JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 25, No. 2: 308~313, Feb. 2021

트윗의 타임 시퀀스를 활용한 DTM 분석 : 2019 남북미정상회동 이벤트를 중심으로

고은지[1] · 최선영[2*]

Tweets analysis using a Dynamic Topic Modeling : Focusing on the 2019 Koreas-US DMZ Summit

EunJi Ko[1] · SunYoung Choi[2*]

[1]Ph.D, Division of Digital Media, Ewha Womans University, Seoul, 03760 Korea
[2*]Visiting Professor, Graduate School of Communication & Arts, YonSei University, Seoul, 03722 Korea

요 약

이 연구는 2019년 판문점 남북미 정상 회동 트윗을 타임 시퀀스와 함께 수집하여 시퀀셜 토픽모델링인 DTM으로 분석하였다. 트위터와 같은 마이크로 블로깅 서비스는 단일 이벤트에 뉴스와 오피니언이 혼재된 비정형 데이터가 대규모로 동시에 발생하고 , 정보와 반응이 동일 메시지 형식으로 생산된다. 때문에 토픽 트렌드를 파악하려면 시퀀셜 데이터의 특성을 반영하여 패턴 분석을 해야 맥락적 의미를 알 수 있다. 토픽 일관성 점수를 구해 LDA를 평가한 후 DTM을 계산한 결과, 뉴스 보도와 오피니언 관련 토픽 30개가 도출되었고, 각 토픽과 키워드는 시간에 따라 발생 확률이 역동적으로 진화하고 있었다. 결론적으로 DTM은 특정 이벤트에 대한 사회 전반에 나타난 통합적 토픽 추이를 시간에 따라 분석하는데 적합한 모델임을 밝혔다.

ABSTRACT

In this study, tweets about the 2019 Koreas-US DMZ Summit were collected along with a time sequence and analyzed by a sequential topic modeling method, Dynamic Topic Modeling(DTM). In microblogging services such as Twitter, unstructured data that mixes news and an opinion about a single event occurs at the same time on a large scale, and information and reactions are produced in the same message format. Therefore, to grasp a topic trend, the contextual meaning can be found only by performing pattern analysis reflecting the characteristics of sequential data. As a result of calculating the DTM after obtaining the topic coherence score and evaluating the Latent Dirichlet Allocation(LDA), 30 topics related to news reports and opinions were derived, and the probability of occurrence of each topic and keywords were dynamically evolving. In conclusion, the study found that DTM is a suitable model for analyzing the trend of integrated topics in a specific event over time.

키워드 : DTM, LDA, 토픽 일관성, 트윗, 2019 남북미정상회동

Keywords : Dynamic topic modeling, LDA, Topic coherence, Tweets, 2019 Koreas – US DMZ summit

Received 8 January 2021, **Revised** 12 January 2021, **Accepted** 16 January 2021

*** Corresponding Author** SunYoung Choi(E-mail:pighairs@gmail.com)
Visiting Professor, Graduate School of Communication & Arts, YonSei University, Seoul, 03722 Korea.

Open Access **http://doi.org/10.6109/jkiice.2021.25.2.308** print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

소셜미디어를 통해 쉽게 실시간 메시지를 생산 및 공유하는 일이 흔해지고 비정형 데이터가 대규모로 발생하면서 정보 확산의 맥락과 흐름을 포착하는 일은 점점 더 중요해지고 있다. 시간에 따라 변화하는 특정 토픽이 여론의 향방을 이끌고 다양한 이해당사자들의 의견 결정에 사회적 영향을 주기 때문이다. 특히 트위터와 같은 마이크로 블로깅(Micro Blogging)은 짧은 문장 몇 개로 뉴스와 정보를 전달할 수 있을 뿐 아니라 오피니언(opinion)도 동일 형식의 메시지 형식으로 생산할 수 있는[1], 혼재된 정보 또는 메시지라고 할 수 있다. 현재 저널리즘 및 미디어 분야에서 컴퓨테이셔널 방법을 통해 데이터를 연구할 때 뉴스나 댓글, 오피니언을 분리해 데이터마이닝하여 각각의 연구를 정태적(static)으로 분석하는 경향이 있는데, 이 경우 특정 이벤트가 시시각각 변화했을 경우 정보와 오피니언의 상호반응에 의한 사회적 주요 여론을 통합적으로 해석하기 어렵다. 언론기관이 정보 수단을 독점했던 과거와 달리 누구나 분, 초 단위의 정보를 생산해 공유하는 것이 가능해지면서 일방향 정보의 흐름보다 실시간 오피니언과 상호작용에 의해 생성된 사회적 여론 파악도 중요해졌다. 특히 비정형 데이터 메시지의 생성 규모가 커졌고 을 공유와 확산 속도도 빨라지면서 통합적 분석이 긴요해진 것이다.

