Journal of the Korean Institute of Industrial Engineers
Vol. 48, No. 2, pp. 235-248, April 2022.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2022.48.2.235

KoBERTSEG: 한국어 BERT를 이용한 Local Context 기반 주제 분리 방법론

소규성 · 이윤승 · 정의석 · 강필성†

고려대학교 산업경영공학부

KoBERTSEG: Local Context Based Topic Segmentation Using KoBERT

Kyoosung So · Yunseung Lee · Euisuk Chung · Pilsung Kang

Department of Industrial & Management Engineering, Korea University

Topic segmentation refers to the work of separating a document consisting of several topics into unit documents, such as paragraphs, with one single topic. Topic segmentation has been considered as one of main preprocessing step prior to performing natural language processing tasks, such as document summary or document classification. This paper proposes a Korean BERT-based news article segmentation method aiming at separating a single news article, in which multiple subjects exist, into news segments, each of which contains a single subject. The proposed model has the advantage of being able to capture a wider range of semantic relationships compared to existing topic segmentation studies by borrowing a structure proposed for document summarization. Experimental results on a Korean news article dataset show that the proposed method outperform the benchmark models for topic segmentation. In addition, we also show that the proposed method can be used for practical news clip summarization task, supporting the possibility of implementing the application service based on Korean topic segmentation model.

Keywords: Natural Language Processing, Topic Segmentation, Text Segmentation, KoBERT, BERTSUM

1. 서 론

일반적으로 문서는 일련의 주제가 유기적으로 연결된 형태를 띠며, 이를 하나의 통일된 주제의 단위 문서로 분리하는 과업을 글자 분리(Text Segmentation), 또는 주제 분리(Topic Segmentation)라고 한다. 주제 분리는 오랫동안 연구된 자연어처리 분야 중 하나로서, 관련된 널리 알려진 방법론으로는 Text Tiling(Hearst, 1997) 알고리즘, C99 알고리즘(Choi, 2000) 등이 있다. 주제 분리는 정보 검색(Information Retrieval), 문서 분류, 문서 요약 등의 후속 과업(Down-stream Task)를 수행함에 있어 활용 가능성이 다분한 실용적인 방법론으로, Solbiati *et al.*(2021)와 같이 회의록에 대한 주제 분리를 통해 이용자들에게 편의를 제공하고자 하는 연구 또한 존재한다.

과거의 주제 분리 연구는 대용량의 레이블링 된 데이터가 부재함에 따라 대부분 비지도 학습을 기반으로 수행되었다. 대표적으로 Text Tiling은 단위 블록 간 어휘 유사도 등을 이용해 측정한 점수를 바탕으로 주제 분리를 수행하였다. Alemi *et al.*(2015)의 경우 Word2Vec, GloVe 등 단어 임베딩에 기반해 문장 단위 표상을 얻어 이를 기반으로 주제 변화 점수를 계산하였다. 이 외에 LDA와 같은 토픽 모델링 기법을 통해 얻은 표

이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.2021-0-00034, 파편화된 데이터의 적극 활용을 위한 시계열 기반 통합 플랫폼 기술 개발)
† 연락저자 : 강필성 교수, 02841, 서울시 성북구 안암로 145 고려대학교 산업경영공학부, Tel: 02-3290-3383, Fax: 02-929-5888, E-mail: pilsung_kang@korea.ac.kr
2021년 11월 17일 접수; 2021년 12월 12일 수정본 접수; 2022년 2월 8일 게재 확정.

상을 기반으로 주제 분리를 수행하는 방법론(Misra *et al.*, 2011; Riedl *et al.*, 2012)도 제안되었다.

보다 최근의 주제 분리 연구는 Wiki-727K(Koshorek *et al.*, 2018)와 같이 주제 분리 과업을 위한 대용량 데이터셋이 등장하면서 신경망 기반의 지도 학습 방법론을 제안하였다. Koshorek *et al.*(2018)은 LSTM에 기반한 주제 분리 모델을 제안하였고, Badjatiya *et al.*(2018)은 합성곱 신경망(Convolutional Neural Network, CNN)과 어텐션 메커니즘을 활용해 주제 분리 여부를 예측하는 분류 모델을 사용하여 주제 분리를 수행하였다. 특히 Bidirectional Encoder Representations from Transformers (BERT) 등 사전 학습 언어 모델이 등장하면서 보다 풍부한 의미론적 임베딩을 활용하는 방법론들 또한 등장하였다(Aumiller *et al.*, 2021; Iikura *et al.*, 2020; Jeon *et al.*, 2019; Pethe *et al.*, 2020).

그러나 어휘 빈도나 단어 단위의 임베딩을 기반으로 하는 방법론은 문장들의 의미론적 관계를 제대로 반영하기 어렵다는 한계점이 있고, LSTM 기반의 방법론은 많은 문장이 입력되는 경우 기울기 소실과 같은 문제가 발생해 정보를 제대로 반영하기 어렵다. 또한 Jeon *et al.*(2019)와 Pethe *et al.*(2020)의 경우 BERT를 이용한 지역적 문맥(Local Context) 기반의 주제 분리를 시도하였으나, 한 쌍의 문장만을 입력 값으로 사용하여 보다 넓은 범위에서 문맥의 흐름을 반영하지 못한다는 한계점을 갖는다. 두 문장 단위의 임베딩만 사용하는 경우 새로운 어휘의 등장과 같은 노이즈가 발생할 때 동일한 문맥임에도 분리를 수행하는 1종 오류(False Positive)의 발생 확률이 높으며 이는 후속 과업으로 문서 요약 등을 수행하는 경우 성능 저하의 큰 원인이 될 수 있다.

주제 분리에 대한 방법론적인 측면이 아닌 언어적 측면을 보면 Choi(Choi, 2000), Wiki727과 같은 데이터셋에 기반한 영어 주제 분리 연구는 다수 진행되었으나 한국어 데이터를 이용한 주제 분리 연구는 Jeon *et al.* (2019) 외에는 매우 미약한 실정이다. 대량의 한국어 위키 데이터를 이용해 학습된 한국어 BERT(KoBERT) 등 한국어 자연어처리 과업을 수행하기 위한 풍부한 자원이 등장하였음에도 한국어 주제 분리 연구가 거의 수행된 바가 없다는 점은 다소 아쉬운 부분이라고 할 수 있다.

본 연구는 이러한 한계점을 보완하고자 문서 요약 모델인 BERTSUM(Liu *et al.*, 2019)의 구조를 차용한 KoBERT 기반의 주제 분리 모델인 KoBERTSEG를 제안한다. 소설의 챕터를 분리(Pethe *et al.*, 2020)하거나 대화문을 분리(Solbiati *et al.*, 2021)하는 기존 연구들과 달리, 본 연구는 다양한 사건을 다루는 긴 기사를 하나의 사건만을 다루는 단위 기사문으로 분할하는 것을 목적으로 한다. KoBERTSEG는 문장마다 [CLS] 토큰을 삽입하여 각 토큰이 문장을 대표하는 표상을 학습하도록 하였고, 이후 문장들에 대한 합성곱 연산을 수행하여 다수 문장들이 입력으로 사용되는 경우에도 효과적으로 지역적 문맥 기반의 주제 분리를 수행할 수 있도록 설계되었다.

제안된 모델을 학습하고 성능을 평가하기 위해서 2020년 AI 학습용 데이터 구축 사업의 일환으로 비플라이소프트 주관 컨소시엄에서 구축한 한국어 뉴스 기사 데이터셋(이하 한국어 뉴스기사 데이터셋)(https://aihub.or.kr/aidata/8054)을 활용한다. 한국어 뉴스기사 데이터셋이 주제 분리를 목적으로 하는 데이터셋이 아님에도 본 연구는 3.1에서 설명하는 간단한 방법론을 통해 레이블을 부여하여 대용량의 학습용 데이터셋을 구축하였고, 이를 통해 높은 성능의 주제 분리 모델을 구축하였다.

추가적으로 본 연구는 주제 분리 방법론이 실제 후속 과업을 수행하는 데에 있어서 갖는 효용성을 보이고자 KoBERTSEG 기반의 유튜브 뉴스 영상 요약 프레임워크를 제안한다. 이는 여러 사건 혹은 주제로 이루어진 긴 뉴스 영상의 핵심 내용을 요약해 제공하는 동시에 각 사건이 다뤄지는 영상 시점을 함께 제공하는 프레임워크이며, 이를 위해서는 전체 뉴스를 하나의 사건만 다루는 짧은 뉴스로 정확하게 분할하는 것이 필수적인 선결요소이다. 본 연구는 KoBERTSEG가 그러한 역할을 효과적으로 수행할 수 있음을 보이고 현실화 가능한 실용적인 응용서비스 프레임워크를 제안함으로써, 한국어 기사 데이터셋에 대한 주제 분리 모델이 높은 효용성을 가짐을 보이고자 한다.

본 논문은 다음과 같이 구성된다. 제2장에서는 주제 분리와 관련된 연구 동향에 대하여 살펴보고, 제3장에서는 본 논문이 제안하는 주제 분리 모델에 대하여 서술한다. 제4장에서는 한국어 기사 데이터셋을 이용한 실험 방법에 대해 자세히 서술하며, 제5장에서는 그 결과를 보이고 이에 대해 논의한다. 제6장에서는 주제 분리 모델에 기반한 유튜브 뉴스 영상 요약 프레임워크를 제안한다. 마지막 7장에서는 본 연구의 결과 및 의미에 대하여 고찰한다.

2. 관련 연구

2.1 주제 분리

주제 분리는 문서 요약, 담론 분석(Discourse Analysis), Question-Answering(QA)과 같은 여러 NLP 과업에 있어 활용성이 높은 만큼 다양한 접근 방법론이 제안되었다. 과거 비지도 학습 방법론에 기초한 주제 분리 방법론으로는 대표적으로 어휘적 응결성(Lexical Cohesion)을 이용한 연구(Eisenstein *et al.*, 2008; Hearst, 1997), 클러스터링을 이용한 연구(Kazantseva *et al.*, 2011), 그리고 토픽 모델링을 이용한 연구(Misra *et al.*, 2011; Riedl *et al.*, 2012)가 수행되었다. Hearst(1997)에서 제안된 Text Tiling은 지역적으로 정의한 블록 단위 간의 어휘 유사도 점수를 계산해 주제가 변화하는 경계면을 검출한 반면, Riedl *et al.* (2012)가 제안한 Topic Tiling은 LDA를 이용해 토픽 분포를 계산한 뒤 블록 단위 간 토픽 유사도에 기반해 경계면을 검출하였다.

