음성 인식기의 N-best 결과를 이용한 한국어 뉴스 음성의 주제 탐지

백 록 선, 양 일 호, 김 명 재, 허 희 수, 유 하 진
서울시립대학교 컴퓨터과학부

Topic detection of Korean news speech using ASR N-best lists

Rock-Seon Baek, IL-Ho Yang, Myung-Jae Kim, Hee-Soo Heo and Ha-Jin Yu
School of Computer Science, University of Seoul,
whites86@naver.com, heisco@hanmail.net, mj@uos.ac.kr,
zhasgone@naver.com, hjyu@uos.ac.kr

Abstract

In this paper, we evaluate the topic detection in Korean news speech using ASR (automatic speech recognition) N-best lists. We use LDA(latent Dirichlet allocation) for finding topics of the spoken news in Korean. For latent variable modeling, we use three kinds of N(1,3,5)-best lists and original text documents. We compare the system using ASR N-best lists with the system using text documents.

1. 서론

초고속 인터넷, 대용량 데이터베이스와 같은 정보기술이 발전함에 따라 대량의 데이터를 보관하고 네트워크를 통해 접근할 수 있게 되었다. 이로 인해 사용자가 원하는 데이터에 보다 신속하고 정확하게 접근할 수 있도록 각 데이터를 효과적으로 분류할 필요가 발생하였다. 주제 탐지(topic detection)는 대량의 문서 데이터를 유사한 주제 단위로 군집화 함으로써 이러한 문제를 해결하기 위한 한 방법이 될 수 있을 것이다.

일반적으로 주제 탐지는 텍스트 형태의 문서에 대해서 수행하나 본 연구에서는 이를 이용하여 한국어 음성 데이터를 분류하고자 하였다. 음성 데이터의 주제 탐지를 위해 주어진 음성에 대한 텍스트 형태의 음성 인식 결과를 획득하고, 이에 대한 주제 탐지를 수행하는 두 단계 작업을 거쳤다.

텍스트 문서에 대해 주제 탐지를 수행하는 경우와 달리 잡음 등으로 인해 악영향을 받아 음성 인식 결과에 오류가 발생할 수 있다. 따라서 이러한 악영향을 완화하기 위한 방안으로 1-best 결과만 사용하는 것이 아니라 N-best 음성 인식 결과를 고려하여 주제를 탐지하였다. N-best 인식 결과를 고려하면 1-best 결과에서 오인식한 단어를 정인식하는 경우가 생겨 중요 단어의 빈도를 높일 수 있어 성능이 향상될 것으로 기대한다.

2. 데이터베이스

2.1 뉴스 동영상 수집

실험을 위해 미디어다음에서 제공하는 MBC와 SBS의 뉴스 동영상 1,056개(2013년 4월 18일부터 4월 27일까지 10일 분량)를 수집하였다. 수집한 동영상의 총 재생 길이는 약 24시간이다. 또한, 비교를 위해 뉴스의 전사 문서도 수집하였다.

2.2 음성-텍스트 변환

FFmpeg을 사용하여 44kHz, 2채널 뉴스 동영상을 16kHz, 1채널 wav 음성 파일로 변환하였다. 하나의 긴 파일을 여러 개의 짧은 음성 샘플로 분할하여 구글 음성 인식기를 사용해 텍스트 문서로 변환하였다. 1-best 결과일 때 단어 단위 음성 인식 정확도는 30.65%이다.

2.3 형태소 분석 및 불용어 제거

한국어의 경우 형태소 단위로 나누어 주제를 탐지해야 한다. 본 논문에서는 오픈 소스 소프트웨어인 한나눔 한국어 형태소 분석기를 사용하여 체언만 추출하였다[1].

의미 없는 단어가 주제 탐지에 사용되는 것을 막기 위해 추출한 체언 중 문서 빈도(document frequency)가 가장 높은 30개 단어를 불용어로 제거하였다.

2.4 실험 데이터베이스

뉴스 텍스트 데이터베이스는 미디어다음에서 제공하는 뉴스 전사 문서로 구성되어 있고, 형태소 분석과 불용어 제거 과정을 거쳐 13,004가지 체언, 88,705개 단어로 구성되었다. N-best 데이터베이스는 뉴스 동영상에서 음성 부분을 추출하여 짧은 음성 샘플로 나눈 뒤 음성 인식을 하여 N-best 결과를 추출하고 형태소 분석과 불용어 제거 과정을 거쳤다. 그 결과 1-best 데이터베이스는 14,566가지 체언, 71,494개 단어, 3-best 데이터베이스는 18,088가지 체언, 213,568개 단어, 5-best 데이터베이스는 20,487가지 체언, 477,653개 단어로 구성되었다.