특히 그림 1의 2019년 6월 30일 남북미 정상 판문점 회동처럼 예측 불허의 이벤트인 경우가 그렇다.

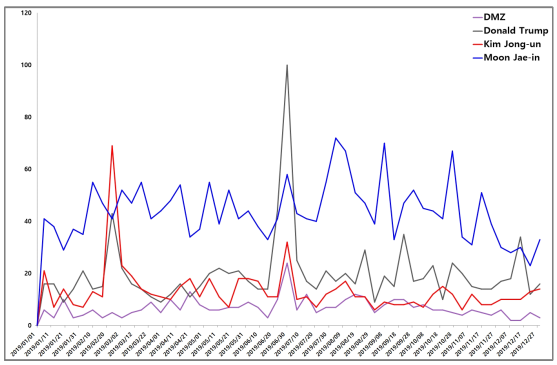

Fig. 1 2019 Koreas-United States DMZ Summit Keywords (Google Trends)

이날 하루 동안 이른바 '트윗(tweets)정치'에 의해 각본 없는 생중계가 이어지면서 뉴스와 소셜 오피니언이 동시에 다량으로 생산됨에 따라 이 이벤트에 다양한 토픽이 발생했을 가능성이 있다. 주요 토픽이 시간에 따라 어떻게 진화(evolution)했는지 확률적 모델링을 통해 비정형 데이터에 내재된 패턴을 분석하면, 정보와 오피니언의 의미적 연결을 도출할 수 있을 것이다.

이 연구의 목적은 정보, 뉴스, 오피니언 등이 혼재된 복합적인 속성의 시퀀셜 비정형 데이터에서 토픽을 추출해 확률적 모델을 계산해 추이를 분석하는 것이다. 이를 통해 마이크로 블로깅, 소셜미디어 메시지 분석에 유용한 방법임을 제안하고자 한다.

Ⅱ. 확률적 토픽 모델

2.1. LDA와 DTM

토픽 모델링은 확률 모델을 통해 문서 집합 내에서 주제(topic)을 찾아 맥락적으로 엮인 단어(Keywords)의 패턴을 알고리즘으로 추론하는 기술이다. LDA(Latent Dirichlet Allocation)는 각 문서(Document)를 발생 확률의 비율로 구성된 토픽의 집합으로 가정하는 기본적인 토픽 모델링으로 토픽은 확률로 계산된 여러 단어로 구성된다[2-4]. LDA는 시간 추이나 트렌드를 반영하지 않는 정태적 문서 집합의 토픽을 확률적으로 추론할 때 흔히 활용한다.

그러나 방법론으로서 토픽 모델링은 데이터의 속성을 고려해 데이터 처리 및 모델의 정확성을 높이는 것이 중요하다[3]. 토픽 발생 확률이 시간대마다 달라질 가능성이 높은 시퀀셜 데이터는 맥락적 트렌드를 분석해야 추론의 정확도를 높일 수 있다. 타임 스탬프가 있는 연속 문서 집합은 시퀀셜(sequential) LDA인 다이내믹 토픽 모델(DTM: Dynamic Topic Model)을 통해 도큐먼트의 형태소에서 '주제의 진화 추이(the time evolution of topics)'를 확률로 구할 수 있다[5]. DTM을 생성하려면 먼저 LDA를 구하고, 이 모델이 실제 관측 값을 제대로 추론하는지 평가 과정을 거쳐야 한다. 문서 내 토픽 출현 확률, 토픽 내에 용어 출현 확률 계산에 있어서 LDA가 일관성 있게 잘 모델링 되었는지 판단해야 하는 것이다[6].

2.2. 토픽 일관성(topic coherence)

확률적 토픽 모델로서 LDA가 도큐먼트의 토픽을 잘 추론했는지, 토픽을 구성하는 키워드 출현 확률이 적절

한지 확인하기 위해 모델링 평가 과정을 거친다. 이는 분석하고자 하는 문서 집합에서 주제 개수(number of topics)를 결정하는 것으로 모델링이 잘 되었을 때 적절한 개수의 토픽이 추출된다.

이 과정에서 연구자는 토픽 모델링 알고리즘에 토픽 개수를 임의적으로 입력해 적절한 수준의 토픽 개수가 무엇인지 반복적으로 확인해야 한다[3]. 알고리즘이 잠재적으로 분포된 토픽과 토픽을 구성하는 단어의 분포를 계산한 결과 값이 토픽의 질(Quality)을 결정하기 때문이다. LDA 모델을 통해 계산된 토픽의 확률 분포가 예측을 잘 해야 DTM도 좋은 결과가 나온다.

이를 위해 토픽의 발생 확률과 토픽의 키워드가 확률적으로 잘 분포하는지 토픽 일관성(topic coherence) 계산으로 점수(score)를 구하는데, 토픽의 상위 키워드가 토픽 내에서 맥락적으로 의미있는지 유사도 계산을 통해 DTM에 필요한 LDA 모델을 결정한다[7,8].

Ⅲ. 분석 방법

3.1. 데이터 수집

본 연구에서 활용한 데이터는 트위터 메시지로 속보 및 정보, 의견 메시지 등을 누구나 '트윗'해 생산 · 공유할 수 있는 특징이 있고, 이벤트 중요도에 따라 실시간 메시지 생산량 변동이 크다. 구체적으로 '2019 판문점 남북미정상 회동' 관련 트윗을 분석 데이터로 선정했는데, DTM 분석에 맞는 타임 스탬프가 기록되는 시퀀셜 데이터이기 때문이다. 이벤트 특성상 시간에 따라 실시간 뉴스와 정보, 의견 등이 맥락적으로 상호작용하면서 토픽 진화가 일어났을 가능성도 높고 단일 이벤트라 토픽 일관성 계산도 잘 도출될 것이라 가정했다.

데이터 수집은 트위터에서 2019년 6월 30일에 발생한 트윗 중 대한민국에서 발생한 한글 트윗에 대해, 7월 5일 하루 동안 해당 날짜의 데이터를 파이선(Python) 트위피(tweepy) 라이브러리를 사용하여 수집하였다. 데이터 수집 조건 키워드로 '북미정상회담', '트럼프', '김정은', '문재인', '판문점', '번개' 등을 지정하였다.

3.2. 연구설계

데이터 처리는 그림 2와 같은 과정을 통해 수행하였다. LDA와 DTM을 계산하기 위해 파이선 3 라이브러리 gensim(https://radimrehurek.com/gensim)을 활용하였고, 트윗 1개를 하나의 도큐먼트로 설정하였다.

그 결과 총 15,043개의 'documents', 151,512개의 'non-zero entries', 14,544개의 중첩되지 않은 형태소 'features'가 생성되었다. 이러한 데이터를 이용하여 토픽 개수를 달리하여 LDA 모델을 반복 계산하여 확정하였고, 각 LDA 모델에 3시간 단위의 타임 시퀀스를 적용하여 DTM을 계산하였다.

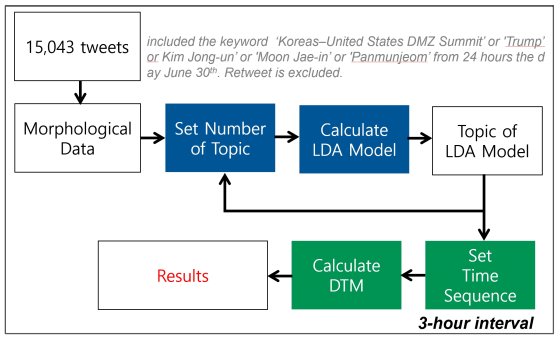

Fig. 2 Data Processing

3.3. LDA 토픽 일관성과 모델 평가

모델 평가를 위해 토픽 개수를 15~50개까지 설정한 후 각각의 LDA 모델의 일관성 점수를 계산하였다[9]. LDA는 트윗 1개를 하나의 문서 단위로 설정하여 gensim 라이브러리를 사용해 계산하였다. 그 결과, 그림 3과 같이 25~30개가 적절한 토픽개수로 나타났다.

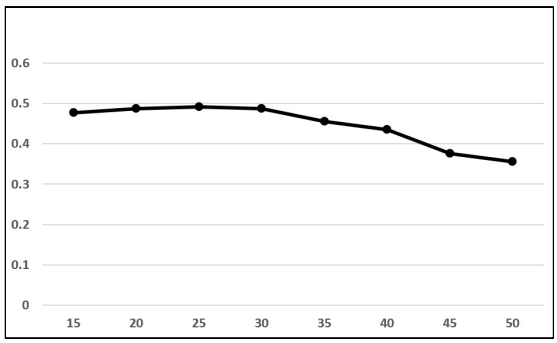

Fig. 3 Topic Coherence Score

Ⅳ. 연구결과

4.1. DTM 확정 및 시각화

토픽 일관성 점수를 토대로 토픽 개수를 결정하기 위해 DTM 계산 후 pyLDAvis로 확인한 결과 그림 4,5와 같이 나타났다. 이 데이터에 있는 잠재적 토픽의 개수는 25개보다 30개의 토픽이 데이터 집합의 확률적 토픽 분포를 더 잘 추정한 것으로 해석할 수 있다.

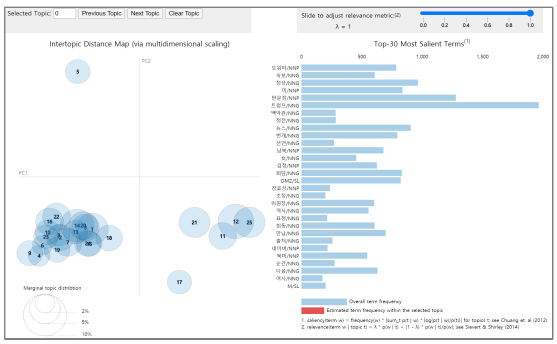

Fig. 4 DTM of 25 topics' pyLDAvis

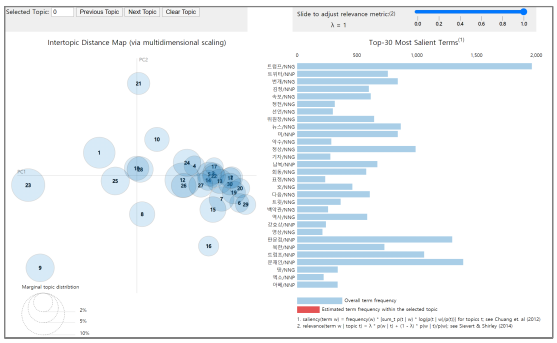

Fig. 5 DTM of 30 topics' pyLDAvis

30개의 토픽 중 확률적으로 잠재 변수로 추론한 주요 토픽과 토픽을 구성하는 키워드는 표 1과 같다. 주요 토픽은 'DMZ 회동', '트위터 정치', '급번개', '백악관 초청', '종전선언기대', '뉴스에 대한 코멘트' 등으로 분포된 것을 볼 수 있다. 주요 이벤트에 대한 토픽 분포가 뉴스 보도 뿐 아니라 보도에 대한 의견이 별도 토픽으로 도출된 점에서 키워드를 통해 의미 연결을 추론할 수 있는 것이다.

Table. 1 Topics with a high probability ratio

Topic(number)	hangul Keywords
DMZ meeting(23)	미국, 판문점, 뉴스, 문, 속보, 남북, 정상회담, 다음, DMZ, 회동, 역사적
Twitter Politics(17)	트럼프, 사랑, 트위터, 트친, 대통령, 김정은, 토착, 왜구, 통역, DMZ, 위험
Instant Meetup(22)	트위터, 여사, 칭찬, 감사, 계정, 번개, 김정은, 트럼프, 문재인, 대통령
White House invitation(1)	위원장, 김정은, 배악관, 대통령, 초청, 사진, 방문, 트럼프, DMZ, 미국, 속보
Armistice negotiation expectations(25)	선언, 정전, 정상, 북미, 판문점, 평화, 대통령, 역사, 종전, 남북
Comments on the news(3)	문재인, 대통령, 기레기, 연합뉴스, 개, 자유한국당, 남측, 김정은, 트럼프,

4.2. 시간 추이에 따른 DTM pyLDAvis 비교 분석

DTM 결과 토픽과 토픽을 구성하는 키워드가 시간 경과에 따라 어떻게 진화하는지 높은 발생 확률의 토픽인 'DMZ 회동'(23번)을 pyLDAvis를 통해 비교한 결과 그림 6,7,8과 같이 나타났다.

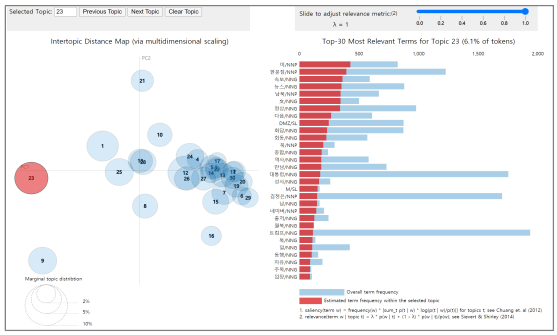

Fig. 6 pyLDAvis of 23th topic(06:00~09:00)

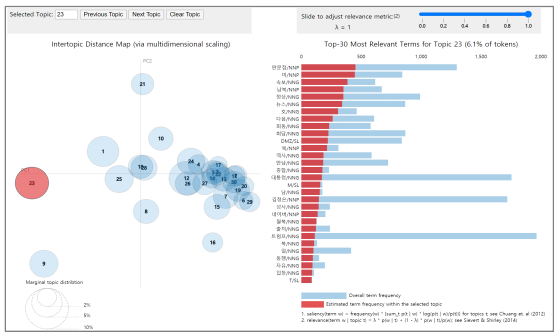

Fig. 7 pyLDAvis of 23th topic(12:00~15:00)

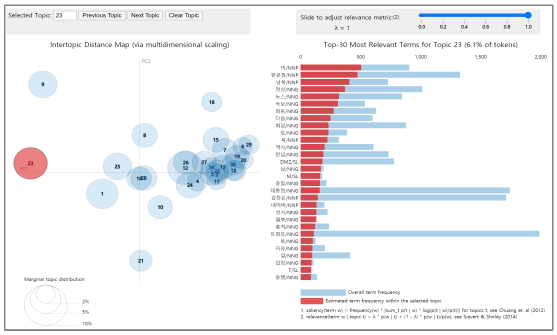

Fig. 8 pyLDAvis of 23th topic(21:00~24:00)

그림 6, 7에서는 '속보', '뉴스' 등이 높은 확률로 나타났지만, 그림 8에서는 이보다 '남북', '정상'과 같은 키워드가 높은 확률로 발생했는데 이는 DTM이 토픽 진화 다이내믹을 제대로 추론했음을 보여주는 것이다.

4.3. DTM 주요 토픽 키워드 발생 확률 분석 결과

토픽 키워드 진화 결과를 확인하기 위해 4개의 주요 토픽을 선정해 분석한 결과는 그림 9~12와 같다.