주제 분리 방법론은 Koshorek *et al.*(2018)를 기점으로 지도 학습에 기반한 방법론이 다수 제안되었다. Koshorek *et al.* (2018)은 분류 관점에서의 지도 학습을 수행하기 위해 레이블링 된 데이터셋 Wiki727을 구축함과 더불어 LSTM 기반의 주

제 분리 모델을 제안하였는데, 단어 단위 LSTM 계층과 문장 단위 LSTM 계층으로 이루어진 계층적 모델을 통해 주제 변화 여부를 예측하였다. Badjatiya *et al.*2018)는 CNN과 Bi-LSTM을 이용한 임베딩에 대해 어텐션 연산을 수행함으로써 중심 문장과 주변 문맥을 포괄적으로 활용해 주제 분리 여부를 예측하는 방법론을 제안하였다.

사전 학습된 언어 모델 등장 이후 언어의 의미적인 요소를 더욱 효과적으로 이용하고자 하는 BERT 기반의 주제 분리 방법론(Aumiller *et al.*, 2021; Iikura *et al.*, 2020; Jeon *et al.*, 2019; Pethe *et al.*, 2020; Solbiati *et al.*, 2021)이 다수 제안되었다. Iikura *et al.*(2020)은 소설 내 두 문장이 동일 문단에 속하는 지 여부를 예측하는 BERT 기반 모델을 구축하였고, 추가적으로 데이터셋의 불균형을 해결하고자 Focal Loss를 활용하였다. Aumiller *et al.*(2021)은 법률 문서의 문단 분리를 수행하기 위해 문단 단위로 BERT 임베딩을 수행해 각 문단이 동일한 내용을 다루는 지 여부를 예측하였다. Jeon *et al.*(2019)은 본 연구와 유사하게 KoBERT에 기반한 문장 단위 임베딩을 활용해 칼럼 데이터셋의 문단을 분리하는 지역적 문맥 기반 주제 분리 모델을 제안하였다. 다만 Aumiller *et al.*(2021)에서 제안하는 방법론은 문단 단위 구분이 불가능한 문서에 적용하기 어렵고, Jeon *et al.* (2019)은 두 문장 단위의 임베딩만을 활용하기 때문에 텍스트 노이즈로 인한 1종 오류 발생 가능성이 높다. 본 연구는 이러한 한계점을 극복할 수 있도록 범용적으로 적용이 가능하며 텍스트 노이즈에 강건한 지역적 문맥 기반 주제 분리 모델을 제안한다.

2.2 사전 학습된 언어 모델

사전 학습된 언어 모델은 수백 기가바이트(GB), 혹은 그 이상의 대규모 말뭉치 데이터셋으로부터 언어적 구조 및 의미를 학습한 언어 모델로서, 다양한 자연어처리 과업의 성능 개선에 큰 역할을 수행하고 있다. 대표적으로, 트랜스포머(Vaswani *et al.*, 2017) 기반의 BERT의 경우 Masked Language Modeling (MLM) 및 Next Sentence Prediction(NSP)의 두가지 비지도 학습 방법론을 활용해 약 33억 개의 단어로 사전 학습되었다. MLM은 마스킹 된 토큰의 실제 토큰을 예측하도록 하는 학습 방법론이며, NSP는 두 문장을 입력하였을 때 선/후 관계를 예측하도록 하는 학습 방법론이다. 이러한 방식으로 학습된 사전 학습 언어 모델은 문서 요약과 같이 언어 이해가 요구되는 과업의 성능을 개선하는 데 주로 사용되며, 일반적으로 특정 과업을 목적으로 두어 학습하면서 가중치의 일부 혹은 전부를 미세 조정하여 활용한다.

2.3 BERTSUM

BERTSUM은 사전 학습 언어 모델 BERT를 기반으로 문서 요약 과업을 수행하기 위해 제안된 구조이다. BERT는 기본적으로 MLM를 활용해 사전 학습되기 때문에 출력 결과물이 문장 단위가 아닌 토큰 단위로 의미를 갖게 된다. 또한 문장 분류, 감성 분류 등의 후속 과업을 수행하기 위해 전체 토큰의 앞에 입력되는 [CLS] 토큰을 활용하는 것이 일반적이다. 하지만 요약 과업의 경우 문장 간의 의미적 관계에 기반해 핵심 요약을 포함하는 문장을 선택하므로, 토큰 단위가 아닌 문장 단위의 의미론적 표상이 더욱 효과적이다. 따라서 BERTSUM은 입력되는 문장마다 [CLS] 토큰을 추가하여 각 문장을 대표하는 표상을 학습하도록 하였고, 추가적으로 각 문장이 구분되도록 0과 1의 값을 갖는 Segment Embedding을 번갈아 입력하였다.

본 연구에서 제안하는 지역적 문맥 바탕의 주제 분리 방법론 또한 문장 간의 의미적인 관계를 충분히 활용하여 입력된 윈도우 내 주제 변화 여부를 예측하여야 한다. 따라서 기본적인 BERT 구조가 아닌 요약 모델인 BERTSUM의 구조를 차용해 구축된 효과적인 주제 분리 모델을 제안한다.

3. 제안 방법론

제3장에서는 KoBERTSEG를 학습하기 위한 문제 정의 및 KoBERTSEG의 구조에 대해 설명하고, 실제 여러 주제를 다루는 뉴스 기사 문서에 대해 주제 분리를 수행하기 위한 방법론을 설명한다.

3.1 문제 정의

본 연구는 뉴스 기사에 대한 주제 분리 문제를 주어진 문장들의 가운데 지점에서 주제가 분리되는지 여부를 예측하는 분류의 관점에서 접근한다. 따라서 제안하는 모델을 f, 주제 분리 여부를 판단하기 위한 입력 문서를 S라고 할 때, S를 f에 입력하여 얻게 되는 로짓값 L은 $L=f(S)$로 정의된다. 결과적으로 본 연구가 제안하는 모델은 입력되는 문서의 가운데 지점에서 주제가 분리되는 경우 L를 최대화, 반대의 경우 L을 최소화하는 방향으로 학습된다.

학습용 데이터를 구축하기 위한 레이블 부여 과정은 <Figure 1>과 같다. 우선 학습에 사용되는 뉴스 기사 데이터셋 집합을 $D=\{D_1, D_2, ..., D_N\}$이라고 정의한다. 이 때 하나의 기사에서 특정 개수의 문장을 추출해 학습 데이터셋 구축에 활용하는데, i번째 기사에서 m개의 문장을 임의 추출한 기사 문을 $D_i^{(m)}$으로 정의한다. 모델의 입력으로 활용하는 합성 기사 D_{syn_i}은 i번째 기사를 기준으로 생성된 하나의 합성 데이터를 의미하여, 다음과 같이 정의될 수 있다.

$$D_{syn_i}=[D_i^{(k_1)},\ D_{i+1}^{(k_2)}],$$
$$where\ i=1,\ 2,\ ...,\ N-1 \text{ and } k_1+k_2=2k$$

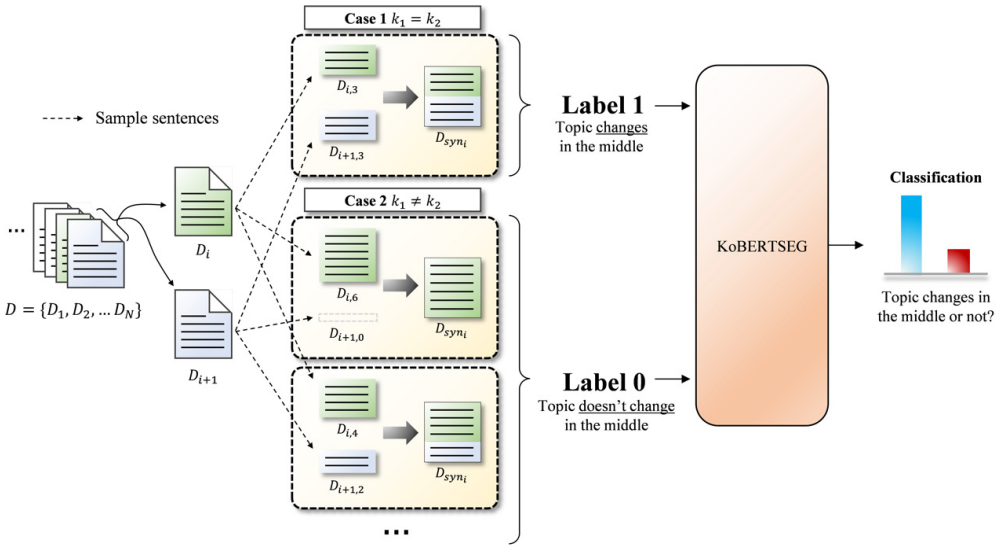

Figure 1. Problem Definition of KoBERTSEG-based Topic Segmentation

여기서 k는 제안하는 모델이 반영하는 지역적 문맥의 정도를 조정하기 위해 설정되는 윈도우의 크기이며, k의 단위는 일반적인 언어학적 의미로서의 문장이다. 위 수식에서 $k_1 = k_2$인 경우 합성된 뉴스 기사 D_{syn_i}의 중앙에서 주제가 변화한다는 1의 레이블이 부여되며, 그렇지 않은 모든 경우 0의 레이블이 부여된다. 결과적으로 KoBERTSEG는 학습 데이터셋 D_{syn_i} $(i = 1, 2, \ldots, N-1)$을 이용해 $2k$개 문장으로 이루어진 합성 기사의 가운데 지점에서의 주제 분리 확률 $p(D_{syn_i})$을 예측하도록 학습되며, 이 때 $p(D_{syn_i})$는 다음과 같이 산출된다.

$$p(D_{syn_i}) = sigmoid(f(D_{syn_i})),\ where\ p(D_{syn_i}) \in [0,1]$$

3.2 KoBERTSEG

본 연구에서 제안하는 주제 분리 모델 KoBERTSEG의 구조는 <Figure 2>와 같으며 각 계층에 대한 세부 설명은 다음과 같다. 우선 입력된 기사 문장들의 의미론적 표상을 추출하기 위

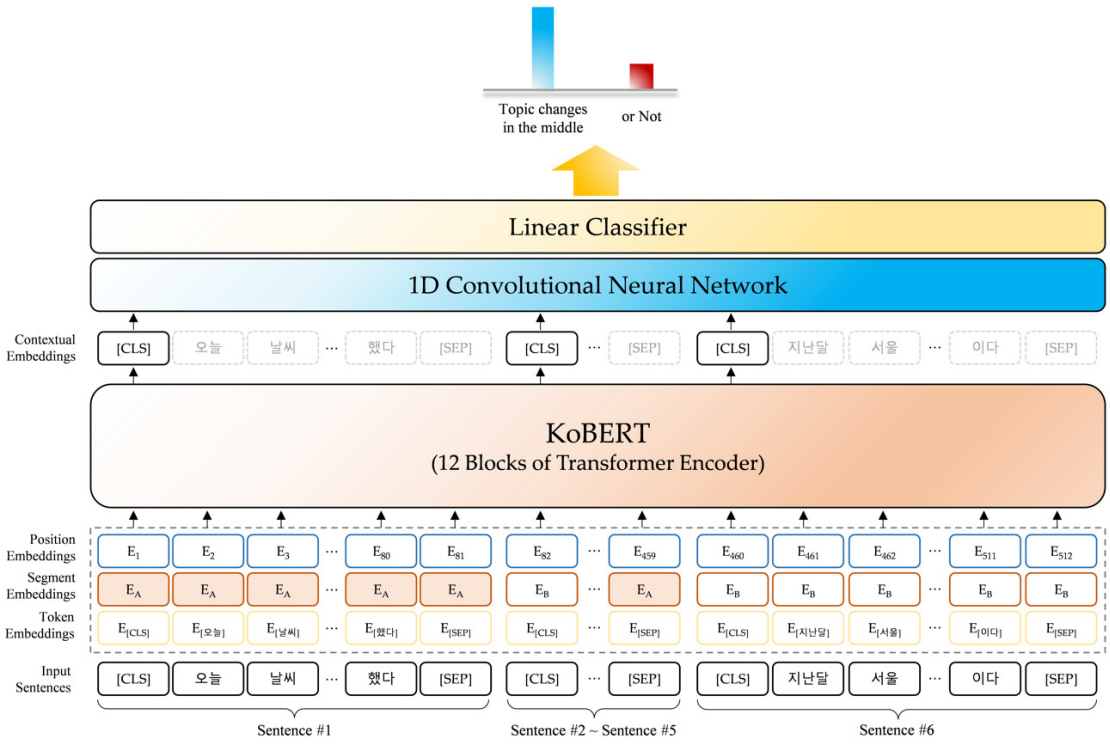

Figure 2. Structure of KoBERTSEG with Window of Size 3

해 사전 학습된 KoBERT 구조를 Backbone으로 활용한다. 다만 문장마다 [CLS] 토큰을 추가한다는 점, 그리고 문장 간 구분을 위해 세그먼트 임베딩(Segment Embedding)이 0과 1의 값을 번갈아 갖는다는 점에서 일반적인 KoBERT 구조와 차이를 갖는다. 따라서 제안하는 KoBERTSEG의 입력은 토큰 임베딩, 세그먼트 임베딩, 포지션 임베딩 세 가지 임베딩의 합으로 구성된다.

설정된 윈도우의 크기와 KoBERT의 임베딩 차원이 각각 k, d_{kobert}이고 입력되는 토큰의 총 개수가 n개라면, n개 토큰이 KoBERT를 통과해 생성되는 임베딩 행렬은 $n \times d_{kobert}$ 크기를 가진다. 이 중 문장별 [CLS] 토큰에 해당하는 $2k$개의 행만 선택하여 $2k \times d_{kobert}$ 크기의 문장 표상 행렬 C_{sent}를 얻을 수 있으며, C_{sent}는 시퀀스 데이터를 처리하기 위한 1차원 합성곱 신경망(1-Dimensional Convolutional Neural Network, 1D CNN) 계층을 통과하게 된다. 이 때 합성곱 연산을 위한 커널의 크기는 $2k-2$로 설정되며, 1D CNN 계층의 출력 차원을 d_{conv}로 정의하면 문장 표상 행렬 C_{sent}는 $3 \times d_{conv}$의 고정 크기 행렬로 변환된다. 해당 행렬은 이후 완전 연결 계층에 입력됨으로써 주어진 $2k$개 문장 가운데 지점에서의 주제 분리 확률값을 출력한다.

이 때 k의 값이 클수록 입력되는 문장 개수가 증가하므로 주제 분리에 더욱 많은 정보를 활용할 수 있다. 하지만 KoBERTSEG는 어텐션 메커니즘 기반의 트랜스포머 구조를 활용하기 때문에 토큰의 개수 증가는 연산량의 이차적인 증가로 이어지며, 따라서 과업에 맞는 적절한 k값을 설정하여야 한다.

3.3 슬라이딩 윈도우 기반 추론

학습된 KoBERTSEG 모델을 실제 주제 분리에 적용하기 위해서는 지역적 문맥 기반의 주제 분리 모델이 보편적으로 활용하는 슬라이딩 윈도우 방법론을 활용한다. 여러 주제를 갖는 기사 문서가 문장 집합 $[sent_1, sent_2, ..., sent_N]$로 주어질 때 만약 설정된 윈도우 크기가 k라면 KoBERTSEG에 입력될 하나의 합성 기사 D_{syn_i}는 다음과 같이 정의된다.

$$D_{syn_i} = [sent_i, ..., sent_{i+k-1}, sent_{i+k}, ..., sent_{i+2k-1}],$$
$$where\ i = 1, 2, ..., N-2k+1$$

결과적으로 하나의 합성 기사 D_{syn_i}를 모델에 입력하여 가운데 지점에서의 문장 분리 확률 $p(D_{syn_i}) = sigmoid(f(D_{syn_i}))$를 얻을 수 있다. <Figure 3>에 나타난 바와 같이 N개의 문장으로 이루어진 기사 문서에 대해 슬라이딩 윈도우 방식으로 KoBERTSEG를 적용함으로써 $N-2k+1$개의 주제 분리 확률 $p(D_{syn_i})(i = 1, 2, ..., N-2k+1)$을 얻게 되며, 그 중 미리 설정된 Threshold 값을 넘는 지점에서 최종적인 주제 분리가 결정된다.

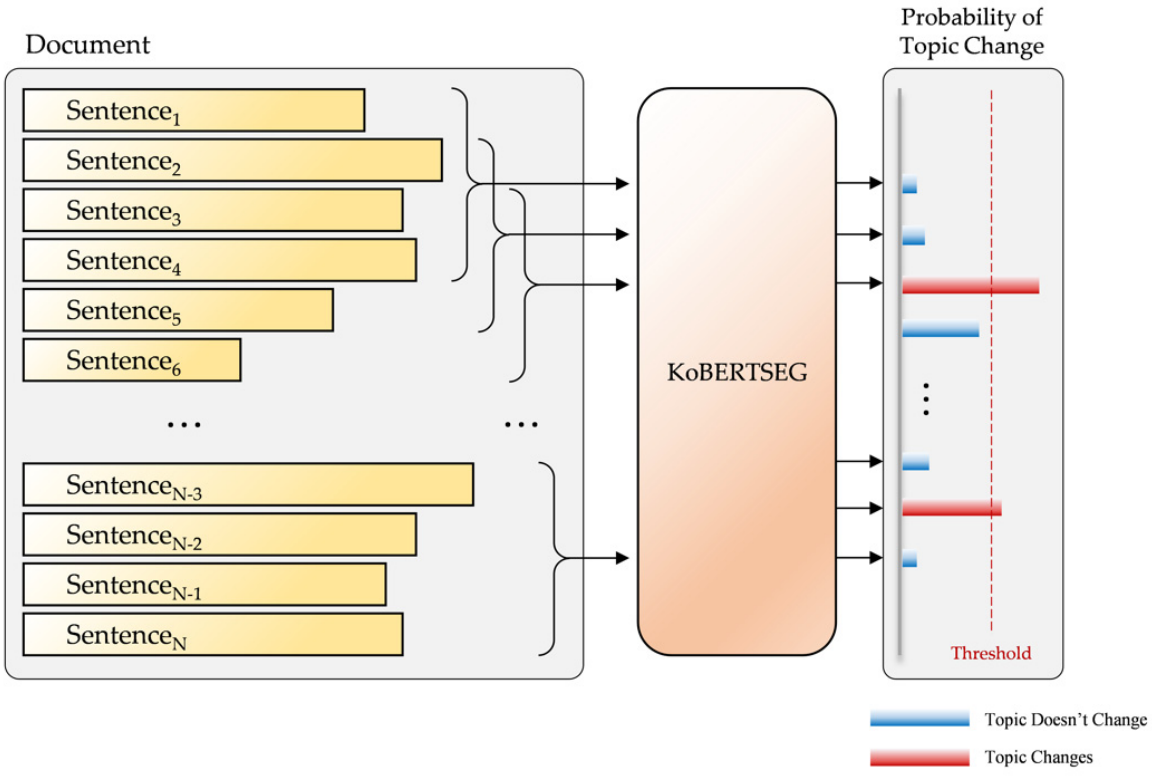

Figure 3. Illustrative Example of Inference Using KoBERTSEG with Window of Size 2

4. 실험 설계

제4장에서는 KoBERTSEG 모델을 학습 및 평가하기 위해 활용한 한국어 기사 데이터셋에 대하여 설명하고, 모델 성능 평가 방법론 및 학습 구현 방법론 등에 대해 상세히 서술한다.

4.1 데이터셋

4.1.1 학습용 데이터셋

영어의 경우 Choi(Choi, 2000) 등 주제 분리 모델의 성능을 평가하기 위한 벤치마크 데이터셋이 존재하나 한국어의 경우 표준 데이터셋이 없는 실정이다. Jeon *et al.*(2019)의 경우 한국어 기반의 주제 분리 모델을 학습 및 평가하기 위해 약 6,000개의 사설/컬럼 데이터를 직접 수집하여 사용한 바 있다.

