KIPS Tr. Software and Data Eng.
Vol.5, No.11 pp.585~592 pISSN: 2287-5905

http://dx.doi.org/10.3745/KTSDE.2016.5.11.585

On-Line Topic Segmentation Using Convolutional Neural Networks

Gyoung Ho Lee† · Kong Joo Lee††

ABSTRACT

A topic segmentation module is to divide statements or conversations into certain topic units. Until now, topic segmentation has progressed in the direction of finding an optimized set of segments for a whole document, considering it all together. However, some applications need topic segmentation for a part of document which is not finished yet. In this paper, we propose a model to perform topic segmentation during the progress of the statement with a supervised learning model that uses a convolution neural network. In order to show the effectiveness of our model, we perform experiments of topic segmentation both on-line status and off-line status using C99 algorithm. We can see that our model achieves 17.8 and 11.95 of Pk score, respectively,

Keywords : Topic Segmentation, Convolutional Neural Network

합성곱 신경망을 이용한 On-Line 주제 분리

이 경 호† · 이 공 주††

요 약

글이나 대화를 일정한 주제의 단위로 나누는 것을 주제 분리라고 한다. 지금까지 주제 분리는 주로 완결된 하나의 문서에서 최적화된 분리를 찾는 방향으로 진행되어 왔다. 하지만 몇몇 응용은 글이나 대화가 진행 중에 주제 분리를 할 필요가 있다. 본 논문에서는 합성곱 신경망을 이용한 교사 학습 모델을 통해 문장의 진행 중에 주제 분리를 수행하는 모델에 대해 제안한다. 그리고 제안한 모델의 성능 검증을 위해 On-line 상황을 가정한 실험과 기존의 C99모델을 결합한 실험을 수행하였다. 실험결과 각각 17.8과 11.95의 Pk 점수를 얻었고, 이를 통해 본 논문의 모델을 통한 On-line 상황에서의 주제 분리 활용의 가능성을 확인하였다.

키워드 : 주제 분리, 합성곱 신경망

1. 서 론

글이나 대화를 일정한 주제의 단위(segment)로 나누는 것을 주제 분리(Topic segmentation, TS)[1]라고 한다. TS는 정보검색, 문서 요약 등 자연언어 처리의 여러 응용에 활용 될 수 있다.

지금까지 TS 연구는 주로 완결된 하나의 문서(document)에서 최적화된 분리를 찾는 방향으로 진행되어 왔다[2, 3]. 하지만 몇몇 응용은 완결된 상태가 아닌, 글이나 대화가 진행되는 중(On-line)에 분리가 필요가 있다. 이러한 예 중 하나가 음성인식 시스템이다. 시스템이 현재 대화의 주제를 알 수 있다면, 음성 인식 과정에서 발생하는 후보 단어들 중 적합한 어휘를 선택하는데 긍정적인 영향을 줄 수 있다[4]. 이러한 예로 Fig. 1과 같은 상황이 있다.

A-1: 응급실 진료비가 얼마 나왔어?
B-1: 잠깐만 확인해 볼게. 백이십 달러 나왔네.
A-2: 다음 검사는 언제야?
B-2: 약 30분 뒤에 CT촬영을 한대.
A-3: 검사 끝나고 저녁 먹자. 근처에 맛있는 집 알고 있니.
B-3: 응 있어. 병원 건너 다리 앞에 보면 음식점이 모여있어.

Fig. 1. Example of Conversation 1

Fig. 1은 A와 B가 병원에서 대화를 하고 있는 상황의 예이다. 인간의 경우 B-2의 내용을 들을 때, 앞서 대화에 나타난 단어 “응급실”, “진료비”, “검사” 등의 단어를 통해 이 대화의 주제가 병원과 관련된 것임을 짐작하고 이를 이용하여 “CT”이라는 단어를 쉽게 인식 할 수 있다. 하지만 일반

※ 본 연구는 한국전자통신연구원 연구운영비지원사업의 일환으로 수행하였음 [16ZS1110, 언어장벽 없는 국가 구현을 위한 자동통번역 산업 경쟁력 강화 사업].
† 준 회 원 : 충남대학교 전자전파정보통신공학과 박사과정
†† 종신회원 : 충남대학교 전파정보통신공학과 교수
Manuscript Received : October 5, 2016
Accepted : October 13, 2016
* Corresponding Author : Kong Joo Lee(kjoolee@cnu.ac.kr)

적인 음성인식기는 AM(Acoustic Model)과 LM(Language Model)을 이용하여 “뒤에” 다음에 인식될 다양한 인식 후보 단어를 생성 하고 이들 중 적절한 단어를 선택한다. 음성인식기의 후보 단어 예는 Fig. 2와 같다.