3. 실험 및 결과

3.1 주제 탐지 실험

추출한 단어들을 사용하여 주제를 탐지하였다. 주제 탐지에 사용한 방법은 토픽 모델링 기법중 하나인 LDA(Latent Dirichlet Allocation)이다. LDA는 각 문서를 주제의 혼합체로 표현하고, 각 주제를 단어의 분포로 표현하며, EM 알고리즘으로 식 (1)을 추정한다.

$$q(\theta, z|\gamma,\phi) = q(\theta|\gamma)\prod_{n=1}^{N} q(z_n|\phi_n) \quad (1)$$

z는 주제, θ는 주제의 가중치, γ는 θ의 Dirichlet 분포 파라미터, ϕ는 z의 다항분포 파라미터를 말한다. E-단계에서 γ와 ϕ를 구하고, M-단계에서 z와 θ를 구한다[2]. 데이터베이스 가운데 856개 뉴스를 학습 데이터로 사용하여 주제를 탐지하고, 200개 뉴스를 테스트 데이터로 사용하였다. EM 알고리즘을 1,000회 반복하고, 주제의 수는 10개, 주제마다 20개의 단어를 탐지하였다. 실험에는 오픈 소스 소프트웨어인 MALLET을 사용했다[3]. 탐지한 주제와 테스트 데이터 사이의 유사도를 계산하였다. 유사도가 가장 높은 주제 단어와 테스트 데이터의 뉴스 제목을 비교하여 같은 단어가 1개 이상 나오거나 비슷한 의미의 단어가 3개 이상 나오면 정확히 인식한 것으로 사람이 체크하여, 주제 인식 정확도를 계산하였다.

3.2 실험 결과

표 1은 학습 데이터 가운데 날씨 관련 뉴스가 가장 많이 포함된 주제로 탐지한 20개 단어 중에 10개를 보여준다. 실험 결과 1-best 데이터베이스의 주제 인식 정확도는 51%, 3-best 데이터베이스는 64.5%, 5-best 데이터베이스는 56.5%, 뉴스 텍스트 데이터베이스는 69.5%가 나왔다.

표 1. 날씨 관련 주제 단어

1-best	3-best	5-best	뉴스 텍스트
어제, 비, 도, 서울, 오후, 지방, 전국, 아침, 날씨, 대구	비, 어제, 도, 오후, 지방, 서울, 전국, 크, 날씨, 아침	비, 어제, 도, 지방, 전국, 오후, 아침, 서울, 크, 날씨	비, 도, 어제, 기온, 지방, 곳, 전국, 날씨, 오후, 아침

4. 결론

본 연구에서는 1종류의 텍스트 문서와 잡음이 있는 한국어 음성 데이터의 3종류 N-best 음성 인식 결과를 사용하여 주제를 탐지한 결과를 비교했다.

우선 텍스트 문서와 음성 인식 결과의 인식 정확도를 비교해 보면, 잡음의 영향으로 음성 인식률이 낮아 음성 인식 결과의 주제 인식 정확도가 낮은 것을 확인할 수 있었다. 또한, 30%라는 낮은 음성 인식률에 비해 음성 인식 결과의 주제 인식 정확도가 많이 떨어지지 않았다는 것도 확인할 수 있다.

N-best 음성 인식 결과의 주제 인식 정확도를 비교해 보면, 1-best 결과를 사용했을 때보다 3-best나 5-best 결과를 사용했을 때가 주제 인식 정확도가 높다는 것을 확인할 수 있다. 이것은 N-best 음성 인식 결과를 사용하면 중요 단어의 출현 빈도가 높아져 주제 탐지에 도움을 줄 수 있다는 것을 보여준다.

향후 연구를 통해 음성 인식 결과로 주제를 탐지하여도 텍스트 문서로 주제를 탐지한 것과 유사한 결과를 얻고자 한다.

참고문헌

[1] Park, S., Choi, D., Kim, E., and Choi, K.-S. (2010). A plug-in component-based korean morphological analyzer. *In Proceedings of HCLT2010*, pages 197 - 201.

[2] Blei, D.M., Ng, A.Y., & Jordan, M.I, (2003). Latent dirichlet allocation, *The Journal of Machine Learning Research*, Vol. 3, 993-1022,

[3] McCallum, & Kachites, A. (2002). MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu.