

스피치 요약을 위한 태그의미분석과 잠재의미분석간의 비교 연구*

Comparing the Use of Semantic Relations between Tags Versus Latent Semantic Analysis for Speech Summarization

김 현 희 (Hyun-Hee Kim)**

목 차

- | | |
|-----------------------------------|--------------------------------------|
| 1. 서론 | 3.2 요약 알고리즘 |
| 1.1 연구 배경과 목적 | 4. 태그의미분석 방법의 분석과 평가 |
| 1.2 연구 문제와 방법 | 4.1 플리커 태그 클러스터 및 워드넷
동어의 분석 |
| 2. 선행 연구 | 4.2 태그의미분석 방법과 잠재의미분석
방법 간의 비교 분석 |
| 3. 비디오 스피치 요약을 위한 모형 및
알고리즘 설계 | 5. 결론 |
| 3.1 요약 모형 | |

초 록

본 연구는 스피치 요약을 위해서 태그를 확장하고 또한 태그 간의 의미적 관계 정보를 이용할 수 있는 태그의미분석 방법을 제안하고 평가하였다. 이를 위해서, 먼저 비디오 태그를 확장하고 태그 간의 의미적 관계를 분석하는데 있어서 플리커의 태그 클러스터와 워드넷의 동의어 정보가 얼마나 효과적으로 이용될 수 있는가 조사해 보았다. 그런 다음 태그의미분석 방법의 특성과 효율성을 조사해 보기 위해서 제안한 방법을 잠재의미분석(Latent Semantic Analysis) 방법과 비교해 보았다. 분석 결과, 플리커의 태그 클러스터는 효과적으로 이용되었지만 워드넷은 효과적으로 이용되지 못한 것으로 나타났다. F측정을 사용하여 두 방법의 효율성을 비교한 결과, 제안한 방법의 F값(0.27)이 잠재의미분석 방법의 F값(0.22)보다 높게 나타났다.

ABSTRACT

We proposed and evaluated a tag semantic analysis method in which original tags are expanded and the semantic relations between original or expanded tags are used to extract key sentences from lecture speech transcripts. To do that, we first investigated how useful Flickr tag clusters and WordNet synonyms are for expanding tags and for detecting the semantic relations between tags. Then, to evaluate our proposed method, we compared it with a latent semantic analysis (LSA) method. As a result, we found that Flickr tag clusters are more effective than WordNet synonyms and that the F measure mean (0.27) of the tag semantic analysis method is higher than that of LSA method (0.22).

키워드: 일반 스피치 요약, 비디오, 태그의미분석, 확장된 태그, 태그 클러스터, 잠재의미분석, F측정, 강의 자료, 플리커, 유튜브, 내재적 평가

Expanded Tags, Latent Semantic Analysis, TED Talks, Flickr Tag Clusters, WordNet

* 본 연구는 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 연구되었음 (NRF-2011-342-B00025).

** 명지대학교 문헌정보학과 교수(kimhh@mju.ac.kr)

논문접수일자: 2013년 7월 16일 최초심사일자: 2013년 8월 5일 게재확정일자: 2013년 8월 5일
한국문헌정보학회지, 47(3): 343-361, 2013. [http://dx.doi.org/10.4275/KSLIS.2013.47.3.343]

1. 서론

1.1 연구 배경과 목적

정보 검색의 미디어가 데스크탑 컴퓨터에서 태블릿 PC 또는 스마트폰으로 빠르게 이동하고 있다. 생산되는 콘텐츠는 기하급수적으로 증가하고 있는데 반해서, 콘텐츠를 표현하는 가상 공간은 좁아지고 있다. 이러한 정보 검색 환경에 맞추어 최근 야후는 뉴스 기사를 400자 내외로 요약해주는 '섬리(Summlly)' 기술을 도입한 새로운 모바일 앱을 출시했다(<http://www.bbc.co.uk/news/technology-21924243>). 뉴스 정보뿐만 아니라 개인, 대학, 기관 등은 방대한 스피치 형태의 교육 자료를 생산하여 인터넷 웹을 통해서 제공하고 있지만, 이용자들은 적절한 요약 정보의 부족으로 이들 자료를 효율적으로 활용하지 못하는 경우가 많아지고 있다. 따라서, 이러한 방대한 정보원에 대한 효율적인 접근을 가능하게 하는 하나의 방법으로 이용자들이 원하는 스피치 비디오를 선택하도록 돕는 요약문을 포함한 메타데이터의 제공이 매우 필요한 시점이 되었다.

최근 비디오나 텍스트 문헌의 요약을 위해서 태그가 사용되고 있다. Heckner, Neubauer, & Wolff(2008)는 비디오 태그가 색인어로서 충분한 가치가 있다는 주장하였고, 김현희(2009; 2012)와 Zhu et al.(2009)은 태그를 이용한 스피치와 웹문헌의 요약 방법을 제안하였다. 그러나 자료의 요약에 태그를 사용하는 태그 방법은 태그의 희박성 때문에 요약을 하기 위한 충분한 태그의 확보가 보장되지 않는 경우가 종종 발생하게 된다. Zhu et al.(2009)은 이러한 태그 희

박성 문제를 해결하기 위해서 연관 마이닝 기술을 이용하여 태그를 확장하고 태그에서의 잡음을 줄이는 태그 랭킹 알고리즘을 이용한 태그 기반 웹 문서 요약 방법을 제안하였다. 김현희(2009)는 태그의 희박성 문제를 해결하기 위해서 표제의 키워드를 태그와 함께 사용하는 방안을 제안하였다. 또한 태그들에 의한 요약은 동일한 점수를 갖는 세그먼트나 문장이 많아서 가중치순으로 문장을 선정하는 작업이 어려워 지기도 한다. 이외에 태그 방법은 태그들이 태그간의 관계를 드러내지 못하는 평면적인 구조로 보여주기 때문에 동의어는 함께 모아지지 않고, 동음이의어는 구별되지 않는다(Matusiak 2006).

전문가들이 문헌을 요약할 경우 빈도수가 높은 단어를 기준으로 요약하기보다는 각 단어와의 관계를 분석하여 기여도가 높고 의미상 연관성이 높은 단어들 위주로 요약을 수행하고 있다고 보고하고 있다(Heu et al. 2013). 많은 연구들(Gong & Liu 2001; Steinberger & Jezek 2004; Hennig 2009; Ozsoy, Alpaslan, & Cicekli 2011)은 이러한 전문가의 요약 패턴에 근접하기 위한 하나의 방법으로 잠재의미분석(Latent Semantic Analysis, LSA)을 이용하여 텍스트를 요약하는 방안들을 제안하고 있다. 잠재의미분석은 일반화된 법칙에 기초한 비지도(unsupervised)(규칙 기반) 방법 중 하나로 문헌과 단어들 간의 표면적인 관계가 아니라 문헌 내에 존재하는 단어들에 내포된 숨겨진 주제들 별로 클러스터들을 만들고 클러스터된 단어와 문장 간의 관계를 분석하여 문장에 점수를 부여하는 방법이다(정영미 2005). 이러한 방법은 비교적 우수한 결과값을 얻을 수 있지만

대규모의 단어-문헌 행렬을 입력 데이터로 하여 요인분석 기법의 일종인 특이값 분해(SVD: Singular Value Decomposition)와 같은 복잡한 계산 과정을 거쳐야 하는 단점이 있다.

본 연구는 스피치 요약을 위해서 태그를 확장하여 태그의 희박성 문제를 해결하고 의미상 연관성이 높은 단어들을 포함한 문장에 더 높은 가중치를 부여하기 위한 태그의미분석 방법을 제안한다. 태그의미분석 방법은 복잡한 계산 과정 없이 태그 간의 의미적 관계 정보를 이용할 수 있는 알고리즘을 활용한다. 제안한 방법은 태그간의 의미적 관계 즉, 동등관계와 연관관계를 파악하기 위해서 워드넷(WordNet) 2.1의 동의어(텍소노미)와 플리커의 관련어 태그(폭소노미) 정보를 이용한다. 이 두 가지 정보원을 이용하는 이유는 플리커의 관련어 태그들은 태그에서 많은 비중을 차지하는 개체명(사람 이름, 장소명, 조직명, 날짜 등)과 같은 고유 명사 또는 신조어를 많이 포함하고 있는데 워드넷과 같은 전통적인 텍소노미는 이러한 고유 명사나 신조어를 많이 포함하고 있지 않기 때문이다. 또한 플리커의 관련어 태그들은 학술 정보 콘텐츠에서 사용한 용어들을 많이 포함하지 못한다는 단점이 있기 때문이다. 실제 플리커의 태그 클러스터는 Kim(2011)의 연구와 Heu et al.(2013)의 연구에서 성공적으로 활용되었다. Kim은 플리커의 태그 클러스터를 유튜브 비디오 탐색시질의 확장을 하기 위해서 사용하였고, Heu et al.는 복수 텍스트 문헌들을 요약할 때, 핵심 키워드들과 키워드 간의 의미적 관련성을 파악하기 위해서 사용하였다.

본 연구의 목적은 스피치 요약을 위해서 태그의미분석 방법을 제안하고 제안한 방법을 잠재

의미분석 방법과 비교해 본 후 태그의미분석 방법의 특성과 효율성을 조사해 봄으로써 효율적인 스피치 요약 방안을 제안하는 데 있다.

1.2 연구 문제와 방법

효율적인 스피치 요약을 구현하는 방안을 모색하기 위해서 다음과 같은 두 가지 연구 문제들을 조사한다.

- **연구 문제 1:** 플리커의 태그 클러스터와 워드넷의 동의어 정보가 태그를 확장하고 비디오 태그 간의 의미적 관계를 분석하는데 얼마나 효과적으로 이용될 수 있는가?
- **연구 문제 2:** 스피치 요약에 태그의미분석 방법과 잠재의미분석 방법을 적용할 때 이 두 방법에 의한 요약문이 품질면에서 어떤 차이를 보일 것인가?

이러한 연구 문제들을 조사하기 위한 표본 비디오 자료는 음성으로 많은 정보를 표현하는 강의, 교육 및 연설 비디오로 정하고 TEDTalks 사이트와 유튜브 사이트에서 총 60개의 영어로 된 비디오들을 선정하였다. 스피치 요약의 효율성을 평가하기 위해서 요약 기법의 성능을 평가하는 내재적 평가를 하였다(정영미 2005). 내재적 평가를 위해서 연구팀은 스피치 대본에서 비디오의 의미를 가장 잘 나타내는 문장들을 추출하여 표준 요약을 구성하였고(자세한 설명은 “4.2 태그의미분석 방법과 잠재의미분석 방법 간의 비교 분석” 참조), 통계 분석을 위해서 SPSS 통계 패키지를 사용하였다.

2. 선행 연구

Yamamoto, Masuda, Ohira, 및 Nagao(2008)는 태그 클라우드에 기초한 비디오 장면 요약 방법을 제안하였다. 이들이 제안한 신비(Synvie) 시스템은 비디오를 시청하는 동안 이용자로 하여금 이용자 댓글로부터 생성되는 태그 클라우드에 포함된 태그를 선택하여 클릭된 태그가 비디오를 보는 시점과 연관되어진다. 이들은 신비 시스템에 의해서 생성된 요약의 적용 범위가 기존의 방법보다 더 크게 나타났고, 태그 공유와 태그 클라우드 방법은 이용자들이 태그를 추가하도록 동기 부여를 한다고 주장하였다.

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 보고하고 있다.

김현희(2009)는 텍스트 요약에 적용된 이론과 방법이 오디오 요약에도 적용될 수 있을 것이라는 가정하에 오디오 요약을 위해서 태그와 표제의 키워드를 함께 사용하는 태그 확장 방법을 제안하여 태그의 희박성 문제를 해결하고자 하였다. 이 연구는 좀 더 효율적인 오디오 요약을 구현하기 위해서는 오디오 정보의 특성에 맞춘 요약 기법에 대한 연구가 요망된다고 기술하고 있다. 또한 김현희(2012)는 스피치 요약의 알고리즘을 구성하기 위해서 이용자 태그 기법, 문장 위치 및 문장 중복도 제거 기법의 효율성을 분석해 보았다. 그런 다음, 분석 결과를 기초

로 하여 스피치 요약 방법을 구성, 평가하여 효율적인 스피치 요약 방안을 제안하였다.

Boydell과 Smyth(2010)는 문헌 요약을 위해서 이용자 피드백과 탐색 엔진의 결합이 효과적이면서 효율적인 문헌 요약을 산출할 것이라는 가정하에 소셜 북마크 태그와 탐색 질의 정보를 사용하는 방안을 제안하였다. 즉, 이용자들이 하나의 페이지를 북마크할 때 사용하는 태그(들)을 각 질의어로 하여 탐색 엔진으로 페이지를 검색한 다음 검색된 페이지의 요약(들)의 각 세그먼트의 가중치를 측정된 후 등급화하여 비디오 요약을 구성하였다. 이들은 제안한 방안을 표층적 자연언어처리기법과 통계적 단어 빈도 방법을 결합한 방식을 채택한 두 개의 텍스트 요약 시스템 즉, 오픈 텍스트 요약기(Open Text Summarizer)와 MEAD 시스템과 비교하였다. 비교 결과, 제안한 방안이 두 개의 시스템들보다 더 높은 품질의 요약문을 생성하는 것으로 나타났다.

Wang et al.(2012)은 태그 지역화(tag localization)와 키샷(키프레임) 마이닝(key-shot mining)에 의해서 이벤트 기반 웹비디오의 요약 방안을 제안하였다. 이들은 다중 인스턴스 학습 방법(multiple instance learning approach: 인스턴스의 집합인 Bag을 학습 단위로 하며 이진 분류를 수행함에 있어 Bag이 긍정이면 적어도 Bag 안에 존재하는 인스턴스 중 하나 이상이 긍정임을, Bag이 부정이면 Bag 안에 모든 인스턴스들이 부정임을 가정함)을 이용하여 각 비디오에 할당된 이용자 태그를 해당되는 샷에 배정하는 태그 지역화 작업을 한다. 그런 다음 샷 단위의 태그들을 이벤트 질의와 매칭시켜 적합성 점수를 계산하였다. 검색된 비디오 집단에

서 자주 출현하는 유사한 샷들은 키샷일 가능성이 높을 것이라는 가정하에 유사 중복 키프레임 검출 알고리즘(near-duplicate keyframe detection algorithm)을 이용하여 키샷들을 확인하였다. 이렇게 추출된 키샷들은 비디오 스킴이나 비주얼-텍스트 스토리보드를 구성하기 위해서 이용하였다. 제안한 비디오 요약 방법을 평가하기 위해서 만여 개의 표본 웹비디오, 각 웹비디오에 할당된 태그들, 그리고 60개의 표본 이벤트 질의를 이용하여 비디오 요약을 구성하였다. 이들은 구성된 비디오 요약의 질을 평가하기 위해서 11명의 이용자들에게 본 연구에서 제안한 방식에 의한 비디오 요약과 Hong et al.(2011)이 제안한 이벤트 기반 웹비디오 요약 방식에 의한 비디오 요약을 서로 비교하게 하였다. 비교 결과, 제안한 방식이 더 우수한 것으로 확인되었다.

Heu et al.(2013)은 전문가들이 문헌을 요약할 경우 빈도수가 높은 단어를 기준으로 하기 보다는 기여도가 높고 의미적으로 연관성이 높은 단어들 위주로 요약을 수행하고 있다고 보고 다음과 같은 알고리즘을 이용하여 문헌을 요약하였다. 첫째, 복수개의 텍스트 문헌의 내용을 요약하기 위해서 문헌의 본문에 출현한 단어와 각 단어의 빈도수를 추출하였다. 그런 다음, 플리커에서 제공된 태그 클러스터들을 이용하여 문헌 내에 존재하는 단어의 빈도수를 갱신하여 문헌내의 중요 단어들을 등급화하였다. 이를 기반으로 문헌 내에 존재하는 단어들의 문헌 내 기여도와 단어들 간의 의미적인 연관성을 분석하여 문헌 내의 핵심이 되는 단어들 위주로 문헌을 요약하는 방안을 제안하고, 제안한 방법의 효율성을 입증하였다.

선행 연구들을 살펴본 결과, 태그 방법이 스피치 요약에 이용될 때 태그 확장은 시도되었지만 태그 간의 의미적 정보를 이용하여 세그먼트(문장)의 가중치를 측정하는 알고리즘을 적용한 연구는 없는 것으로 나타났다. 따라서, 본 연구는 스피치 요약을 위해서 플리커의 태그 클러스터와 워드넷의 동의어 정보를 이용하여 태그를 확장하여 태그의 희박성 문제를 해결하고, 더 나아가 전문가의 요약 패턴에 근접하기 위해서 의미적으로 연관성이 높은 단어들을 포함한 세그먼트에 더 높은 가중치를 부여하게 하는 알고리즘(태그의미분석 방법)을 제안하고 그 효율성을 평가해 보고자 한다.

3. 비디오 스피치 요약을 위한 모형 및 알고리즘 설계

스피치 요약 알고리즘을 구성하기 위해서 요약 모형을 설계한다. 제안한 요약 모형은 태그 집합과 가장 관련 있는 문장이면서 의미적으로 서로 관련 있는 태그들을 많이 포함한 문장에 더 높은 가중치를 부여하여 주요 문장을 선택하도록 하는 스피치 요약 알고리즘에 이용된다.

3.1 요약 모형

요약 모형(공식 1)은 오리지널 및 확장 태그 집합과 가장 관련 있는 문장이면서 동시에 의미적으로 서로 관련 있는 태그들을 많이 포함한 문장에 더 높은 가중치를 부여하여 주요 문장을 선택하도록 다음과 같이 구성한다.

$$W_{tot}(S_i) = (Sim(S_i, T_{ex}) + W_{sr}(S_i)) / 2$$

..... (1)

여기서, $W_{tot}(S_i)$ 는 문장 S_i 의 최종 가중치, T_{ex} 는 태그 집합, Sim 는 문장 S_i 와 T_{ex} 간의 유사도이며, 코사인 유사계수를 사용하여 측정한다. $W_{sr}(S_i)$ 는 문장 S_i 에 포함된 태그들 간의 의미적 관계를 수치로 표현한 값이다.

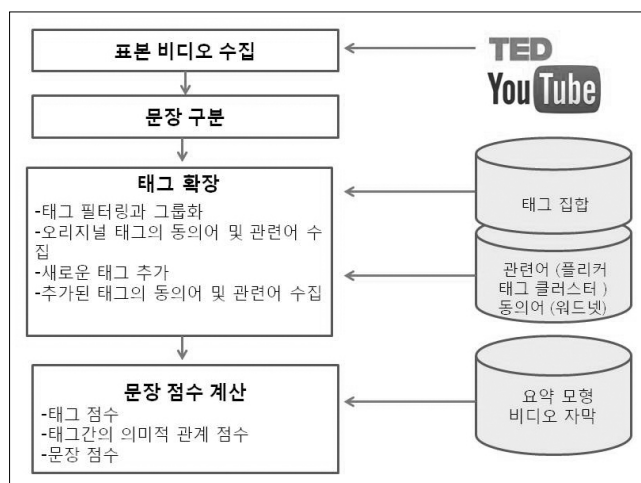
3.2 요약 알고리즘

스피치 요약 생성 절차는 다음과 같이 크네 단계로 구분한다(〈그림 1〉 참조). 요약 생성 과정의 중요 부분들은 프로그램들을 작성하여 자동으로 수행하도록 하였다.

제1단계-표본 비디오 수집: 비디오 태그 및 스피치 내용 분석을 위해서 60개의 영어로 된 표본 비디오를 선정하였다. 표본 비디오로 영어 비디오를 선정한 이유는 소셜 태그를 풍부하게 포함하고 있기 때문이며, 또한 본 연구에서 제

안한 스피치 요약 알고리즘은 언어에 관계없이 모두 적용될 수 있기 때문이다. 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. 60개의 표본 비디오를 분석한 결과, 오리지널 태그 평균수가 14.4개로 나타났다. 구체적으로, 단일어 태그의 평균수는 12.6개, 복합어 태그의 평균수는 1.8개이다. 태그가 해당 비디오 자막에 출현한 경우만을 계산했을 때 평균수가 14.4개에서 5.1개로 줄어들었고, 확장된 태그를 포함했을 때 최종적으로 평균 태그수가 5.6개가 되었다. 6개의 비디오들(10.0%)은 단지 하나의 태그만이 해당 비디오 자막에 출현하였고 한 비디오(1.7%)는 태그가 전혀 출현하지 않았다.

제2단계-문장 구분: 각 표본 비디오의 스피치는 스크립트를 이용하여 영문 텍스트로 변환하고, 변환된 텍스트는 마침표 등을 기준으로 하여 문장 단위로 구분한다. 그런 다음 각 문장에



〈그림 1〉 스피치 요약 생성 절차

순서대로 번호를 매긴다. 제3단계에 들어가기 전에 세 단어 이하로 구성된 짧은 문장은 분석에서 제외시켰다. 예를 들어서, 비디오 11(“Steve Jobs’ 2005 Stanford commencement address”)은 142개의 문장을 포함하고 있다.

제3단계-태그 확장: 다음과 같이 네 가지 절차를 통해서 태그를 확장한다.

(1) 태그 필터링과 그룹화: 각 비디오에 할당된 태그들 중 스팸 태그는 삭제하고 단복수, 같은 의미의 다른 철자 사용 단어 등을 그룹화시킨다. 예를 들어서, 태그 필터링 및 그룹화를 수행한 결과, 표본 비디오 11번에 할당된 14개의 오리지널 태그에서 비디오 자막에 출현하지 않은 6개의 태그를 제외한 8개의 태그(T₁-T₈: [Stanford, commencement, Apple, graduation, NeXt, Pixar, cancer, & computer])로 구성된 초기 태그 리스트를 만들었다(〈표 1〉 참조).

(2) 오리지널 태그의 관련어 및 동의어 수집:

각 오리지널 태그의 관련어 및 동의어로 구성된 용어 클러스터를 수집하기 위해서 두 가지 용어 데이터베이스를 이용하였다.

첫째, 태그의 관련어를 수집하기 위해서, 플리커의 태그 클러스터를 이용하였다(<http://www.flickr.com/services/api/explore/flickr.tags.getClusters>). 태그 클러스터는 태그의 동시출현(co-occurrence) 빈도에 기초하여 생성되며(Specia & Motta 2007), 태그의 동의어의어를 구분하는데 유용하게 사용된다. 예를 들어서, ‘java’라는 태그는 여러 가지 의미 즉, 인도네시아섬, 커피 또는 프로그래밍 언어를 의미하는데 플리커는 두 개의 클러스터 즉, 인도네시아섬을 의미하는 태그 클러스터(indonesia, volcano, Borobudur)와 커피를 의미하는 태그 클러스터(coffe, cup, joe)를 갖고 있다.

비디오 11의 태그 ‘commencement’는 두 개의 플리커 태그 클러스터(태그 클러스터 1:

〈표 1〉 비디오 11의 태그 가중치 테이블

번호	태그 (=T _i)	가중치 (W _i)	관련어 및 동의어 태그 클러스터 (=TC _i)	
			플리커 관련어	위드넷 동의어
1	Stanford	1	-	-
2	commencement	1	graduation, university, college	commencement exercise*, commencement ceremony*, graduation, graduation exercise*
3	Apple	1	-	-
4	graduation	1	university, college, commencement	-
5	NeXT	1	-	-
6	Pixar	1	-	-
7	cancer	1	-	-
8	computer	1	Apple, mac*, laptop*	-
9	university	1	college, campus, architecture*	-
10	college	1	university, campus, school	-

* 밑줄 친 태그는 초기 태그 리스트에 있는 용어들이며, 별표를 윗첨자로 사용한 태그들은 비디오 11의 자막에 출현하지 않은 용어들이다. 볼드체의 단어들은 새로운 태그로 추가된 것들이다.

graduation, university, college: 태그 클러스터 2: Tacoma, bay, Washington)를 갖고 있다. 본 연구는 의미적으로 관련 있는 적합한 태그 클러스터(들)를 선택하는 조건으로 태그 클러스터에 속한 용어(태그)가 초기 태그 리스트에 있으면 관련된 태그 클러스터로 판단하고 해당 태그 클러스터를 선정한다. 비디오 11의 경우 두 개의 클러스터 중에서 태그 클러스터 1이 초기 태그 리스트에 있는 'graduation'을 갖고 있기 때문에 선택되었다.

둘째, 태그의 동의어를 수집하기 위해서 워드넷을 사용했다. 워드넷은 프린스턴 대학에 의해 개발된 영어를 위한 가장 큰 어휘 데이터베이스이다. 워드넷은 명사, 동사, 형용사, 부사를 인지(cognitive) 동의어로 분류하고, 인지(cognitive) 동의어 간에 개념적 의미 관계와 어휘적 관계를 제공한다(fellbaum 1998). 태그 'commencement'은 세 개의 용어 클러스터(용어 클러스터 1: beginning, first, outset...: 용어 클러스터 2: commencement exercise, commencement ceremony, graduation, graduation exercise: 용어 클러스터 3: beginning, start)를 갖고 있다. 앞의 플리커 태그 클러스터와 같이 클러스터에 속한 용어가 초기 태그 리스트에 있다면 해당 용어 클러스터를 선정하였다. 이에 따라서, 용어 클러스터 2도 선정되었다. <표 1>에서 확인할 수 있는 것처럼 플리커 태그 클러스터 1과 워드넷 용어 클러스터 2가 관련어 및 동의어 태그 클러스터 2(TC₂)에 저장된다. 나머지 7개의 오리지널 태그에 대해서 앞에서 기술한 절차대로 분석하여 태그 및 용어 클러스터를 수집하였다.