INDEX	전단어 인식 결과	인식 후보
INDEX:1	뒤에	시디
INDEX:2	뒤에	시티
INDEX:3	뒤에	CT
INDEX:4	뒤에	CD
INDEX:5	뒤에	SEED
INDEX:6	뒤에	시기
INDEX:7	뒤에	식기
....		

Fig. 2. Example of Speech Recognition 1

음성인식기가 대화의 주제를 인식 하고 각 후보 단어와 주제의 관련성을 평가할 수 있다면, “응급실”, “진료비”, “검사”가 쓰인 글에 “식혜”, “식기”, “CD”보다 “CT”가 더 적합한 선택임을 알 수 있다. 이러한 과정을 통해 음성인식기가 다른 단어들 보다 “CT”에 가중치를 더 부여 하도록 할 수 있고 이를 통해 음성인식기 성능향상에 도움을 주는 것이 가능하다.

하지만 대화 진행 중에 A-3과 B-3의 대화처럼 대화의 주제가 바뀌는 경우가 있다. “다리 앞에”라는 발화를 인식할 때 인식 후보들의 일부는 다음 Fig. 3과 같다.

INDEX	이전 인식 결과	인식 후보
INDEX:1	다리	앞에
INDEX:2	다리	어폐
INDEX:3	다리	어때
INDEX:4	다리	없데
INDEX:5	다리	아퍼

Fig. 3. Example of Speech Recognition 2

인간은 A-3의 발화에서 “근처”, “어디”, “가야” 등의 단어를 통해 대화의 주제가 위치와 관련된 내용임을 알 수 있다. 그래서 “다리 앞에”를 인식할 때, 다리 다음에 위치와 관련된 “앞에”가 인식되는 것이 자연스럽다. 하지만 대화 주제가 변한 것을 음성인식기가 알지 못하였다면, Fig. 3의 INDEX:5 “다리 아퍼”와 같이 잘못된 단어에 가중치가 부여될 수 있다. 이러한 문제를 피하기 위해 대화가 진행하는 중에 대화의 주제가 변화하였다는 것을 인식해야 한다.

대화에서 주제 변화는 종종 발생하는 일이다. 만약 시스템이 인식하고 있는 주제가 이미 지나가버린 다른 주제라면, 위의 예와 같이 후보단어 선택에 부정적인 영향을 끼치게 된다. 그렇기 때문에 대화가 진행되는 동안 주제가 변화하였다는 것을 인식해야 할 필요가 있다. 대화 진행 중 주제가 변한 것을 적절한 시기에 인식 할 수 있다면, 이전까지 인식된 주제의 정보의 반영을 제한하고 새로운 주제 정보를 탐지하여 시스템이 이를 활용하게 할 수 있다.

하지만 문서 작성이나 대화가 진행되는 중간에 주제변화를 감지하는 것은 어려운 문제이다. 현재 위치 문장과 이전 문장 간의 유사도 차이가 전체 글의 흐름에서 용인 가능 한 수준인지, 그 수준을 넘어서는 것인지 판단하기 어렵다. 그렇기 때문에, 이전의 많은 TS알고리즘들은 하나의 완결된 글에서 전역적(global)으로 최적화된 분리 위치를 찾도록 연구가 되어 왔다[2, 3]. 하지만 이러한 알고리즘들은 작성이 진행 중인 문서 또는 대화에 적용하기 어렵다.

본 연구에서는 이러한 문제에 적용할 수 있는 Local 정보 기반의 주제 변화 인식 모델을 제안한다. 구체적으로, 문장의 경계 위치 p에서의 주제 분리 여부를 결정하기 위해, 좌우의 일정 범위($window$)에 속한 문장들을 문장 블록(sb_l, sb_r)으로 삼고, 이들의 관계를 이용하여 주제 분리 여부를 판단하는 모델에 대해 제안한다. 하지만 이러한 경우, 전체 문서와 문장 블록들 간의 관계를 주제 분리에 활용할 수 없다. 이러한 문제를 교사학습을 통해 완화하고자 한다.