본 연구는 한국어 기사에 대한 주제 분리 모델을 학습하기 위해 서론에서 언급된 대용량의 한국어 뉴스기사 데이터셋을 활용하였다. 한국어 뉴스기사 데이터셋은 종합, 정치, 경제, 사회, 문화 등 다양한 주제로 이루어져 있으며, 2017년 1월부터 2019년 12월 사이의 기사를 포함하고 있다. 데이터셋에 관한 기술 통계량 및 기사 예시는 <Table 1> 및 <Table 2>와 같다. 본 연구는 모델 학습 및 평가를 위해 약 26만 5천 개의 전처리를 거친 기사문을 활용하였으며, <Figure 1>에 묘사된 방식을 기반으로 하여 약 26만 개의 학습용 합성 기사 데이터셋을 생성하였다. 이 때 실제 문단 분리 상황에서의 레이블 불균형 및 학습의 안정성을 고려하여 레이블이 1인 데이터의 개수와 레이블이 0인 데이터 개수 간 비율은 3:7로 설정하였다.

4.1.2 평가용 데이터셋

주제 분리 모델의 성능을 평가하기 위해서는 하나의 주제를 갖는 여러 단위 문서들의 집합이 필요하다. 본 연구는 기사 데이터를 활용하기 때문에, 서로 다른 주제를 다루는 여러 기사가 합성되었을 때 모델로 하여금 합성 경계면을 분리 지점으로 예측하도록 하여 평가를 수행하였다.

평가용 데이터셋을 구축하기 위해서는 모델이 학습 과정에서 보지 못한 5,000개의 한국어 기사 데이터셋을 이용하였다. 최소 10개, 최대 20개 문장을 갖는 기사 5개를 통합하여 하나의 합성 기사를 생성하였고, 결과적으로 총 1,000개의 합성 기사 데이터셋을 통해 모델의 주제 분리 성능을 평가하였다.

Table 1. Descriptive Statistics of the Korean News Article Dataset

Number of Documents	Number of Documents per Category					Length of Articles	
	General	Political	Economic	Social	Etc.	Mean	Std.
280,697	197,022	18,007	27,286	19,649	18,733	12.66	5.39

Table 2. Example of the Korean News Article Dataset

Title	초음파로 악성종양·특정세포 제거… 초음파 수술 시대 열리나
Date	2019년 1월 22일
Media	서울경제
Category	Economic
Text	한·영 공동 연구진이 외과 수술 없이 종양을 치료하거나 특정 세포만 제거하는 초음파 수술법의 가능성을 제시하 눈길을 끈다. 한국과학기술연구원(KIST) 바이오닉스연구단 박기주·김형민 박사팀은 22일 영국 런던대(UCL) 기계 공학과 네이더 사파리 교수팀과 함께 초강력 초음파로 절개 없이 몸 안의 종양 등 연조직을 제거하는 메커니즘을 처음으로 규명했다고 밝혔다. 강력한 초음파 영역에서 발생하는 음향 공동현상을 예측하는 수학적 모델을 개발하고 집속초음파의 연조직 제거효과 메커니즘을 밝힌 것이다. 의학계에서는 초강력 초음파 에너지를 한 곳에 집중하 초점 부위의 조직을 제거하거나 치료하는 고강도 집속초음파(HIFU) 기술에 주목하고 있다. 외과 수술 없이 종양을 제거할 수 있어 정상조직의 부작용이 적고 회복 시간도 빨라 실제 임상 적용을 위한 다양한 연구가 진행되고 있다. 이 기술은 대기압의 수백배인 수십 메가파스칼(MPa)의 압력을 갖는 고강도 집속초음파가 1,000분의 1초 정도에 초점 부위 온도를 끓는점까지 올릴 때 초점에서 발생한 수증기 기포가 수술칼 역할을 해 주변 세포조직을 제거하는 원리다. 하지만 치료에 적용하려면 초점 부위에서 발생하는 수증기 기포의 크기와 운동 등 관련 현상을 완벽하게 설명하고 조절할 수 있어야 하나 아직 완벽하게 파악되지 않았다. 박기주(왼쪽, 제1저자·공동교신저자)·김형민(공동교신저자) 박사. 연구팀은 먼저 강력한 초음파 영역에서 발생하는 음향 공동현상의 운동변화를 예측하는 수학적 모델을 만들고 이를 이용해 세포조직의 변형률을 계산했다. 이어 겔(gel)로 」만든 인체 모사조직으로 실험하면서 이때 발생한 공동현상을 초고속카메라로 촬영했다. 그 결과 고강도 집속초음파에 의해 발생한 기포의 운동 강도는 연조직을 파괴할 수 있을 만큼 강하면서도 혈관을 파괴할 수 있는 강도보다는 약한 것으로 밝혀졌다. 박기주 KIST 박사는 "수학 모델링 기법으로 최적화된 초음파 조사조건을 찾으면 외과적 수술 없이도 종양치료와 특정 세포의 선택적 제거가 가능할 것"이라고 내다봤다. 이 연구는 국가과학기술연구회 창의형 융합연구사업과 KIST 기관고유사업 지원으로 수행됐으며 '초음파학 음향화학' 최신호에 게재됐다.

4.2 평가

4.2.1 평가 지표

주제 분리 모델의 평가 지표로는 정밀도($Precision$), 재현율($Recall$), $F-1$ $Score$와 더불어 p_k(Beeferman *et al.*, 1999)와 $WindowDiff$(Pevzner *et al.*, 2002)가 가장 널리 사용된다.

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

$$Recall = \frac{TP}{(TP+FN)} \quad (2)$$

$$F-1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3)$$

$$p_k(ref,\ hyp) = \frac{1}{N-k} \sum_{i=1,j=i+k}^{N-k} (\delta_{ref}(i,j) \oplus \delta_{hyp}(i,j)) \quad (4)$$

$$WindowDiff(ref,hyp) = \frac{1}{N-k} \sum_{i}^{N-k} (|b(ref_i,\ ref_{i+k}) - b(hyp_i,\ hyp_{i+k})| > 0) \quad (5)$$

식 (1)과 식 (2)에서 $TP(True\ Positive)$는 실제 분리 지점을 모델이 올바르게 예측한 경우의 수를 의미하며, $FP(False\ Positive)$와 $FN(False\ Negative)$는 각각 실제 분리 지점이 아닌 곳, 실제 분리 지점인 곳에 대해 모델이 예측에 실패한 경우의 수를 의미한다.

식 (4)에서 $\delta_{ref}(i,j)$와 $\delta_{hyp}(i,j)$는 각각 정답과 예측에 대해 문서의 i번째 문장과 j번째 문장이 동일한 주제에 속하는 경우 1, 반대의 경우 0의 값을 갖는다. $\oplus$는 XNOR 연산자이며, 따라서 정답과 예측의 δ 지표가 서로 다른 값을 갖는 경우 페널티가 부여된다. 결과적으로 p_k는 k 크기의 윈도우를 문서 전체에 슬라이딩하여 얻는 페널티의 합을 0과 1 사이의 값으로 표준화한 지표이다.

하지만 Pevzner *et al.*(2002)은 p_k가 1종 오류에 비해 2종 오류에 더욱 큰 페널티를 부여한다는 점, 경계면의 개수를 반영하지 못한다는 점 등을 지적하면서 이를 보완하기 위한 지표로서 식 (5)의 $WindowDiff$를 제안하였다. 식 (5)에서 $b(i,j)$는 지점 i와 지점 j 사이에 존재하는 주제 분리 경계면의 개수를 의미하며, 따라서 정답과 예측 간 동일한 윈도우 내 경계면의 개수가 다른 경우 페널티가 부여된다. p_k와 마찬가지로 k 크기의 윈도우를 슬라이딩하여 총 페널티를 계산하고, 이를 0과 1 사이의 값으로 표준화한다. $WindowDiff$와 p_k 간 페널티 계산 방식의 차이는 <Figure 4>에 자세히 묘사되어 있으며, 두 지표 모두 값이 작을수록 방법론의 성능이 우수함을 의미한다.

다만 KoBERTSEG는 윈도우의 크기 k에 의존적인 학습을 수행하며, <Figure 3>에서 볼 수 있듯이 윈도우의 크기가 k일 때 문서의 앞과 뒤 $k-1$ 크기 만큼은 분리 지점 예측이 불가능하다. 본 연구는 최소 2, 최대 5 크기의 윈도우를 이용하였기 때문에, 윈도우 크기 간 성능 비교 혹은 다른 방법론과의 형평성 있는 성능 비교를 위하여 문서의 앞, 뒤 각각 4개 지점에서는 주제 분리 예측을 수행하지 않도록 하였다. 결과적으로 모든 크기의 윈도우에 대한 KoBERTSEG 및 타 방법론은 N개 문장으로 이루어진 문서에 대해 $N-8$개로 동일한 개수의 후보 분리 지점을 가진다.

4.2.2 비교 대상 방법론

성능 비교 대상 방법론으로는 우선 Beeferman *et al.* (1999)에서 제안된 Random과 Even을 이용하였다. Random은 전체

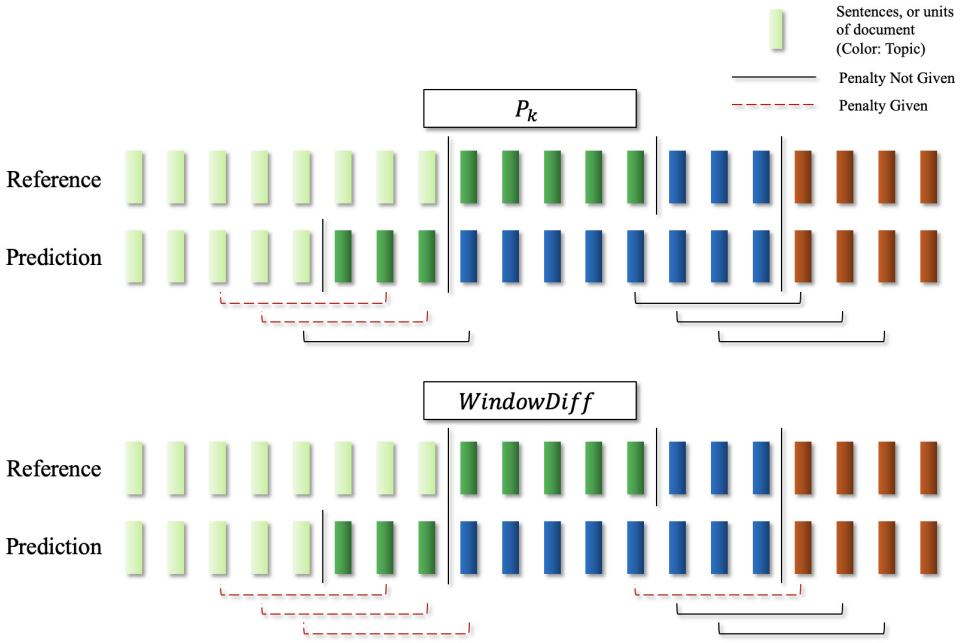

Figure 4. Illustrative Example Showing How and Differ in Penalizing

가능 분리 지점 중 특정 개수의 지점을 Uniform 임의 추출하는 방법론이며, Even은 주어진 문서에 대해 특정 길이 마다 분리 예측을 수행하도록 한다. 이 때 합성에 사용되는 기사의 수가 5개이므로 Random은 4개의 분리 지점을 임의 추출하였고, Even은 한 기사의 평균 문장 개수인 13개 문장마다 분리 지점을 형성, 평가에 사용하였다.

다음으로는 Koshorek *et al.*(2018)에서 제안한 LSTM 기반의 방법론을 비교 대상 모델로써 사용하였다. 해당 방법론의 경우 Word2Vec 기반의 단어 단위 임베딩을 구한 후 이를 단어 및 문장 수준의 2개 계층으로 이루어진 LSTM 구조에 입력하여 주제가 분리되는 지점을 예측하였다. 본 연구에서는 한국어 기사 데이터셋에 대해 LSTM 기반 모델을 학습 및 평가하기 위하여 대용량의 네이버 뉴스 말뭉치로 사전 학습된 Word2Vec 모델(https://github.com/dongjun-Lee/kor2vec.)을 활용하였다.

마지막으로 Jeon *et al.*(2019)에서 제안한 BERT based Text Segmentation(BTS) 모델을 비교 대상 모델로 사용하였다. BTS는 문서 내 한 쌍의 문장을 KoBERT에 입력하여 문장 사이에서의 주제 분리 여부를 예측하며, LSTM 기반의 방법론(Koshorek *et al.*, 2018) 등에 비해 높은 성능을 달성한 바 있다. 모델 구조적인 관점에서는 두 문장 단위의 임베딩을 이용한다는 점, 합성곱 신경망을 활용하지 않는다는 점 등에서 본 연구에서 제안하는 KoBERTSEG와 차이를 가진다.

4.3 구현 상세

본 연구는 KoBERTSEG의 Backbone으로써 KoBERT를 활용하였다. KoBERT는 기존 BERT의 한국어 성능 한계를 극복하기 위해 개발된 모델로, 위키피디아, 뉴스 등에서 수집한 대규모 한국어 말뭉치를 이용해 학습되었다. 한국어의 불규칙한 언어 변화의 특성을 반영하기 위해 데이터 기반 토큰화 기법을 적용하였으며 감성 분석, 객체명 인식 등 한국어 기반 후속 과업에 있어서 다국어 BERT에 비해 높은 성능을 달성하였다(https://github.com/SKTBrain/KoBERT).

Gururangan *et al.*(2020)은 Task-Adaptive Pretraining(TAPT)이 타겟 과업 성능 개선에 효과적임을 증명한 바 있다. 본 연구는 해당 연구 결과에서 아이디어를 차용하여 유사한 과업으로 미세 조정된 가중치 또한 타겟 과업 모델의 초기 가중치로서 효과적일 것이라는 가정 하에, 요약 과업을 위해 학습된 한국어 요약 모델 KoBERTSUM의 가중치를 KoBERTSEG의 초기 가중치로 사용하는 경우(KoBERTSEG-SUM)를 실험에 추가하였다. KoBERTSUM 또한 Backbone으로써 KoBERT를 활용하며, 주제 분리 모델 학습 시 활용한 한국어 기사 데이터셋을 이용해 요약 과업에 학습되었다.

KoBERTSEG 학습을 위한 GPU는 RTX2080ti를 이용하였고, 최대 150,000번의 Step으로 학습되었으며 그 중 $WindowDiff$가 가장 낮은 모델을 최종 평가에 사용하였다. 학습률의 경우 KoBERTSEG는 0.001, KoBERTSEG-SUM은 그보다 작은 0.0001를 이용하였고, 옵티마이저는 자연어처리 과업에 널리 활용되는 Adam Optimizer를 활용하였다. 모델 학습 시 윈도우 크기 k는 2부터 5까지 설정하여 실험하였는데, 연산 효율성과 더불어 $2k$개의 문장을 KoBERTSEG에 입력하였을 때 입력 토큰 수 제한(512)에 걸리는 경우를 방지하고자 각 문장의 최대 토큰 개수를 평균 토큰 개수($\left\lfloor \frac{512}{2k} \right\rfloor$)로 제한하였다.

5. 실험 결과

<Table 3>은 윈도우 크기에 따른 제안 모델의 주제 분리 성능을 나타낸 것이며, 총 네 가지의 후보군 중에서 윈도우 크기가 5일 때 가장 우수한 분리 성능을 나타냈다. 윈도우가 커짐에 따라 주제 분리 성능 또한 증가하는 경향성이 존재하며, 특히 재현율이 일정 수준에서 유지되는 반면 정밀도가 유의미하게 상승함을 확인할 수 있다. 이는 모델이 더욱 넓은 지역적 문맥 정보를 활용함에 따라 입력 텍스트가 갖는 노이즈에 견고해지고, 따라서 실제 분리 지점이 아닌 지점을 분리하도록 오판하는 1종 오류(False Positive)가 감소하는 현상으로 해석할 수 있다.

하지만 앞서 언급한 바와 같이 KoBERTSEG가 입력으로 사용하는 토큰의 개수 증가는 연산량의 이차적인 증가로 이어지며, 따라서 성능과 연산 효율 간 Trade-off가 존재한다. 또한 제안하는 모델을 적용할 실제 과업의 특성을 고려해 윈도우 크기를 설정하는 것이 중요한데, 예를 들어 6장에서 제안하는 뉴스 영상 요약 과업의 경우 한 주제가 세 문장, 혹은 네 문장 정도로 적은 개수의 문장으로 이루어지는 경우가 존재한다. 이러한 경우 2 또는 3 크기의 윈도우가 적합하며, 이후 실험에서는 성능 및 모델의 범용성을 고려하여 윈도우 크기가 3인 KoBERTSEG 모델을 활용하였다.

<Table 4>는 합성 기사 데이터에 대한 주제 분리 성능을 방

Table 3. Topic Segmentation Result of KoBERTSEG with Different Window Sizes

Window Size	Precision	Recall	F-1 Score	p_k	WindowDiff
2	94.93	98.65	96.41	0.0232	0.0271
3	95.69	98.50	96.79	0.0193	0.0233
4	97.37	98.50	97.71	0.0157	0.0173
5	**97.45**	**98.75**	**97.91**	**0.0129**	**0.0149**

Table 4. Results of Topic Segmentation Using Different Methods

	Precision	Recall	F-1 Score	p_k	WindowDiff
Random	6.40	6.30	6.34	0.5077	0.4877
Even	7.76	7.75	7.75	0.4118	0.4118
LSTM(Koshorek *et al.*, 2018)	67.66	89.80	76.25	0.1526	0.1531
BTS(Jeon *et al.*, 2019)	90.02	**98.90**	93.65	0.0355	0.0457
KoBERTSEG(window=3)	95.69	98.50	96.79	0.0193	0.0233
KoBERTSEG-SUM(window=3)	**97.38**	98.78	**97.86**	**0.0136**	**0.0157**

법론 간 비교한 결과이며, KoBERTSEG 모델이 다른 방법론에 비해 우수한 성능을 보임을 확인할 수 있다. 특히 BTS와의 비교 결과 재현율은 소폭 감소하는 반면 정밀도가 큰 폭 향상되었는데(90.02%→95.69%), 이를 통해 <Table 3>의 결과와 마찬가지로 1종 오류가 감소하였음을 알 수 있다. 이는 두 문장 단위의 정보를 이용하는 BTS 모델 대비 넓은 범위의 지역적 문맥을 이용하는 KoBERTSEG 모델이 새로운 어휘 등장과 같은 텍스트 노이즈에 대해 더욱 견고하기 때문이며, Appendix의 정성적인 예측 결과(<Table A>)에서도 이러한 특성을 확인할 수 있다. 무엇보다 1종 오류를 줄이는 것은 분리된 문단에 대한 문서 요약 등의 후속 과업 품질에 큰 영향을 미친다는 점에서 기존 방법론 대비 큰 개선점을 제공한 것이라 볼 수 있다.

추가적으로 요약 과업에 학습된 KoBERTSUM 모델의 가중치를 KoBERTSEG의 초기 가중치로 활용하는 경우(KoBERTSEG-SUM), 처음부터 학습된 KoBERTSEG에 비해 높은 성능을 보이고 있다. 이러한 결과는 KoBERTSUM 모델의 KoBERT 구조 가중치가 한국어 기사 데이터셋에 미세 조정되었다는 점 뿐만 아니라, 요약 과업이 문장 간의 의미적 관계를 충분히 이용하여야 한다는 점에서 주제 분리와 맥락을 같이 하는 유사 과업이기 때문이라고 해석할 수 있다.

본 연구에서 제안하는 KoBERTSEG는 Backbone으로써 12개 계층의 인코더 블록으로 구성된 BERT 구조를 활용한다.

