 수 있게 하는 수정된 MMR 모형을 활용한 요약 방법을 제안, 평가해 보았다. 비디오 스피치 본문은 재생시간에 따라서 일반 문헌 텍스트에 비해서 정보량이 훨씬 방대하다. 따라서 복잡한 텍스트 처리가 요구되는 방법은 실제 비디오 요약에 적용하기에 많은 어려움이 예상된다. 따라서 스피치 본문의 세밀한 분석 없이 이용자가 할당한 태그, 문장 위치와 같은 구조적 정보만을 활용한 제안된 방법은 실제 비디오 데이터에 쉽게 적용될 수 있는 이점이 있다고 할 수 있다. 마지막으로, 제안된 스피치 요약 방법을 평가하기 위해서 이 방법을 스피치 본문의 단어 출현 및 단어 위치 정보를 적용하여 상대적으로 복잡한 어휘 처리를 한 Extractor 시스템과 비교해 보았다. 비교 결과, 제안된 방법이 통계적으로 유의미한 차이를 보이며 평균 정확률(0.18→0.36)은 더 높았고, 평균 재현율(0.24→0.45)은 더 높았지만 통계적으로 유의미한 차이를 보이지는 못했다. 앞으로 방대한 비디오 데이터를 활용한 시뮬레이션을 통해서 MMR 모형의 좀 더 적절한 λ값과 문장 위치 가중치($w(S_i)$)값을 구하는 작업이 필요해 보인다.

본 연구에서 제안한 스피치 요약 알고리즘은 비디오 대본의 텍스트 요약은 물론 스피치 요약의 구성에 적용되어 스마트폰이나 PDA와 같은 소규모 디스플레이 장치가 주류를 이루는 미래의 정보 환경에서 특히 유용하게 이용될 수 있을 것이다. 이외에 제안된 알고리즘은 비디오의

색인어 추출 및 주제 분류에도 활용될 수 있을 것으로 생각된다. 궁극적으로 제안된 스피치 요약 방법이 본 연구에서 다루지 않은 음향학/운율적인 정보와 결합되어 좀 더 나은 스피치 요약을 위한 이론적인 틀을 구성하기 위한 기초 자료로 활용될 수 있기를 바란다.

참 고 문 헌

[1] 김현희. 2009. 비디오의 오디오 정보 요약 기법에 관한 연구.『정보관리학회지』, 26(3): 169-188.

[2] 김현희. 2011. 비디오 의미 파악을 위한 멀티미디어 요약의 비동시적 오디오와 이미지 정보간의 상호 작용 효과 연구.『한국문헌정보학회지』, 45(2): 97-118.

[3] 정영미. 2007.『정보검색연구』. 서울: 구미무역출판부.

[4] 이한성 외. 2010. 멀티모달 방법론과 텍스트 마이닝 기반의 뉴스 비디오 마이닝.『정보과학회논문지: 데이타베이스』, 37(3): 127-136.

[5] Boydell, O., & Smyth, B. 2010. "Social summarization in collaborative web search." *Information Processing and Management*, 46(6): 782-798.

[6] Chen, B., & Lin, S. 2012. "A risk-aware modeling framework for speech summarization." *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1): 211-222.

[7] Christensen, H., et al. 2003. "Are extractive text summarisation techniques portable to broadcast news?" In *Proceedings of Automatic Speech Recognition and Understanding Workshop*, 489-494. St. Thomas, USA.

[8] Chung, M., Wang, T. & Sheu, P. 2011. "Video summarisation based on collaborative temporal tags." *Online Information Review*, 35(4): 653-668.

[9] Goldstein, J., et al. 2000. "Multi-document summarization by sentence extraction." In *Proceedings of the 2000 NAACL-ANLP Workshop on Automatic Summarization(NAACL-ANLP-AutoSum '00)*, Vol.4: 40-48. Stroudsburg, PA, USA: Association for Computational Linguistics.

[10] Hannon, J., et al. 2011. "Personalized and automatic social summarization of events in video." In *Proceedings of the 16th International Conference on Intelligent User Interfaces*, 335-338. Palo Alto, California, USA.

[11] Heckner, M., Neubauer, T., & Wolff, C. 2008. "Tree, funny, to read, google: What are tags supposed to achieve?" In *Proceedings of the 2008 ACM Workshop on Search in Social Media*, 3-10. Napa Valley, California, USA.

[12] Hirohata, M., et al. 2006. "Sentence-extractive automatic speech summarization and evaluation

techniques." *Speech Communication*, 48(9): 1151-1161.

[13] Kim, H. 2011. "Toward video semantic search based on a structured folksonomy." *Journal of the American Society for Information Science*, 62(3): 478-492.

[14] Liu, Y., & Hakkani-Tur, D. 2011. "Speech summarization." In *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*. G. Edited by Hakkani-Tur and R. Mori. Hoboken, NJ: Wiley, 357-392.

[15] Maskey, S., & Hirschberg, J. 2006. "Summarizing speech without text using Hidden Markov Models." In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers(NAACL-Short'06)*, 89-92. Stroudsburg, PA, USA: Association for Computational Linguistics.

[16] Marchionini, G., et al. 2009. "Multimedia surrogates for video gisting: Toward combining spoken words and imagery." *Information Processing and Management*, 45(6): 615-630.

[17] Murray, G., Renals, S., & Carletta, J. 2005. "Extractive summarization of meeting recordings." *Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH)*, 593-596. Lisbon, Portugal.

[18] Song, Y., & Marchionini, G. 2007. "Effects of audio and visual surrogates for making sense of digital video." In *Proceedings of CHI 2007*, 867-876. San Jose, CA, USA.

[19] Turney, P. 2000. "Learning algorithms for keyphrase extraction." *Information Retrieval*, 2(4): 303-336.

[20] Xie, S., & Liu, Y. 2008. "Using corpus and knowledge-based similarity measure in maximum marginal relevance for meeting summarization." *IEEE International Conference on Acoustics, Speech and Signal Processing*, 4985-4988.

[21] Xie, S., et al. 2009. "Integrating prosodic features in extractive meeting summarization." *Proceedings of Automatic Speech Recognition & Understanding, 2009*.

[22] Zechner, K. 2002. "Automatic summarization of open-domain multiparty dialogues in diverse genres." *Computational Linguistics*, 28(4): 447-485.

[23] Zhang, J., et al. 2007. "A comparative study on speech summarization of broadcast news and lecture speech." In *INTERSPEECH-2007*, 2781-2784.

[24] Zhu, J., et al. 2009. "Tag-oriented document summarization." *Proceedings of the 18th International Conference on World Wide Web*, 1195-1196.

[25] Zhu, X., Penn, G., & Rudzicz, F. 2009. "Summarizing multiple spoken documents: Finding evidence from untranscribed audio." *Proceedings of ACL/AFNLP*, 549-557.

• **국문 참고자료의 영어 표기**
(English translation / romanization of references originally written in Korean)

[1] Kim, Hyun-Hee. 2009. "Investigating the efficient method for constructing audio surrogates of digital video data." *Journal of the Korean Society for Information Management*, 26(3): 169-188.

[2] Kim, Hyun-Hee. 2011. "A study on the interactive effect of spoken words and imagery not synchronized in multimedia surrogates for video gisting." *Journal of the Korean Society for Library and Information Science*, 45(2): 97-118.

[3] Chung, Young Mee. 2007. *Information Retrieval Research*. Seoul: Gumi Trading Publisher.

[4] Lee, Hansung, et al. 2010. "A news video mining based on multi-modal approach and text mining." *Journal of KISS: Databases*, 37(3): 127-136.