(3) 새로운 태그 추가: 세 개의 관련어 및 동

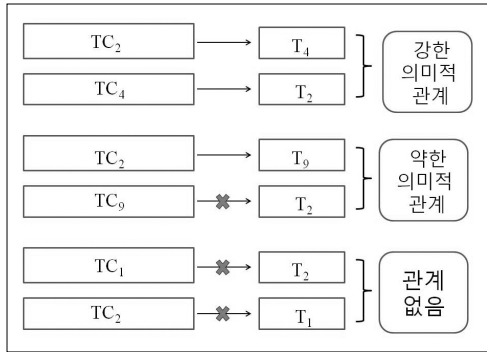
의어 태그 클러스터(TC₂, TC₄, TC₈)에서 수집한 10개 용어들 중에서 두 개의 단어(university & college)가 새로운 태그로 선정되었다. 5개의 단어(commencement exercise, commencement ceremony, graduation exercise, mac, & laptop)가 선정되지 않는 것은 이들 단어들이 비디오 자막에 출현하지 않았기 때문이며, 또 다른 3개의 단어(graduation, commencement, & Apple)는 초기 태그 리스트에 이미 있기 때문에 선정되지 않았다.

(4) 추가된 태그의 관련어 및 동의어 수집: 두 개의 추가된 태그의 플리커 태그 클러스터와 워드넷 용어 클러스터를 수집하였다. 수집한 결과, 비디오 자막에 출현하는 새로운 태그로 'campus'와 'school'이 확인되었으나 이들을 두 번째 태그 클러스터 수집 단계에서는 초기 태그 리스트에 추가하지 않았다. 최종적으로 10개의 태그로 구성된 최종 태그 리스트를 얻었다. 그리고 단일어 태그에는 가중치 '1'을 복합어 태그에는 가중치 '2'를 부여한다. 다음 공식(2)은 태그 가중치 테이블(Tag Weight Table, TWT)을 기술한 것이다.

$$TWT = \{(t_1, w_1), (t_2, w_2), \dots, (t_{n-1}, w_{n-1}), (t_n, w_n)\} \dots\dots\dots (2)$$

여기서 t_i은 한 문헌에서 i번째 태그를 나타내고, w_i은 i번째 태그의 가중치를 나타낸다. <표 1>은 비디오 11의 태그 가중치 테이블을 기술한 것이다.

제4단계-문장 점수 계산: 각 문장의 점수는 세 가지 절차를 통해서 측정하였다. 비디오 11의 스크립트에 속한 각 문장의 점수는 태그 점수와



<그림 2> 태그 간의 의미적 관계

$$W_{sr}(S_i) = \sum_{i, j \in \text{trt}(s), \text{trt}(s) \subseteq \text{TRT}(D)} R_{i,j} \dots\dots\dots (3)$$

여기서, 만약 $T_j \in TC_i$ 이면 $R_{i,j}$ 의 값이 0.1 이 되고 $T_j \notin TC_i$ 이면 $R_{i,j}$ 의 값이 0이 된다. $\text{TRT}(D_i)$ 은 문헌(비디오) D_i 의 태그 관계 테이블이고, $\text{trt}(s_i)$ 는 문헌(비디오) D_i 에 속한 문장 (S_i)의 태그 관계 테이블이다. 예를 들어서, 비디오 11의 문장 1(S_1)에 있는 태그 간의 의미적 관계 점수($W_{sr}(S_1)$)을 계산하기 위해서 $\text{TRT}(D_{11})$ 의 부분 집합인 $\text{trt}(S_1)$ 을 구성하였다(<표 3> 참조). 여기서 문장 1(S_1)은 두 개의 태그 (commencement [T_2] & university [T_9])을 포함하고 있다. $\text{trt}(S_1)$ 으로부터 얻은 태그 관계값을 $W_{sr}(S_1)$ 에 대입하면, 단지 하나의 태그쌍(T_2 and T_9)만이 약한 의미적 관계를 갖고 있기 때문에 $W_{sr}(S_1)$ 의 값으로 0.1(= $R_{2,9}$)을 얻는다.

<표 3> 비디오 11의 S_1 의 태그 관계 테이블 ($\text{trt}(S_1)$)

	T_2	T_9
TC_2	-	○
TC_9	-	-

(3) 문장의 최종 가중치 계산: 각 문장의 최종 가중치($W_{tot}(S_i)$)를 계산하기 위해서 요약 모형 (공식 1)에 있는 $\text{Sim}(S_i, T_{ex})$ 과 $W_{sr}(S_i)$ 의 평균값을 구한다. 앞에서 언급한 것처럼 비디오 11의 문장 1의 태그값이 0.45이고 태그 간의 의미적 관계값이 0.10이므로 문장의 최종 가중치는 $0.28([0.45+0.10]/2)$ 이 된다. 문장 가중치가 높은 순서대로 비디오의 재생시간에 따라서 2~8개의 문장을 선정한다. 최종적으로 선택된 문장들을 텍스트에서 출현한 순서대로 요약문을 구성한다. 최종적으로 선택된 비디오의 요약문들은 TEDTalks의 번역 사이트(<http://www.ted.com/OpenTranslationProject>) 또는 관련된 인터넷 사이트 등을 활용하여 영문 요약문에서 한글 요약문으로 변환한다(<표 6> 참조). 나머지 59개의 표본 비디오들에도 위와 같은 절차를 적용하여 스피치 요약문을 구성한다.

4. 태그의미분석 방법의 분석과 평가

본 장에서는 표본 비디오 60개를 이용하여 연구 문제 1(플리커의 태그 클러스터와 워드넷의 동의어 정보가 비디오 태그를 확장하고 태그 간의 의미적 관계를 분석하는데 얼마나 효과적으로 이용될 수 있는가)을 분석해 본다. 그런 다음 잠재의미분석 방법이 어떤 방식에 의해서 중요한 문장들을 추출하는지 기술한 후 연구 문제 2(스피치 요약에 태그의미분석 방법과 잠재의미분석 방법을 적용할 때 이 두 방법에 의한 요약문이 품질면에서 어떤 차이를 보일 것인가)를 검증해 본다.

4.1 플리커 태그 클러스터 및 워드넷 동의어의 분석

플리커의 태그 클러스터가 비디오 태그를 확장하고 태그 간의 의미적 관계를 분석하는데 얼마나 효과적으로 적용될 수 있는지 그리고 워드넷의 동의어 정보가 태그를 확장하고 비디오 태그 간의 의미적 관계를 분석하는데 얼마나 효과적으로 적용될 수 있는지 분석해 보았다.

60개 비디오에 포함된 총 336개 태그를 플리커 태그 클러스터에 매칭했을 때 매칭 기준(태그 클러스터에 속한 태그가 초기 태그 리스트에 있다면 해당 태그 클러스터를 선정함)에 적합한 태그는 118개로 35.0%의 매칭율을 나타냈다. 이에 반해서 336개 태그를 워드넷 동의어 클러스터에 매칭했을 때 매칭 기준에 적합한 태그는 30개로 8.9%의 상대적으로 낮은 매칭율을 나타냈다. 30개 중 14개 태그(4.2%)가 플리커 태그 클러스터에 동시에 매칭되었고 워드넷에만 매칭된 태그는 16개(4.8%)에 불과하였다. 이외에 태그가 확장된 경우는 60개 비디오 중 16개의 비디오(26.7%)이었고 초기 태그 리스트가 12개 이상의 태그를 갖고 있는 경우는 태그를 확장하지 않았는데 3개의 비디오가 이 경우에 해당되었다. 태그 간의 의미적 관계를 나타낸 문장을 하나 이상 포함한 비디오는 27개(45.0%)이었다.

결론적으로 플리커의 태그 클러스터는 태그를 확장하고 비디오 태그 간의 의미적 관계를 분석하는데 효율적으로 이용되었지만 워드넷의 동의어 정보는 효율적으로 이용되지 못한 것으로 나타났다. 이와 같은 분석 결과로 볼 때 규모가 큰 비디오 표본 데이터를 통해서 검증해 볼

필요는 있지만 TEDTalks, 유튜브 등과 같은 대중을 대상으로 한 강의 자료를 분석할 경우 태그 간의 의미적 관계 정보를 얻기 위해서 플리커 태그 클러스터만을 사용해도 큰 무리는 없는 것으로 보인다.

4.2 태그의미분석 방법과 잠재의미분석 방법 간의 비교 분석

4.2.1 잠재의미분석 방법

잠재의미분석 방법에 의한 스피치 요약하기 위해서 MATLAB 7.1 패키지를 이용하였다. 비디오 20("Clay Shirky: How cellphones, Twitter, Facebook can make history")의 자막에서 출현 빈도가 3 이상인 단어들 중 불용어와 기능어를 제외한 15개의 키워드를 최종 추출하였다. 키워드 추출을 위해서는 Wordcounter 사이트(<http://www.wordcounter.com/>)를 이용하였다. <표 4>는 단어-문장 행렬로 15개의 단어가 34개의 문장에 분포된 것을 기술한 것이다.