본 논문의 구성은 아래와 같다. 2장에서는 주제 분리와 관련된 연구동향에 대하여 살펴본다. 3장에서는 본 논문에서 제안하는 모델에 대한 설명과 학습 방법에 대하여 설명한다. 4장에서는 이러한 정보를 학습 할 수 있도록 하는 학습데이터를 구성하는 방법과 이를 모델에 적용하여 주제 분리를 수행한 결과를 나타낸다. 마지막으로 5장에서 본 논문에 결과에 대하여 고찰하였다.

2. 관련 연구

주제 분리는 자연언어 처리 분야의 오랜 연구 대상이었다. 이 분야의 대표적인 알고리즘으로는 TextTiling(TT)가 있다[2]. TT는 인접한 두 블록의 유사도를 계산하고, 전체 문장에서 블록 유사도들의 지역 최소점을 주제 분리 위치로 결정한다. TT에서는 각 블록에서 나타난 단어들을 이용하여 블록을 표현하고, 이를 이용하여 유사도를 계산한다. TT 이후, 여러 알고리즘들이 TT를 기반으로 TS성능을 향상 시켜 왔다. LcSeg[5]가 이러한 종류의 알고리즘이다. 기존의 TT가 단순히 단어의 출현 여부와 빈도만을 이용하였다면, Lcseg는 문장 표현의 자질로 문장을 구성한 단어의 tf-idf 가중치를 사용하였고 이를 통해 TS의 성능을 향상시켰다.

단어 기반의 자질은 희소성 문제를 가지게 된다. 이러한 문제를 완화하기 위하여 LDA 도입을 시도한 연구들이 있었다[6, 7]. 후보 Segment들에 대한 표현을 단어 기반의 표현이 아닌 LDA를 통해 주제 벡터로 표현함으로써, 단어 기반이 가지는 희소성 문제를 완화 할 수 있고 LDA가 가진 주제에 대한 표현이라는 장점을 TT에 적용할 수 있다. 이러한 연구로는 TopicTiling[6]이 있다. Topic-tiling은 Segment

의 최소 단위를 개별 문장으로 사용한다. 학습된 LDA 모델을 이용하여 문장을 구성하는 각 단어들이 가질 수 있는 토픽을 결정하고, 이를 문장의 자질로 사용하였다.

TT분야의 또다른 대표적인 알고리즘으로 C99이 있다[3]. C99알고리즘은 Similarity matrix(S)와 Rank matrix(R)를 이용하여 전체 문서에 전역적으로 최적화된 주제 분리를 수행한다. Similarity matrix 의 S_{ij}는 i 번째 문장과 j번째 문장의 유사도를 나타낸다. S_{ij}는 각 문장을 단어 빈도수(tf)을 사용하여 표현하고 이를 이용해 계산한 코사인 유사도를 사용한다. 이렇게 계산된 S의 대비(Contrast)를 강화하기 위하여 R을 계산한다. R_{ij}는 일정한 범위 안에서 i 번째 문장과 j번째 문장의 유사도 보다 낮은 유사도를 갖는 원소들 개수를 의미한다. 이렇게 계산된 R에 Hierarchical Clustering 알고리즘을 적용하여 전체 문서에 대한 Segmentation을 수행한다. 여러 연구들이 C99을 기반으로 성능향상을 위해 연구되었다. 기존의 C99이 유닛의 자질로 tf를 사용하였지만, C99LDA[7]는 기본 유닛의 표현에 LDA를 이용하였다. [8]에서는 각각의 유닛 표현을 단어 임베딩[9]으로 대체, 이를 C99에 적용하였다.

그 외에 Dynamic programming 기법을 이용한 전역 최적화 알고리즘들이 최근 좋은 TS 결과를 나타내고 있다[10]. 하지만 이 방법들은 문장의 표현 보다는 분리 과정에 초점을 맞춘 연구이기 때문에 본 연구와의 직접적인 관련성은 낮다.