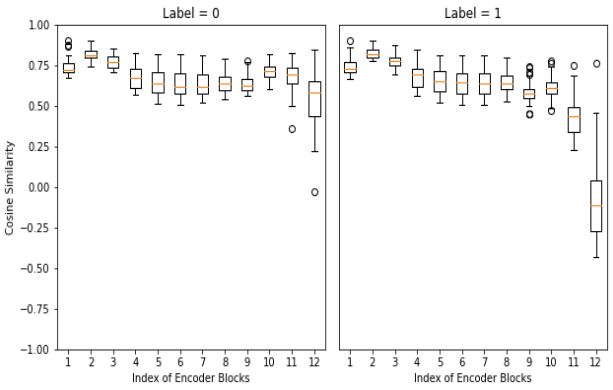

Figure 5. Cosine Similarity Calculated between Average Embeddings of the First and Last 3 Sentences of 1,000 Synthesized Articles with Window of Size 3

<Figure 5>는 윈도우 크기가 3인 1,000개의 합성 기사에 대해 앞과 뒤 3개 문장들의 [CLS] 토큰 임베딩 값의 평균을 구하고, 서로 간의 코사인 유사도(Cosine Similarity)를 KoBERT의 인코더 블록마다 시각화한 결과이다. 이 때 $Label = 1$인 경우는 주어진 6개 문장의 가운데에서 주제가 분리됨을 의미하며, $Label = 0$인 경우는 가운데에서 주제가 분리되지 않음을 의미한다. <Figure 5>에 따르면 10번째 인코더 블록까지는 두 경우 모두 3개 문장 단위 임베딩 평균 간 큰 차이를 보이지 않으나, $Label = 1$인 경우 11번째 및 12번째 인코더 블록을 거치면서 3개 문장 단위 임베딩 간 코사인 유사도가 급격히 감소함을 확인할 수 있다.

어텐션 메커니즘 기반의 BERT 구조는 낮은 계층에서는 언어의 표면적 정보를, 깊은 계층에서는 의미론적인 정보를 처리한다고 알려져 있으며(Jawahar *et al.*, 2019), 위 결과를 통해 KoBERTSEG가 활용하는 KoBERT 구조 또한 깊은 계층으로 갈수록 의미적인 관계 정보를 바탕으로 주제 분리를 위한 문장 단위 임베딩을 효과적으로 수행하고 있음을 확인할 수 있다.

KoBERTSEG는 KoBERT가 출력한 문장 단위의 임베딩 행렬을 합성곱 신경망 계층에 입력한 뒤, 그 결과물을 완전 연결 계층으로 구성된 분류기에 넣어 주제 분리 여부를 예측한다. <Table 5>는 KoBERT의 출력 결과물을 분류기에 입력하기 전 합성곱 연산을 수행하는 경우, 트랜스포머 연산을 수행하는 경우, 트랜스포머와 합성곱 연산을 모두 수행하는 경우, 그리고 어떠한 연산도 수행하지 않는 경우 간 성능을 비교한 결과이다.

이 때 BERT 구조가 출력한 문장 단위 임베딩을 트랜스포머에 입력해 문장 간 어텐션 연산을 추가적으로 수행하는 구조(Transformer)는 문서 요약을 위해 Liu *et al.* (2019)에서 제안되었다. 다만 Table 5의 결과에 따르면 트랜스포머 계층을 추가하는 것이 합성곱 신경망 만을 활용하는 경우에 비해 성능 개선 효과가 없음을 확인할 수 있다. 또한 KoBERT의 결과물을 바로 분류기에 입력하는 경우(None)에 비해서도 유의미한 성능 개선이 없는 것으로 나타나는데, 이는 문서 요약에 비해 상대적으로 간단한 주제 분리 과업에 대해 많은 파라미터를 적합시킴에 따라 학습 데이터에 대해 오버피팅(Over-fitting)이 발생하는 것으로 해석 가능하다.

Table 5. Ablation study for KoBERTSEG with window of size 3

Intermediate Layer	Precision	Recall	F-1 Score		WindowDiff
Conv (Ours)	95.69	98.50	**96.79**	**0.0193**	**0.0233**
Transformer	94.22	**98.85**	96.13	0.0248	0.0284
Transformer + Conv	95.27	98.28	96.41	0.0229	0.0271
None	**96.14**	97.48	96.49	0.0230	0.0265

6. 응용: 유튜브 뉴스 영상 요약 및 키포인트 매칭 프레임워크

앞선 실험에서 본 연구가 제안하는 KoBERTSEG 모델이 기사문을 효과적으로 주제별로 분리할 수 있음을 보였다. 이러한 주제 분리 방법론은 다른 후속 과업을 수행하는 데에 있어 전처리 과정의 하나로서 유용하게 활용될 수 있으나, 대부분의 연구는 주제 분리를 다른 과업에 적용한 결과를 직접 보여주지는 않았다. 본 연구는 한국어 기사에 대한 주제 분리 모델을 유튜브 뉴스 영상 요약 과업에 적용한 결과를 제시함으로써, 주제 분리 방법론의 효과 및 효용성을 추가적으로 보이고자 한다.

우선 유튜브 뉴스 영상 요약 과업을 수행하는 배경은 다음과 같다. 온라인 동영상 플랫폼이 급격하게 발전함에 따라 분당 약 400시간 분량의 영상 데이터가 업로드 되고 있으며, 이에 플랫폼 이용자들은 원하는 정보를 찾기 위해 많은 시간을 할애한다. 이러한 과정에서의 불편함을 해소하기 위해 효과적인 서비스 도입이 필요한 상황이나, 현실은 그러한 서비스가 매우 부족한 실정이다. 본 연구는 이러한 상황을 적극 반영하여, 현실 적용이 가능한 서비스로서 뉴스 영상 요약 및 키포인트 매칭 프레임워크를 제안하는 바이다.

주제 분리 기반의 뉴스 영상 요약 및 키포인트 매칭은 뉴스 영상을 사건 단위로 분리해 각 뉴스에 대한 생성 요약문을 제시하며, 추가적으로 각 사건의 영상에서의 시작 시간을 제공함으로써 사용자가 영상의 내용을 효율적으로 파악하도록 하는 일련의 프레임워크이다. 해당 프레임워크는 <Figure 6>과 같이 크게 오디오 텍스트화, 주제 분리, 생성 요약의 세 가지 모듈로 구성된다.

6.1 음성-텍스트 변환

뉴스 영상 요약 및 키포인트 추출 프레임워크를 구축하기 위해 1차적으로 필요한 요소는 여러 사건 혹은 주제를 연속적으로 다루는 뉴스 영상과, 해당 영상의 내용에 대해 주제 분리 및 요약 모델을 적용할 수 있도록 하는 자연어 스크립트이다. 본 연구는 여러 사건을 다루는 종합 뉴스 영상을 선정하고 이에 대해 Naver Clova의 Speech-to-Text(STT) API[1]를 적용해 자연어 스크립트를 추출하였다. 해당 API의 경우 인식된 발화 내용에 대한 Confidence 값을 추가적으로 출력하며, 외국어이거나 발음이 불명확한 경우 낮은 값을 갖는다. 따라서 Confidence 값에 대해 Threshold를 적용함으로써 모델이 해석 가능한 한국어 스크립트를 선택하였으며, "OOO 기자입니다"와 같이 의미가 없는 짧은 문장을 제거하는 등의 전처리를 수행하여 제안하는 프레임워크의 입력으로 활용될 스크립트를 생성하였다.

6.2 주제 분리

연속적으로 다양한 주제의 기사가 언급되는 뉴스 영상을 한 번에 요약 모델에 입력하게 되면 모델이 모든 토큰을 처리하기 어려울 뿐만 아니라 단일 주제 요약 목적으로 학습된 요약 모델이 의미 있는 요약문을 생성하기 어렵다. 따라서 전체 스크립트에 대해 의미 있는 요약문을 생성하기 위하여, 본 연구가 제안하는 KoBERTSEG 모델을 뉴스 영상 스크립트에 적용해 최대 하나의 사건 만을 포함하는 단위 기사로 분할하였다. 이 때 윈도우 크기는 3으로 설정하였고, Threshold 값은 0.7로 설정하였다. 분리 결과 예시는 <Table 6>과 같으며, 뉴스 영상 스크립트 내 사건이 변화하는 지점을 포착하여 적절히 분리하고 있음을 확인할 수 있다.

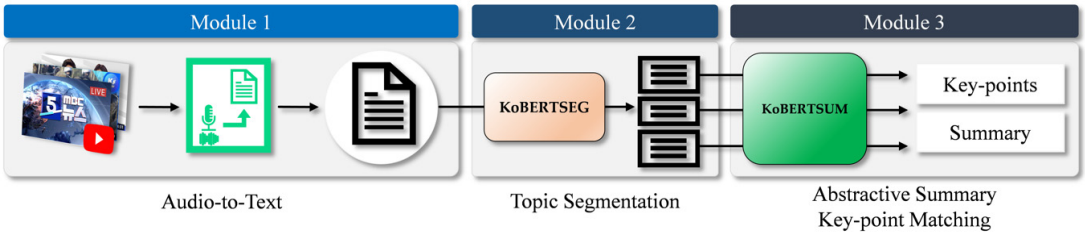

Figure 6. Youtube News clip Summarization & Key-point Matching Framework

1) https://www.ncloud.com/product/aiService/clovaSpeech.

Table 6. Result of Topic Segmentation on Youtube News Scripts

Sentences & Segmentation points
• 오늘 새벽부터 오전 사이에 중부지방을 중심으로 강하고 많은 비가 올 것으로 보입니다.
• 기상센터 연결해서 자세한 상황 알아보겠습니다.
… (중략) …
• 오늘 낮 기온은 서울이 21도에 머물겠습니다.
• 대전은 23도 전주 24도 대구 25도 부산은 21도 예상됩니다 이번 주 석가탄신일에는 대체로 맑은 날씨가 예상되지만 주 후반부터 충청 이남 지역을 중심으로는 또 한 차례 비 소식이 들어 있습니다.
========== Predicted Split Point ==========
• 코로나19 신규 확진자 수는 오늘 600명 안팎 예상됩니다.
• 하지만 여전히 모임과 직장 학교 등 일상생활 곳곳에 집단 감염이 전국에서 끊이지 않아서 불안감이 줄어들지 않고 있습니다.
… (중략) …
• 현장조사 결과 연습실 공간은 면적이 165제곱미터 가량으로 넓은 편이었으나 환기 상태가 좋지 않았던 것으로 파악되었습니다.
========== Predicted Split Point ==========
• 밤사이 서울 강동구의 한 건물에서 불이 났습니다.
• 큰 폭발음이 수차례 들리고 지나가던 시민들이 대피를 돕기도 했습니다.
• 사건 사고 소식 박찬범 기자 건물 틈 사이로 시뻘건 불길이 보이기 시작합니다.
… (중략) …
• 불은 20분 만에 꺼졌고 인명피해는 발생하지 않았습니다.
========== Predicted Split Point ==========
• 이스라엘군과 팔레스타인 무장정파 하마스 간의 대규모 무력충돌이 이어지는 가운데 가자지구 내 외신들이 입주한 건물이 이스라엘군의 공습으로 붕괴됐습니다.
• 조 바이든 미 대통령은 양측 정상과 각각 통화했습니다
… (중략) …
• 유엔 안전보장이사회는 오늘 무력분쟁 해소 방안을 논의하기 위해 화상회의를 가질 예정입니다.