<표 5>는 <표 4>를 입력 데이터로 하여 MATLAB 7.1의 특이값 분해(SVD) 기능을 이용하여 단어-문장 행렬을 2차원으로 축소한 결과이다. 다시 말해서 15개의 단어로 표현된 주제(15 × 34)를 2개의 개념(2 × 34)으로 축소한 것이다. 개념 1은 제1주제를 나타내고 개념 2는 제2주제를 나타낸다. <표 5>에서 개념을 잘 나타낸 문장(들)을 추출하는 방법으로 교차 방법(cross method)을 이용하였다. 교차 방법은 각 개념의 평균값을 계산한 후 평균값 이하인 값을 영(0)으로 변환한 후 두 개념의 합을 계산한 후 합계가 높은 값 순으로 문장을 선정하는 방안이다. 교차 방

〈표 4〉 단어-문장 행렬

	s1	s136	s195	s83	s118	s165	s197	s117	s19	s20	s35	s40	s43	s47	s62	s66	s69	s80	s81	s84	s85	s91	s98	s114	s147	s190	s191	s41	s53	s184	s67	s159	s168	s180			
media	1	1	2	0	1	1	1	0	1	1	1	0	1	1	0	0	0	0	0	0	0	0	1	1	1	0	0	0	2	2	0	0	0	0			
group	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2	0	0	0	0	2		
internet	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	2	0	0	0	
message	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0		
world	1	0	2	0	1	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
conversat ion	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
obama	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	2	0	
citizen	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	
china	0	0	0	0	1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0
social	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
landscape	1	0	0	0	0	0	0	0	0	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0
network	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0
quake	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0
chinese	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0
twitter	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0

〈표 5〉 개념-문장 행렬

문장 번호	s1	s136	s195	s83	s118	s165	s197	s117	s19	s20	s35	s40	s43	s47	s62	s66	s69										
개념 1	0.28	0.19	0.45	0.01	0.21	0.18	0.23	0.09	0.17	0.19	0.19	0.05	0.19	0.01	0.19	0.19	0.17										
개념 2	0.02	0.33	0.06	0	0	0.05	0.07	0.04	-0.03	-0.09	-0.09	0.03	-0.09	0	-0.09	-0.09	-0.03										
총점	0.28	0.52	0.51	0	0.21	0.23	0.3	0	0.17	0.19	0.19	0	0.19	0	0.19	0.19	0.17										
문장 번호	s80	s81	s84	s85	s91	s98	s114	s147	s190	s191	s41	s53	s194	s67	s159	s168	s180	평균									
개념 1	0.03	0.07	0	0.01	0	0.15	0.15	0.19	0.02	0.02	0.06	0.3	0.34	0.03	0.01	0.01	0.02	0.13									
개념 2	0	0.03	0	0.01	0	-0.04	-0.04	-0.09	0.32	0.32	0.57	-0.02	-0.12	-0.01	0.08	0.16	0.49	0.05									
총점	0	0	0	0	0	0.15	0.15	0.19	0.32	0.32	0.57	0.3	0.34	0	0.08	0.16	0.49										

법을 사용한 이유는 교차 방법이 다른 잠재의 미분석 기반 접근 방법들에 비해서 효율적인 것으로 나타났기 때문이다(Ozsoy, Alpaslan, &

Cicekli 2011). 교차 방법에 의해서 비디오 20의 경우 총점이 0.32 이상인 8개(s41, s136, s190, s191, s194, s195, s197, & s180)의 문장이 비디

오의 주제를 잘 나타내는 문장들로 선택되었다 (<표 6> 참조). 동일한 방법으로 나머지 59개의 표본 스피치의 요약문을 구성하였다.

4.2.2 비교 분석 결과

(1) 표준 요약

태그의미분석 방식의 효율성을 평가하기 위해서 각 비디오의 표준 요약문 두 명의 연구자들이 공동으로 작성하였다. 작성 방법은 연구자들이 비디오를 시청한 다음 유튜브 사이트에 있는 텍스트 요약 및 메타데이터 내용을 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 요약문을 구성하였다.

(2) 요약문 내용 비교

표본 비디오의 전체를 대상으로 한 제안한 방법과 잠재의미분석 방법의 효율성을 비교하기 전에 4개 비디오의 요약문 본문을 비교해 보기로 한다. <표 6>에서 확인할 수 있는 것처럼 선정된 문장의 수는 같지만 잠재의미분석 방법은 제안한 방법에 비해서 상대적으로 긴 문장들을 선정하였다. 밑줄 친 문장은 표준 요약문과 일치하는 문장을 나타낸다. 두 개의 방식에 의해서 선정된 동일한 문장의 비율은 60개의 요약문을 비교해 결과, 12.0%로 그다지 높지 않게 나타났다.

요약문에 의해서 비디오의 주제를 나타내는 정확성은 비디오에 따라서 다른 차이를 보였다.

스티브 잡스의 스탠포드 졸업식 연설문을 나타내는 비디오 11의 경우, 제안 방식에 의한 요약문은 대학 졸업식에서 한 연설문이라는 사실을 알려 주면서 화자(스티브 잡스)와 애플사와의 관계를 알려 주는 반면 잠재의미분석 방식에 의한 요약문에서는 화자가 애플사를 중심으로 어떤 활동과 변화를 경험했는지에 대한 정보만 제공하고 있다. 변화된 미디어 지형을 다룬 비디오 20은 제안한 방식의 요약이 잠재의미분석 방식의 요약문보다 좀 더 구체적으로 어떻게 소셜 미디어가 역사를 만들어내는가에 대한 것을 보여주고 있는 것으로 보인다.

한편, 팀 버나스리의 링크드 데이터에 대한 비디오 15의 경우는 잠재의미분석 방식의 요약문이 제안한 방식의 요약문보다 링크드 데이터에 대한 개념과 실현 방법에 대한 더 많은 정보를 제공하고 있다. 이외에 디지털 도서관을 다룬 비디오 60의 경우는 잠재의미분석 방식의 요약문이 제안한 방식의 요약문보다 디지털 도서관에 대한 화자의 생각을 좀 더 자세하게 보여주고 있는 것으로 보인다.

(3) 효율성 비교

60개의 전체 표본 비디오를 사용하여 제안한 방식과 잠재의미분석 방식을 표준 요약문을 기준으로 하여 정확률과 재현율을 대체하는 하나의 척도인 F 측정($(2 \times \text{재현율} \times \text{정확률}) / (\text{재현율} + \text{정확률})$)을 이용하여 비교해 보았다. 비교한 결과, 제안한 방식의 평균 F값(0.27)이 잠재의미분석방식의 평균 F값(0.22)보다 높게 나타났다으나 t-검증 결과 통계적으로 유의미한 차이는 없었다($p(=0.21) > 0.05$)(<표 7> 참조).