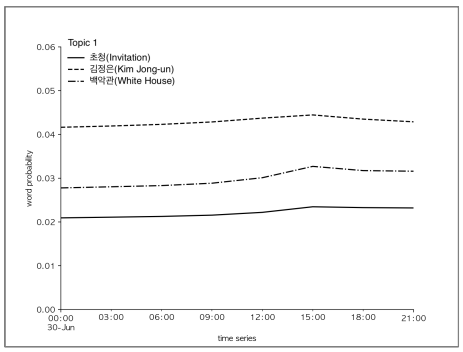

Fig. 9 topic 1

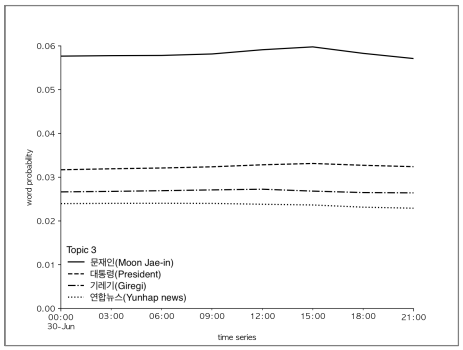

Fig. 10 topic 3

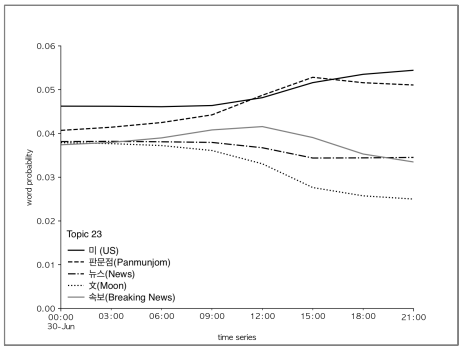

Fig. 11 topic 23

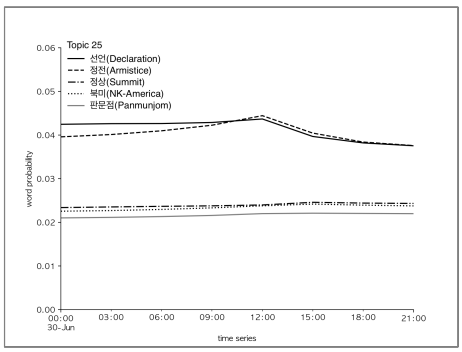

Fig. 12 topic 25

그림 9는 혼란스럽게 미확인 기대감을 높인 토픽이었으나, 관련 언급이 외신을 통해 알려지자 키워드 발생 확률이 다소 높아졌음을 보여준다. 그림 10은 언론보도에 대한 의견 평가로, 특정 키워드가 높은 확률을 유지하고 있다. 그림11은 가장 높은 발생확률의 토픽 23으로 '미국', '판문점'은 시간에 따라 발생확률이 높아진 반면 '뉴스', '속보', '문'의 낮아지고 있는데 이는 그림 6~8의 결과와도 일치한다. 그림 12는 '종전선언기대'라는 토픽으로 종전 선언에 대한 오전의 추측이 실제 회동 후에 발생확률이 낮아졌음을 알 수 있다.

V. 결 론

이 연구는 트윗과 같이 뉴스나 오피니언이 혼재된 비정형 시퀀셜 데이터에서 토픽을 추출해 분석할 때 일반적인 토픽 모델링보다 DTM이 유용함을 밝혔다.

이 연구의 의의는 첫째, 뉴스 보도와 소셜 오피니언을 분리하지 않고 토픽 추론 방법론을 제안한 점이다. 둘

째, DTM을 통해 토픽 내에서 주요 키워드 발생 확률의 진화를 확인하였고 셋째, 모델 평가를 일관성 점수 계산과 시각화를 통해 밝혔다.

단일 사례 연구로서 특정 이벤트 데이터이고 토픽이 예측된다는 점에서 한계는 있지만, 컴퓨테이셔널 방법론으로서 DTM이 사회과학적 융합연구와 후속연구로 발전하기를 기대한다.

ACKNOWLEDGEMENT

This research was supported by KABS(Korean Association for Broadcasting & Telecommunication Studies)as an excellent researcher program in 2018.

REFERENCES

[1] S. A. A. Hridoy, M. T. Ekram, M. S. Islam, F. Ahmed, and R. M. Rahman, “Localized twitter opinion mining using sentiment analysis,” *Decision Analytics*, vol. 2, no. 1, pp. 1-19, 2015.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *Advances in neural information processing systems*, vol. 14, pp. 601-608, 2001.

[3] S. Y. Choi and E. J. Ko, “Analysis of 〈Korean Journal of Journalism & Communication Studies〉 from 1960 to 2018 using Metadata with Dynamic Topic Modeling,” *Korean Journal of Journalism & Communication Studies*, vol. 63, no. 4, pp. 7-42, Aug. 2019.

[4] S. Y. Choi and E. J. Ko, “Real-time Participative Democracy through Media Multitasking and Online Community Gamification - Analysis on the Online Posts Using a Dynamic Topic Model,” *Korean Journal of Broadcasting and Telecommunication Studies*, vol. 31, no. 3, pp. 78-113, May. 2017.

[5] D. M. Blei and J. D. Lafferty, “Dynamic topic models,” in *Proceedings of the 23rd international conference on Machine learning*, ACM, pp. 113-120. 2006.

[6] D. M. Blei, “Probabilistic Topic Models(review article),” *Communications of the ACM*, vol. 55, no. 4, pp. 77-84. 2012.

[7] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, “ Automatic evaluation of topic coherence,” in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 100-108, 2010.

[8] D. O’callaghan, D. Greene, J. Carthy, and P. Cunningham, “An analysis of the coherence of descriptors in topic modeling,” *Expert Systems with Applications*, vol. 42, no. 13, pp. 5645-5657, Aug. 2015.

[9] F. Morstatter and H. Liu, “In search of coherence and consensus: measuring the interpretability of statistical topics,” *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6177-6208, 2017.

고은지(EunJi Ko)

이화여자대학교 디지털미디어학부 미디어공학 석사
이화여자대학교 공학박사
연세대학교 정보대학원 인지공학스퀘어 박사 후 연구원
※관심분야 : 피지컬 컴퓨팅, 컴퓨테이셔널 방법론, 아이트래킹

최선영(SunYoung Choi)

이화여자대학교 영상미디어박사
나노비전 대표 역임
이화여자대학교 에코크리에이티브 특임교수 역임
※관심분야 : 컴퓨테이셔널 방법론, OTT서비스, 소셜미디어, 비주얼커뮤니케이션