6.3 생성요약 및 키포인트 매칭

제안하는 프레임워크의 마지막 모듈은 생성 요약 및 키포인트 매칭으로서, 앞서 주제 분리의 결과로 얻은 각 뉴스 사건에 대해 요약문을 생성하고, 나아가 각 사건이 영상에서 시작되는 시점과 매칭하는 단계이다. 본 연구는 생성 요약 모델로써 한국어 기사 데이터셋에 학습된 KoBERTSUM 모델을 활용하였고, 해당 모델이 출력한 생성 요약 및 키포인트 매칭 결과는 <Table 7>과 같다. 뉴스 영상에 대해 이와 같은 사건 단위의 요약문을 제공함으로써 동영상 플랫폼 이용자들이 동영상을 모두 시청하지 않음에도 전반적인 정보를 획득하는 데에 활용될 수 있다. 또한 영상 상에서 하나의 사건이 시작하는 시점을 제공함으로써, 동영상 플랫폼 이용자가 특정 사건에 대한 정보만을 획득하는 데에도 효과적인 보조 역할을 수행할 수 있다.

Table 7. Result of Abstractive Summarization and Key-point Matching

Key-point	Abstractive Summary
00:09	오늘 새벽부터 오전 사이에 중부지방을 중심으로 시간당 30 ~ 최고 50mm의 매우 강한 비가 내일 아침부터 낮 사이에 전국에 비가 내릴 수 있으며, 비의 양은 수도권과 강원 충청북부와 경북 북부를 중심으로 50 ~ 최고 100mm의 비가 예상된다.
02:00	코로나19 신규 확진자 수는 사흘 만에 600명대로 내려왔지만 일평균 지역 발생 확진자 수도 591명으로 여전히 사회적 거리 두기 2.5단계 범위를 벗어나지 못하고 있다.
03:47	어젯밤 10시 10분쯤 서울 강동구 천호동의 한 건물에서 불이 났는데, 폭발음 소리가 곧이어 들리고 시커먼 연기가 건물 주변을 가득 메웠으며, 소방당국은 건물 배전반에서 불길이 시작된 것으로 보고 정확한 화재 원인을 조사하고 있다.
04:37	어제 저녁 7시 50분쯤 인천 부평구 삼산동의 한 아파트 단지 내 정전이 발생해 전기 공급이 9시간 가까이 중단돼 입주민들이 밤새 큰 불편을 겪었는데 정전은 아파트 전기 설비 문제로 발생한 것으로 추정되고 있다.
05:27	현지시간으로 어제 AP통신과 카타르 국영방송 알자지라 등 외신 언론사들이 입주한 가자지구 내 12층짜리 건물이 이스라엘군의 공습으로 파괴됐는데, 이스라엘군은 해당 건물이 하마스에 의해 군사적으로 사용되고 있다고 주장했고, 조바이든 미국 대통령도 이스라엘과 팔레스타인 정상과 각각 통화하고 양측의 무력 충돌 해결 노력을 촉구하는 한편 언론의 안전을 보장할 필요성도 강조했다

7. 결 론

본 연구는 대용량의 한국어 기사 데이터셋과 KoBERT를 활용한 지역적 문맥 기반의 주제 분리 모델 KoBERTSEG를 제안하였다. KoBERTSEG는 문서 요약에 활용되는 구조를 차용하여 기존 대비 넓은 지역적 문맥 정보를 활용함으로써 높은 주제 분리 성능을 달성하였음을 실험적으로 증명하였다. 특히 텍스트 노이즈에 대해 강건함을 지님으로써 주제 분리에 있어서의 1종 오류를 낮추었으며, 이는 문서 요약, 문서 분류, 정보 검색 등의 후속 과업에 있어서 효용성이 높은 방법론임을 의미한다. 또한 본 연구는 주제 분리 방법론을 뉴스 영상 요약에 활용하는 실용적 프레임워크를 제안함으로써, 한국어 주제 분리 방법론의 효용성을 보여주고 나아가 실현 가능성이 높은 서비스를 제시하였다는 의미를 갖는다.

다만 본 연구의 제안 방법론은 지역적 문맥을 기반으로 하기 때문에 문서 전체의 토픽 분포 등을 활용하지 않는다는 한계점을 갖는다. 추후 본 연구의 제안 방법론에 더해 토픽 모델링 등 전역적인 정보를 활용하는 기법을 활용함으로써 문서의 토픽 분포를 함께 고려하는 방법론이 제안될 수 있을 것으로 기대한다. 또한 어휘 발생이 일관적이어서 상대적으로 주제 분리가 용이한 기사 데이터셋을 활용하였다는 점에서도 한계점을 가지지만, 언어 이해에 적합한 사전 학습된 BERT 구조를 활용하기 때문에 더욱 복잡한 주제 분리 과업에 대해서도 충분히 좋은 성능을 보일 수 있을 것으로 기대한다.

참고문헌

Alemi, A. A. and Ginsparg, P. (2015). Text segmentation based on semantic word embeddings. arXiv preprint arXiv:1503.05543.

Aumiller, D., Almasian, S., Lackner, S., and Gertz, M. (2021). Structural Text Segmentation of Legal Documents.

Badjatiya, P., Kurisinkel, L. J., Gupta, M., and Varma, V. (2018). Attention-based neural text segmentation. *Paper presented at the European Conference on Information Retrieval*.

Beeferman, D., Berger, A., and Lafferty, J. (1999), Statistical models for text segmentation, *Machine Learning*, **34**(1), 177-210.

Choi, F. Y. (2000). Advances in domain independent linear text segmentation. arXiv preprint cs/0003083.

Eisenstein, J. and Barzilay, R. (2008), Bayesian unsupervised topic segmentation, *Paper presented at the Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. (2020). Don't stop pretraining: adapt language models to domains and tasks. arXiv preprint arXiv:2004.10964.

Hearst, M. A. (1997), Text Tiling: Segmenting text into multi-paragraph subtopic passages, *Computational Linguistics*, **23**(1), 33-64.

Iikura, R., Okada, M., and Mori, N. (2020), Improving BERT with Focal Loss for Paragraph Segmentation of Novels, *Paper presented at the International Symposium on Distributed Computing and Artificial Intelligence*.

Jawahar, G., Sagot, B., and Seddah, D. (2019), What does BERT learn about the structure of language? *Paper presented at the ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Jeon, J. M., Choi, W. Y., Choi, S. J., and Park, S. Y. (2019), BTS: Text Segmentation using KoBERT, *Proceedings of the Korean Information Science Society Conference*, 413-415.

Kazantseva, A. and Szpakowicz, S. (2011), Linear text segmentation using affinity propagation, *Paper presented at the Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*.

Koshorek, O., Cohen, A., Mor, N., Rotman, M., and Berant, J. (2018), Text segmentation as a supervised learning task, arXiv preprint arXiv:1803.09337.

Liu, Y. and Lapata, M. (2019), Text summarization with pretrained encoders. arXiv preprint arXiv:1908.08345.

Misra, H., Yvon, F., Cappé, O., and Jose, J. (2011), Text segmentation: A topic modeling perspective, *Information Processing & Management*, 47(4), 528-544.

Pethe, C., Kim, A., and Skiena, S. (2020), Chapter captor: Text Segmentation in Novels, arXiv preprint arXiv:2011.04163.

Pevzner, L. and Hearst, M. A. (2002), A critique and improvement of an evaluation metric for text segmentation, *Computational Linguistics*, **28**(1), 19-36.

Riedl, M. and Biemann, C. (2012), Topictiling: a text segmentation algorithm based on lda, *Paper presented at the Proceedings of ACL 2012 Student Research Workshop*.

Solbiati, A., Heffernan, K., Damaskinos, G., Poddar, S., Modi, S., and Cali, J. (2021), Unsupervised Topic Segmentation of Meetings with BERT Embeddings, arXiv preprint arXiv:2106.12978.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017), Attention is all you need. Paper presented at the Advances in neural information processing systems.

저자소개

소규성 : 고려대학교 통계학과에서 2018년 학사학위를 취득하였다. 현재는 고려대학교 산업경영공학부에서 석사과정으로 재학중이다. 연구 분야는 비정형 데이터를 활용한 데이터마이닝이다.

이윤승 : 고려대학교 심리학과에서 2020년 학사학위를 취득하였다. 현재는 고려대학교 산업경영공학부에서 석사과정으로 재학중이다. 연구 분야는 이미지 데이터를 활용한 이상치 탐지이다.

정의석 : 고려대학교 산업경영공학과에서 2020년 학사학위를 취득하였다. 현재는 고려대학교 산업경영공학부에서 석사과정으로 재학 중이다. 연구 분야는 비정형 데이터 및 시계열 데이터를 활용한 데이터마이닝이다.

강필성 : 서울대학교 산업공학과에서 2003년 학사, 2010년 박사학위를 취득하였다. 이후 현대카드 과장, 서울과학기술대학교 조교수로 근무하였으며, 현재는 고려대학교 산업경영공학부 부교수로 재직중이다. 연구 분야는 정형 및 비정형 데이터를 활용한 데이터마이닝 및 기계학습 알고리즘 개발 및 제조/IT/공공분야 응용이다.