<표 6> 요약문 사례

비디오(재생시간)	표준 요약	태그의미분석 방법	잠재의미분석 방법
(11) Steve Jobs' 2005 Stanford commencement address (15:04)	<p>S1: 감사합니다. 오늘 세계 최고의 명문들 중 하나인 이 대학의 학위수여식에서 여러분과 함께 하게 되어 영광입니다.</p> <p>S8: 첫 번째 이야기는 인생의 전환점들을 있는 일입니다.</p> <p>S46: 무언가를 믿어야 합니다. 그게 용기든, 운명이든, 인생이든, 인연이든, 그 무엇이든 간에 말입니다.</p> <p>S48: 두 번째 이야기는 사랑과 상실에 관한 것입니다.</p> <p>S68: 당시에는 몰랐지만 결과적으로 애플사에서 해고당한 것이 제 인생 최고의 전환점이었던 걸로 드러났습니다.</p> <p>S80: 사랑할 만한 것을 찾으십시오.</p> <p>S90: 세 번째 이야기는 죽음에 관한 것입니다.</p> <p>S120: 여러분에게 주어진 시간은 유한합니다. 남의 인생을 사느라 그 시간을 낭비하지 마십시오.</p>	<p>S1: 감사합니다. 오늘 세계 최고의 명문들 중 하나인 이 대학의 학위수여식에서 여러분과 함께 하게 되어 영광입니다.</p> <p>S3 : 이번이 제가 대학 졸업식장에 가장 가까이 와 본 경우입니다.</p> <p>S19: 하지만 참 순진하게도 이곳 스탠포드만큼이나 등록금이 비싼 대학을 골랐고, 노동자층이었던 양부모님이 평생 모으신 돈이 모두 제 등록금으로 소모되었습니다.</p> <p>S71: 그 후 5년간 저는 「넥스트」라는 회사와 「픽사」라는 또 다른 회사를 차렸습다. 그리고 나중에 제 아내가 될 대단한 여성을 만나 사랑에 빠졌습니다.</p> <p>S72: 픽사는 발전을 거듭하여 세계 최초의 컴퓨터 애니메이션 영화인 「토이스토리」를 제작했고, 현재 전세계에서 가장 성공한 애니메이션 스튜디오가 되었습니다.</p> <p>S73: 놀라운 발전이 일어나 애플이 넥스트를 인수했고 저는 애플로 돌아왔습니다. 그리고 넥스트에서 개발했던 기술이 바로 오늘날 애플의 부흥을 이루어 낸 핵심입니다.</p>	<p>S51: 10년 만에 애플은 달랑 두 명뿐인 차고에서 4000명이 넘는 종업원을 거느린 20억 달러 규모의 대기업으로 성장했습니다.</p> <p>S54: 어떻게 자신이 창업한 회사에서 해고당할 수 있을까요? 애플사가 성장하면서 회사를 운영해 갈 대단한 재능을 지닌 것으로 여겨지던 사람을 채용했고, 첫 1년여 기간은 순조로웠습니다.</p> <p>S68: 당시에는 몰랐지만 결과적으로 애플사에서 해고당한 것이 제 인생 최고의 전환점이었던 걸로 드러났습니다.</p> <p>S71: 그 후 5년간 저는 「넥스트」라는 회사와 「픽사」라는 또 다른 회사를 차렸습다. 그리고 나중에 제 아내가 될 대단한 여성을 만나 사랑에 빠졌습니다.</p> <p>S73: 놀라운 발전이 일어나 애플이 넥스트를 인수했고 저는 애플로 돌아왔습니다. 그리고 넥스트에서 개발했던 기술이 바로 오늘날 애플의 부흥을 이루어 낸 핵심입니다.</p> <p>S75: 제가 애플에서 해고되지 않았더라면 이런 일들이 하나도 이루어지지 않았을 것이라 확신합니다.</p>
(15) Tim Berners-Lee: The next Web of open, linked data (6:51)	<p>S44: 여전히 커다란 좌절감이 존재합니다. 현재 웹에 널려있는 정보가 대체로 "문서"이지 "데이터"가 아니기 때문입니다.</p> <p>S61: 단지 두 조각의 데이터 혹은 그가 했듯이 여섯 조각의 데이터가 연결되는 것이 아니라 누구나 데이터를 웹에 올리고 연결하는 생각을 해보자는 겁니다. 실질적으로 우리가 상상할 수 있는 거의 모든 것이 웹에 존재하게 되겠죠. 데이터가 모두 연결되는 것입니다. 이런 걸 "Linked Data"라고 합니다.</p> <p>S63: 웹에 무언가를 올린다고 생각해볼까요? 여기엔 세 가지 규칙이 있습니다. 첫번째 규칙은 바로 그러니까 HTTP: 으로 시작하는 주소가 있다는 겁니다. 요즘은 오직 웹페이지 문서에만 HTTP 주소가 필요한 게 아닙니다. 최근엔 문서를 뛰어넘어 문서가 다루는 "실체"에도 주소가 생기기 시작했습니다.</p> <p>S88: 그는 위키피디아의 그 상자들로부터 데이터를 추출하는 프로그램을 썼습니다. 그러곤 그 데이터들을 다른 링크데이터 문체에 연결시켰죠. 그것을 그는 Dbpedia라 불렀습니다.</p> <p>S97: 정부의 데이터, 기업의 데이터는 매우 중요합니다. 과학적 데이터도 있고 개인적 데이터도 있습니다. 기후 데이터도 있고 사건들에 대한 데이터도 있습니다. 이야기(Talk)에 대한 데이터도 있고, 뉴스도 있고, 데이터는 모든 형태를 망라합니다.</p> <p>S165 : 정보가 유의미하게 연결되면 그건 곧 "힘"이 됩니다. 문서 단위나 페이지 단위로 링크했던 기존의 웹과는 다른 방식으로 말합니다.</p>	<p>S61: 단지 두 조각의 데이터 혹은 그가 했듯이 여섯 조각의 데이터가 연결되는 것이 아니라 누구나 데이터를 웹에 올리고 연결하는 생각을 해보자는 겁니다. 실질적으로 우리가 상상할 수 있는 거의 모든 것이 웹에 존재하게 되겠죠. 데이터가 모두 연결되는 것입니다. 이런 걸 "Linked Data"라고 합니다.</p> <p>S63: 웹에 무언가를 올린다고 생각해볼까요? 여기엔 세 가지 규칙이 있습니다. 첫번째 규칙은 바로 그러니까 HTTP: 으로 시작하는 주소가 있다는 겁니다. 요즘은 오직 웹페이지 문서에만 HTTP 주소가 필요한 게 아닙니다. 최근엔 문서를 뛰어넘어 문서가 다루는 "실체"에도 주소가 생기기 시작했습니다.</p> <p>S88: 그는 위키피디아의 그 상자들로부터 데이터를 추출하는 프로그램을 썼습니다. 그러곤 그 데이터들을 다른 링크데이터 문체에 연결시켰죠. 그것을 그는 Dbpedia라 불렀습니다.</p> <p>S97: 정부의 데이터, 기업의 데이터는 매우 중요합니다. 과학적 데이터도 있고 개인적 데이터도 있습니다. 기후 데이터도 있고 사건들에 대한 데이터도 있습니다. 이야기(Talk)에 대한 데이터도 있고, 뉴스도 있고, 데이터는 모든 형태를 망라합니다.</p> <p>S165 : 정보가 유의미하게 연결되면 그건 곧 "힘"이 됩니다. 문서 단위나 페이지 단위로 링크했던 기존의 웹과는 다른 방식으로 말합니다.</p>	<p>S7: html이라는 것의 초안을 만들게 됐습니다. 그걸 주고받는 통신 규약인 HTTP도요, 어떤 정보의 위치를 담는 주소의 개념으로 URL이라는 것도 생각해 봤습니다. 잘 아시다시피 http://로 시작하죠.</p> <p>S61: 단지 두 조각의 데이터 혹은 그가 했듯이 여섯 조각의 데이터가 연결되는 것이 아니라 누구나 데이터를 웹에 올리고 연결하는 생각을 해보자는 겁니다. 실질적으로 우리가 상상할 수 있는 거의 모든 것이 웹에 존재하게 되겠죠. 데이터가 모두 연결되는 것입니다. 이런 걸 "Linked Data"라고 합니다.</p> <p>S63: 웹에 무언가를 올린다고 생각해볼까요? 여기엔 세 가지 규칙이 있습니다. 첫번째 규칙은 바로 그러니까 HTTP: 으로 시작하는 주소가 있다는 겁니다. 요즘은 오직 웹페이지 문서에만 HTTP 주소가 필요한 게 아닙니다. 최근엔 문서를 뛰어넘어 문서가 다루는 "실체"에도 주소가 생기기 시작했습니다.</p> <p>S66: 두 번째 규칙은 예를 들면 이런 것입니다. 사람들은 정보를 검색합니다. 웹에서 자기가 원하는 정보를 담고 있는 HTTP 주소를 찾는 거죠. HTTP라는 프로토콜(규약)을 통해서 말합니다. 주소를 찾아가 약속과 형식에 따라 저장된 정보를 얻어 온다는 것이죠. 행사든 물건이든 뭐가 됐든 우리에게 유용한 정보가 프로토콜(규약)에 따라 저장되어 있고 우리는 그 정보를 얻어 오는 겁니다.</p> <p>S103: 투명성 때문만은 아닙니다. 정부의 투명성은 중요하지요. 하지만 여기에서 데이터란 정부의 모든 기관으로부터의 데이터입니다. 미국인이 어떤 삶을 사는지를 말해 주는 중요한 정보가 얼마나 많았어요.</p>

비디오(재생시간)	표준 요약	태그의미분석 방법	잠재의미분석 방법
(20) Clay Shirky: How cellphones, Twitter, Facebook can make history (17:03)	<p>S1: 제가 얘기하고자 하는 것은 변화된 미디어 지형에 관한 것과, 또한 전세계로 전하고자 하는 메시지가 있는 모든 이들에게 그런 변화가 지니는 의미입니다. S20: 그래서 이제 우리가 마주하기 시작한 미디어 지형에서는 모든 곳에 걸쳐 혁신이 일어나고 있습니다. S47: 인터넷은 생겨날 때부터 집단과 사회적 대화를 모두 지원하는 역사상 최초의 매체입니다. S61: 세 번째의 큰 변화는, Dan Gilmore의 표현을 빌자면 한 때는 청중에 속했던 이들이 이제는 소비자가 아닌 생산자도 될 수 있다는 것입니다. S83: BBC는 트위터를 통해서 중국의 지진을 처음 알게 되었습니다. S136: 20세기의 고전적인 미디어 문제는 하나의 조직이 자신의 메시지를 네트워크의 가장자리에 분산되어 있는 여러 사람들에게 어떻게 전달할 수 있을 것인가입니다. S165: 사회적 미디어의 가장 창의적인 활용법을 우리는 오바마 대선 캠페인 동안에 볼 수 있었죠. S199: 지금 우리가 직면하고 있는 문제는 이제까지 익숙하게 해 온 방식을 변화시켜 이 미디어를 잘 활용하는 방안을 생각하는 것입니다.</p>	<p>S4: 지난 11월에, 대통령 선거가 있었습니다. S72: 지난 5월, 중국 사천성 지역에는 진도 7.9짜리 끔찍한 지진이었습니다. 리히터 스케일 기준에서 보듯, 넓은 지역에 막대한 피해를 입혔죠. S78: 트위터로 말했습니다. S83: BBC는 트위터를 통해서 중국의 지진을 처음 알게 되었습니다. S101: 사람들이 점차 알아내기 시작한 바로는 사천성에서 그 많은 학교들이 붕괴한 이유로 지진이 비극적이게도 학교수업시간 중에 일어났기 때문이라는 것입니다. 그러니까 학교건물들이 많이 무너진 이유는 부패한 공무원들이 뇌물을 받고 부실공사를 했기 때문이라는 것이었습니다. S117: 중국은 아마도 세계에서 가장 성공적인 인터넷 검열 관리자일 것입니다. 흔히 인터넷 만리장성(Great Firewall of China)이라고 부르기도 합니다. S165: 사회적 미디어의 가장 창의적인 활용법을 우리는 오바마 대선 캠페인 동안에 볼 수 있었죠. S168: 오바마가 한 것들 중에 유명한 것 중 하나는 이런 것입니다. 그러니까 오바마 캠페인 진영은 MyBarakobama.com 과 myBO.com을 만들었습니다.</p>	<p>S41: 집단에게 말을 걸고자 하면, 똑같은 메시지를 집단 속의 모두에게 전달합니다. S136: 20세기의 고전적인 미디어 문제는 하나의 조직이 자신의 메시지를 네트워크 상에 분산되어 있는 많은 사람들에게 어떻게 전달할 수 있을 것인가입니다. S180: 만들어진 후 수일 내에 myBO.com에서 가장 빠른 성장을 보이는 모임이 되었습니다. S190: 사람들은 오바마가 그들을 결코 침묵시키지 않았다는 것을 알아차렸습니다. S191: 오바마 캠페인 진영의 누구도 그 모임을 숨기거나 가입을 더 어렵게 만들거나 존재를 부정하거나 지우거나 사이트에서 내려버리려고 하지 않았죠. S194: 미디어, 우리가 과거에 알고 있는 미디어 지형은 무척 익숙했었고, 그 개념에 있어서 직업적 생산자들이 메시지를 아마추어들에게 전달한다는 발상을 쉽게 다룰 수 있었던 그 미디어는 이제 점점 사라지고 있습니다. S195: 미디어가 전지구적이고 사회적이고 어디에나 있고, 저렴한 세계에서, 한 때의 청중들이 이제는 점점 더 참여자가 되어가는 미디어의 세계에서 미디어는 점점 개인이 소비하도록 고안된 단일한 메시지를 만드는 일과는 거리가 멀어지게 됩니다. S197: 지구상의 누군가 들어주기를 바라는 메시지를 가진 사람이 내려야 할 선택은 현재 이런 변화가 우리가 활동하고 싶어 하는 미디어 환경인지 아닌지가 아닙니다.</p>
(60) Brewster Kahle: A digital library, free to the world	<p>S4: 엘리트, 부모, 도서관 사서, 전문가 등이 실제로 하고 있는 일련의 행동들은, 우리 주변에서 쉽게 얻을 수 있는 지식 또는 폭넓은 지식들을 다음 세대에 알려주기 위해 노력하는 것입니다. S9: 저는 사서입니다. 그리고 제가 하고자 하는 일은 지식의 모든 저작물들을 그것을 읽고 싶어 하는 가능한 많은 사람들에게 제공하는 것입니다. S22: 제가 하려는 방법은 아마존닷컴 웹사이트와 비슷합니다. -책, 음악, 비디오부터 미디어 종류들을 하나씩 하나씩 해나가는 겁니다. 자, 어떻게 해야 할까요? 만약 책부터 시작한다면, 우리는 지금 어디에 있는 걸까요? 가장 먼저, 엔지니어처럼 문제의 범위를 살펴봐야겠죠. S202: 이것이 우리가 진정으로 살고 싶은 세상인가요? 이러한 일들에, 일반기업과 비교해서 공공기관이 할 수 있는 역할은 무엇일까요? 어떻게 하면 미래에도 도서관과 출판업계가 공존하는 세상을 만들 수 있을까요? 우리가 자라면서 기본적으로 누려왔던 것처럼 말입니다. 모든 지식에 대한 자유로운 접근, 저는 이것을 인류의 가장 위대한 성과 중 하나라고 생각합니다. 달에 착륙한 인류나 구텐베르크 성서, 또는 알렉산드리아 도서관처럼 업적이 될 수 있습니다. S205: 이 나라의 위대한 자본가 중 한 사람인 키네기는 그의 유산 위에 이렇게 세웠습니다. "모든 이에게 무료입니다."</p>	<p>S13: 이집트인들은 부지런함으로 알렉산드리아 도서관을 건설할 수 있었습니다. 전 세계 모든 사람들이 가지고 있는 책들을 보관하겠다는 생각이었죠. S22: 제가 하려는 방법은 아마존닷컴 웹사이트와 비슷합니다. -책, 음악, 비디오부터 미디어 종류들을 하나씩 하나씩 해나가는 겁니다. 자, 어떻게 해야 할까요? 만약 책부터 시작한다면, 우리는 지금 어디에 있는 걸까요? 가장 먼저, 엔지니어처럼 문제의 범위를 살펴봐야겠죠. S60: 그리고 약 30일 동안 살리콘 벨리에서 몇몇 사람들을 구했습니다. 그들은 우간다로 가서, 차를 사고, 우간다 국립 도서관에 인터넷을 연결시키고, 우간다 사람들이 원하는 것을 파악한 후에, 우간다 시골 지역에서 책들을 만드는 프로그램을 시작했습니다. S187: 이것은 사실 알렉사 인터넷(Alexa Internet: 회사명)이 그들의 자료를 인터넷 저장소에 기증하면서 시작되었습니다. S195: 이 사진은 알렉산드리아 도서관에 있는 인터넷 저장소의 모습입니다.</p>	<p>S13: 이집트인들은 부지런함으로 알렉산드리아 도서관을 건설할 수 있었습니다. 전 세계 모든 사람들이 가지고 있는 책들을 보관하겠다는 생각이었죠. S23: 얼마나 규모가 클까요? 만약 모든 출판물들을 온라인에 올려놓으려고 한다면, 그래서 누구나 이용할 수 있도록 하려면, 얼마나 크게 일을 벌여야 할까요? 우리는 확실히 모르지만, 세계에서 가장 큰 규모가 도서관은 미의회 도서관입니다. S54: 이것은 이집트의 알렉산드리아 도서관이 개장하던 날의 사진입니다. S128: 보스던 공립 도서관에서 절판된 책들을 빌려주고 있습니다. 우즈 홀 해양과학연구소와 기타 여러 도서관들이 프로그램에 참여하기 시작했고, 도서관과 서점이 해야 할 일인지 구분하려고 노력하고 있습니다. S202: 이것이 우리가 진정으로 살고 싶은 세상인가요? 이러한 일들에, 일반기업과 비교해서 공공기관이 할 수 있는 역할은 무엇일까요? 어떻게 하면 미래에도 도서관과 출판업계가 공존하는 세상을 만들 수 있을까요? 우리가 자라면서 기본적으로 누려왔던 것처럼 말입니다. 모든 지식에 대한 자유로운 접근, 저는 이것을 인류의 가장 위대한 성과 중 하나라고 생각합니다. 달에 착륙한 인류나 구텐베르크 성서, 또는 알렉산드리아 도서관처럼 업적이 될 수 있습니다.</p>

〈표 7〉 비교 결과

방 법	F 측정	재현율	정확률
태그의미분석	0.27	0.29	0.29
잠재의미분석	0.22	0.22	0.23

태그의미분석 방식의 요약문 품질이 평균 키워드수(태그수: 5.6개, 본문키워드수: 12.4개)가 상대적으로 적음에도 불구하고 잠재의미분석 방식의 요약문 품질과 동일한 또는 조금 우수한 것으로 나타난 것은 다음과 같은 이유 때문으로 생각된다. 첫째, 이용자가 비디오의 주제를 인지한 후 할당된 태그를 사용한 태그 방식이 본문 텍스트 전체를 처리하여 중요한 개념을 추출한 잠재의미분석 방식보다 비디오의 주제를 좀 더 효율적으로 표현한다고 추측할 수 있다. 이러한 결과는 비디오 태그가 색인어로서 충분한 가치가 있다는 주장한 Heckner, Neubauer, & Wolff(2008)과 김현희(2011)의 연구를 확인시켜 준다. 둘째, 태그는 본문 키워드보다 개체명(named entities)을 더 많이 포함하고 있는 점이다. Christensen et al.(2003)은 개체명은 텍스트 요약보다 스피치 요약에서 더 중요한 기준이 된다고 보고하고 있다. 60개의 표본 비디오에 할당된 총 336개 태그와 720개의 본문 키워드(각 표본 비디오 자막에서 출현빈도가 3번 이상이면서 불용어 및 기능어를 제외한 단어)가 개체명(인명, 지명, 기관명 및 연도)을 얼마나 포함하고 있는지 분석해 보았다. 분석 결과, 태그의 경우는 16.4%가 본문 키워드의 경우는 4.9%가 개체명에 속해 있었다. 즉, 태그가 본문 키워드보다 대략 세 배 정도 더 많이 개체명을 포함하고 있었다.