<Appendix>

A. Qualitative Result of Topic Segmentation

<Table A>는 BTS 모델과 윈도우 크기가 3인 KoBERTSEG 모델을 합성 기사에 적용하여 얻은 정성적인 예측 결과이다. 활용된 합성 기사는 두개의 기사를 합쳐 생성되었으므로 두 기사가 합쳐진 지점 하나 만을 분리 지점으로 예측해야 한다. KoBERTSEG와 달리 BTS 모델은 정답 분리 지점이 아닌 곳을 분리 지점으로 예측하는 1종 오류를 범하고 있으며, 이를 통해 KoBERTSEG 모델이 BTS 모델에 비해 어휘 변화와 같은 텍스트 노이즈에 더욱 강건함을 확인할 수 있다.

Table A. Qualitative Result of Topic Segmentation Using BTS and KoBERTSEG with Window of Size 3

Model	Result of Topic Segmentation
BTS (Jeon *et al.*, 2019)	• 세종시가 '국립세종과학관' 건립을 추진 중이다. • 사진은 대전 국립중앙과학관에 있는 '미래기술관' 모습이다. • 세종시는 27일 홈페이지에 올린 공고를 통해 "국립세종과학관 건립 기본계획 수립 및 타당성 조사 연구용역을 발주하기로 했다"며 "이에 따라 구성할 '제안서 평가위원회(위원 7명)'에서 활동할 예비평가위원 21명을 6월 5일까지 공개 모집한다"고 밝혔다. • 시에 따르면 이번 용역의 사업비는 총 4천만 원, 수행 기간은 착수일부터 4개월이다. ========== Predicted Split Point ========== (**False Positive**) • 지원서는 경제정책과 과학기술담당(044-300-4023)에서 받는다. • 과학기술정보통신부 산하 국립중앙과학관은 대전 대덕연구단지(유성구 구성동 32)에 있다. • 중앙과학관은 과천,부산,대구,광주 등 전국 4개 도시에 분원 형태의 국립과학관을 두고 있다. • 하지만 현재 정부과천청사에 있는 과학기술정보통신부는 오는 8월 세종시로 이전한다. • 또 세종시에는 KDI(한국개발연구원) 등 15개 기관으로 이뤄진 국책연구단지가 있고, 산학연(産學硏) 클러스터 중심의 세종테크밸리도 조성되고 있다. • 이에 따라 세종시는 국립과학관이 들어서기에 충분한 조건을 갖춘 도시라고 볼 수 있다. ========== Predicted Split Point ========== (**True Positive**) • 전라북도 예술인들의 큰 잔치 '제58회 전라예술제'가 5월 8일부터 12일까지 고창 모양성 잔디광장에서 열린다. • (사)한국예총 전북연합회(회장 선기현)가 주최하고 전라북도와 고창군이 후원하는 이번 예술제는 제56회 전북도민체전 기간에 열려 예향전북의 이미지를 부각시키고, 관광객에게 볼거리를 제공하는 동시에 종합예술축제의 면면을 과시한다. • '빛나라 전라예술 신나라 도민체전!'이라는 슬로건으로 열리는 올해 예술제의 개막식은 5월 8일 오후 7시에 열린다. • 개막식과 함께 열리는 개막공연은 전북음악협회(회장 이석규)가 준비하는 '비상하는 전북, 천년의 소리'이다. • 이 무대에는 가수 진성을 비롯해 남성4중창단 빅브라더스, 전북팝오케스트라, 피아노 3중주의 한울트리오, 소프라노 장수영, 테너 윤호중 씨가 함께한다. • 이어 예술제에서는 국악, 무용, 연극, 연예 등 4개 공연단체가 매일 오후 2시와 밤 7시에 공연을 갖는다. • 전북국악협회(회장 소덕임)는 5월 10일 오후 1시 30분 '우리 소리의 향기'를 통해 다양한 장르의 국악의 세계로 관객들을 초대한다. • 모악한우리농악단의 풍물한마당으로 문을 열고, 시조와 판소리, 부채춤, 가야금병창, 창극, 민요, 타악 퍼포먼스까지 관객들이 추임새를 넣고, 어깨춤을 추며 한바탕 신명 나게 즐길 수 있는 시간을 선물한다. • 전북무용협회(회장 염광옥)는 5월 11일 오후 7시 30분 '100년의 춤, 봄을 맞다'로 수준 높은 예술문화와 감성의 시야를 한 단계 높여 줄 무대를 선보인다. • 태무용단의 '음・양', 전주시지부의 '춘화', 익산시지부의 '그날', 남원시지부의 '소고춤, 군산시지부의 '진쇠춤', 정읍시지부의 '써니'등 창작과 전통을 넘나드는 무대가 과거와 미래의 축복을 기원한다. ========== Predicted Split Point ========== (**False Positive**) • 전북연예예술인협회(회장 김용철)는 5월 12일 오후 7시 30분에 '제26회 전라예술가요제'를 펼친다. • 전문 연주인들로 구성된 빅밴드의 고품격 연주와 함께 치열한 경쟁 끝에 무대에 오른 아마투어 가수들이 우리 가요의 참맛을 선사하며, 초대가수로 진국이, 신송, 김덕진 등이 무대에 오른다. • 이 기간 중 오후 3시 30분에는 고창농악, 영산작법, 고창오거리당산제, 김만경외애밋들노래, 전주기접놀이 등 전라북도 주요 민속작품과 호흡하는 '우리민속페스티발'이 축제 속 축제로 열려 예술제를 꽉 채운다. • 이 밖에도 건축협회(회장 문창호), 문인협회(회장 류희옥), 미술협회(회장 김영민), 사진협회(회장 전종권) 등 4개 협회는 오전 10시부터 오후 7시까지 야외전시장에서 작품전시회를 갖는다.

Model	Result of Topic Segmentation
KoBERTSEG (Window=3)	• 세종시가 '국립세종과학관' 건립을 추진 중이다. • 사진은 대전 국립중앙과학관에 있는 '미래기술관' 모습이다. • 세종시는 27일 홈페이지에 올린 공고를 통해 "국립세종과학관 건립 기본계획 수립 및 타당성 조사 연구용역을 발주하기로 했다"며 "이에 따라 구성할 '제안서 평가위원회(위원 7명)'에서 활동할 예비평가위원 21명을 6월 5일까지 공개 모집한다"고 밝혔다. • 시에 따르면 이번 용역의 사업비는 총 4천만 원, 수행 기간은 착수일부터 4개월이다. • 지원서는 경제정책과 과학기술담당(044-300-4023)에서 받는다. • 과학기술정보통신부 산하 국립중앙과학관은 대전 대덕연구단지(유성구 구성동 32)에 있다. • 중앙과학관은 과천, 부산, 대구, 광주 등 전국 4개 도시에 분원 형태의 국립과학관을 두고 있다. • 하지만 현재 정부과천청사에 있는 과학기술정보통신부는 오는 8월 세종시로 이전한다. • 또 세종시에는 KDI(한국개발연구원) 등 15개 기관으로 이뤄진 국책연구단지가 있고, 산학연(産學硏) 클러스터 중심의 세종테크밸리도 조성되고 있다. • 이에 따라 세종시는 국립과학관이 들어서기에 충분한 조건을 갖춘 도시라고 볼 수 있다. ========== Predicted Split Point ========== (**True Positive**) • 전라북도 예술인들의 큰 잔치 '제58회 전라예술제'가 5월 8일부터 12일까지 고창 모양성 잔디광장에서 열린다. • (사)한국예총 전북연합회(회장 선기현)가 주최하고 전라북도와 고창군이 후원하는 이번 예술제는 제56회 전북도민체전 기간에 열려 예향전북의 이미지를 부각시키고, 관광객에게 볼거리를 제공하는 동시에 종합예술축제의 면면을 과시한다. • '빛나라 전라예술 신나라 도민체전!'이라는 슬로건으로 열리는 올해 예술제의 개막식은 5월 8일 오후 7시에 열린다. • 개막식과 함께 열리는 개막공연은 전북음악협회(회장 이석규)가 준비하는 '비상하는 전북, 천년의 소리'이다. • 이 무대에는 가수 진성을 비롯해 남성4중창단 빅브라더스, 전북팝오케스트라, 피아노 3중주의 한울트리오, 소프라노 장수영, 테너 윤호중 씨가 함께한다. • 이어 예술제에서는 국악, 무용, 연극, 연예 등 4개 공연단체가 매일 오후 2시와 밤 7시에 공연을 갖는다. • 전북국악협회(회장 소덕임)는 5월 10일 오후 1시 30분 '우리 소리의 향기'를 통해 다양한 장르의 국악의 세계로 관객들을 초대한다. • 모악한우리농악단의 풍물한마당으로 문을 열고, 시조와 판소리, 부채춤, 가야금병창, 창극, 민요, 타악 퍼포먼스까지 관객들이 추임새를 넣고, 어깨춤을 추며 한바탕 신명 나게 즐길 수 있는 시간을 선물한다. • 전북무용협회(회장 염광옥)는 5월 11일 오후 7시 30분 '100년의 춤, 봄을 맞다'로 수준 높은 예술문화와 감성의 시야를 한 단계 높여 줄 무대를 선보인다. • 태무용단의 '음 · 양', 전주시지부의 '춘화', 익산시지부의 '그날', 남원시지부의 '소고춤, 군산시지부의 '진쇠춤', 정읍시지부의 '써니'등 창작과 전통을 넘나드는 무대가 과거와 미래의 축복을 기원한다. • 전북연예예술인협회(회장 김용철)는 5월 12일 오후 7시 30분에 '제26회 전라예술가요제'를 펼친다. • 전문 연주인들로 구성된 빅밴드의 고품격 연주와 함께 치열한 경쟁 끝에 무대에 오른 아마투어 가수들이 우리 가요의 참맛을 선사하며, 초대가수로 진국이, 신송, 김덕진 등이 무대에 오른다. • 이 기간 중 오후 3시 30분에는 고창농악, 영산작법, 고창오거리당산제, 김만경외애밋들노래, 전주기접놀이 등 전라북도 주요 민속작품과 호흡하는 '우리민속페스티발'이 축제 속 축제로 열려 예술제를 꽉 채운다. • 이 밖에도 건축협회(회장 문창호), 문인협회(회장 류희옥), 미술협회(회장 김영민), 사진협회(회장 전종권) 등 4개 협회는 오전 10시부터 오후 7시까지 야외전시장에서 작품전시회를 갖는다.