태그의미분석 방식이 잠재의미분석 방식과

비교하여 동일한 또는 더 효율적으로 나타난 것은 여러 가지 면에서 의미가 있다. 첫째, 제안한 방식은 플리커의 태그 클러스터 데이터베이스를 이용하여 간단한 알고리즘으로 구현할 수 있는 반면 잠재의미분석 방식은 대규모의 단어-문헌 행렬과 같은 입력 데이터가 필요하고 복잡한 특이값 분해 계산이 요구된다. 둘째, 태그 간의 의미적 관계 정보는 스피치를 요약하는 것 외에 스피치 비디오의 검색시 질의 확장 또는 비디오 검색 결과를 브라우징하는 정보검색 환경에서 유용하게 사용될 수 있다는 점이다.

끝으로, 요약문 내용 비교의 결과에서 확인할 수 있는 것처럼 태그의미분석 방식은 잠재의미분석 방식과는 다른 특성을 갖는 요약문을 추출하는 것으로 생각된다. 특히 잠재의미분석 방법은 몇몇 중요한 개념을 집중적으로 다루는 문장을 선택하는 경향이 있는 반면, 태그의미분석 방식은 상대적으로 더 많은 개념 또는 개체명을 표현하는 문장이 선택되는 경향을 보인다. 따라서, 대규모 표본 비디오들을 대상으로 한 실험과 분석을 통해서 이러한 경향에 대한 좀 더 심층적인 검증이 필요해 보인다.

5. 결론

본 연구는 스피치 요약을 위해서 비디오 태그를 확장하고 태그 간의 의미적 관계 정보를 이용할 수 있는 태그의미분석 방법을 제안하였다. 그런 다음, 플리커의 태그 클러스터와 워드넷의 동의어 정보가 태그 간의 의미적 관계를 분석하는데 얼마나 효과적으로 이용될

수 있는지 분석하고 제안한 방법의 특성과 효율성을 조사해 보기 위해서 이를 잠재의미분석 방법과 비교해 보았다. 그 결과를 요약하면 다음과 같다.

첫째, 비디오 태그를 확장하고 태그 간의 의미적 관계를 분석하는데 플리커의 태그 클러스터는 효과적으로 이용되었지만 워드넷의 동의어 정보는 효과적으로 이용되지 못한 것으로 나타났다. 이와 같은 분석 결과로 볼 때 규모가 큰 표본 데이터를 통해서 다시 검증해 볼 필요는 있지만 일반 대중을 대상으로 한 스피치 자료를 분석할 경우 태그 간의 의미적 관계 정보를 얻기 위해서 플리커 태그 클러스터만을 사용해도 큰 무리는 없는 것으로 보인다.

둘째, 제안한 방법의 F값(0.27)이 잠재의미분석 방법의 F값(0.22)보다 높게 나타났으나 통계적으로 유의미한 차이는 없었다. 이와 같이 평균 5.6개의 태그와 태그 클러스터 데이터베이스를 이용하여 구성된 요약문의 품질이 평균 12.4개의 본문 키워드를 입력 데이터로 하여 복잡한 특이값 분해 계산 과정을 통해서 구성된

요약문의 품질과 유사하게 나타난 사실은 효율성 면에서 매우 고무적인 일이다.

끝으로 제안한 태그의미분석 방법에서 자막 분석을 위해서 태그와 함께 표제의 키워드를 함께 사용한다면 좀 더 높은 품질의 요약물을 생성할 수 있을 것으로 생각된다. 본 연구에서 제안한 알고리즘은 실제 자막에 출현한 태그가 최소한 2개 이상이 되어야만 태그를 확장하고 태그 간의 의미적 관계를 분석할 수 있다. 60개 표본 비디오를 분석해 본 결과, 6개의 비디오들(10.0%)은 단지 하나의 태그만이 해당 비디오 자막에 출현하였고 한 비디오(1.7%)는 태그가 전혀 출현하지 않아서 요약을 할 수 없었다. 따라서 태그의미분석 방법은 태그를 표제의 키워드와 함께 활용하여 이런 문제점이 보완된다면 소셜 메타데이터가 생성되는 환경에서 스피치 요약에 유용하게 이용될 수 있을 것이다. 또한 태그 간의 의미적 관계 정보는 스피치 비디오를 검색할 때 질의 확장이나 비디오 검색 결과를 브라우징하는 정보검색 환경에서 유용하게 사용될 수 있을 것이다.

참 고 문 헌

- [1] 김현희. 2009. 비디오의 오디오 정보 요약 기법에 관한 연구. 『정보관리학회지』, 26(3): 169-188.
- [2] 김현희. 2012. 이용자 태그를 활용한 비디오 스피치 요약의 자동 생성 연구. 『한국문헌정보학회지』, 46(1): 163-181.
- [3] 정영미. 2005. 『정보검색연구』. 서울: 구미무역출판부.
- [4] Boydell, O., & Smyth, B. 2010. "Social summarization in collaborative web search." *Information Processing and Management*, 46(6): 782-798.
- [5] Christensen, H. et al. 2003. "Are extractive text summarisation techniques portable to broadcast news?" *Proceedings of Automatic Speech Recognition and Understanding Workshop*, St. Thomas,

- USA, 489-494.
- [6] Gong, Y., & Liu, X. 2001. "Generic text summarization using relevance measure and latent semantic analysis." *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 19-25.
- [7] Heckner, M., Neubauer, T., & Wolff, C. 2008. "Tree, funny, to_read, Google: What are tags supposed to achieve?" *Proceedings of the 2008 ACM Workshop on Search in Social Media*, Napa Valley, California, USA.
- [8] Hennig, L. 2009. "Topic-based multi-document summarization with probabilistic latent semantic analysis." *Proceedings of International conference on Recent Advances in NLP*, 144-149.
- [9] Heu, J. et al. 2013. Multi-document summarization exploiting semantic analysis based on tag cluster. In S. Li et al. (Eds.), *Advances in Multimedia Modeling, Lecture Notes in Computer Science*, 7733, 479-489.
- [10] Hong, R. et al. 2011. "Beyond search: Event driven summarization for web videos." *ACM Transactions on Multimedia Computing, Communications, and Applications*, 7S(1): 24-43.
- [11] Kim, H. 2011. "Toward video semantic search based on a structured folksonomy." *Journal of the American Society for Information Science*, 62(3): 478-492.
- [12] Matusiak, K. 2006. "Towards user-centered indexing in digital image collections." *OCLC Systems & Services: International digital library*, 22(4): 283-298.
- [13] Ozsoy, M., Alpaslan, F., & Cicekli, I. 2011. "Text summarization using Latent Semantic Analysis." *Journal of Information Science*, 37(4): 405-417.
- [14] Specia, L., & Motta, E. 2007. Integrating folksonomies with the semantic Web. In E. Franconi, M. Kifer, & W. May (Eds.), *The Semantic Web: Research and Applications, Lecture Notes in Computer Science*, 4519, 624-639.
- [15] Steinberger, J., & Jezek, K. 2004. "Using latent semantic analysis in text summarization and summary evaluation." *Proceedings of ISIM '04*, 93-100.
- [16] Yamamoto, D., Masuda, T., Ohira, S., & Nagao, K. 2008. "Collaborative video scene annotation based on tag cloud." *Proceedings of the Advances in Multimedia Information Processing*, Tainan, Taiwan, 397-406. <<http://tk-www.elcom.nitech.ac.jp/~daisuke/pdf/conf/pcm2008.pdf>>.
- [17] Zhu, J. et al. 2009. "Tag-oriented document summarization." *Proceedings of the 18th international conference on World Wide Web*, 1195-1196.
- [18] Wang, M. et al. 2012. "Event driven Web video summarization by tag localization and key-shot identification." *IEEE Transactions on Multimedia*, 14(4), 975-985. <<http://137.132.145.151/lms/sites/default/files/publication-attachments/wang-06135507.pdf>>.

• 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. 2012. "Investigating an automatic method in summarizing a video speech using user-assigned tags." *Journal of the Korean society for Library and Information Science*, 46(1): 163-181.
- [3] Chung, Y. 2005. *Information retrieval research*. Seoul: Gumi Trading Publisher.