3. 주제 분리 모델

본 논문의 주제 분리 모델은 문서 전체가 아닌 위치 p에서 좌우 window범위의 블록을 이용하여 주제 분리를 결정한다. 문장을 주제 분리의 최소 단위로 사용하였고 $window$를 2로 설정하여 시스템 설계와 실험을 진행하였다. 윈도우사이즈를 2 이상으로 설정하면 대화가 변화한 그 순간 주제 변화를 감지하지 못한다. 하지만 일반적으로 사람들 사이의 대화에서도 몇 번의 대화가 흐른 후에 주제 변화에 적응하는 경우가 있다. 그렇기 때문에 본 논문에서는 윈도우 사이즈를 2로 설정하여 대화 주제 변화에 대한 즉각적인 검출 보단 약간의 지연이 있더라도 좀 더 정확하게 주제 변화를 찾을 수 있도록 시스템을 설계하였다.

다음과 같이 문서 $D=\{s_1,s_2,...,s_n\}$ (s_k: 문장, n: 문서의 문장 수) 가 있을 때, 이 문서는 $Seg=\{seg_0, seg_1, seg_2, ..., seg_n\}$의 경계 레이블을 갖는다. 이때 $seg_k=1$은 이 위치에서 주제가 분리되었음을, $seg_k=0$은 앞뒤 문장이 서로 같은 주제임을 나타낸다. seg_0은 문서의 시작을 의미하고 seg_n는 문서의 끝 위치이다.

문장 s_p와 s_{p+1}사이의 경계 레이블 seg_p는 다음과 같이 계산한다.

Seg0	
S1	응급실 진료비가 얼마 나왔어?
Seg1	
S2	잠깐만 확인해 볼게, 백이십 달러 나왔네.
Seg2	
S3	다음 검사는 언제야?
Seg3	
S4	약 30 분 뒤에 CT촬영을 한대.
Seg4	
S5	검사 끝나고 저녁 먹자. 근처에 맛있는 집 알고 있니?
Seg5	
S6	응 있어. 병원 건너 다리 앞에 보면 음식점이 모여 있어
Seg6	

Fig. 4. Example of Document Structure

$$seg_p = M_{Model}(sb_l, sb_r, threshold) \quad (1)$$

seg_p의 경계 레이블은 해당 위치의 앞 뒤 블록 sb_l, sb_r의 관계와 $threshold$를 통해 결정된다. 이때 sb_l은 p의 앞에 등장한 $window$크기 범위의 문장들($s_{p-w+1} \sim s_p$)를 의미하고 sb_r은 뒤에 등장한 문장들($s_{p+1} \sim s_{p+w}$)를 의미한다. M_{model}은 sb_l, sb_r에 대한 자질을 생성하고 관계를 계산하는 모델이다. M_{model}에서 계산된 관계점수에 threshold를 적용하여 주제 분리 레이블을 결정한다. 본 논문에서는 아래 2가지 모델에 대해 제안하고 결과를 살펴보았다.

M_{base}

위치 p의 앞뒤에 위치한 문장들은 그 위치에서 주제 분리 여부를 결정하는데 중요한 역할을 한다. 앞뒤 문장들의 관계가 서로 밀접하다면 문장들이 같은 주제일 가능성이 높다. 반면 관계가 서로 멀다면 다른 주제의 글일 가능성이 높다. p에서의 레이블 결정을 위해 p의 앞쪽에 나타난 문장 블록 sb_l을 하나의 자질로 표현하고($feature_{sbl}$), sb_r을 같은 방식으로 표현하였다($feature_{sbr}$).

p의 앞뒤 블록의 표현이 결정되면, 두 블록간의 관계점수는 cosine distance 로 계산된다.

$$score_p = \mathrm{CosineDistance}(\mathrm{feature}_{sbl}, \mathrm{feature}_{sbr}) \quad (2)$$

위치 p에서의 주제 분리 레이블은 다음과 같이 결정된다.

$$label_p = \begin{matrix} 1 \mathrm{if} Score_p > threshold \\ 0 \mathrm{else} \end{matrix} \quad (3)$$

즉, 두 블록 사이의 cosine distance 가 일정 수준 이상이면 이 모델은 두 블록 사이의 주제가 변화하였다고 (label = 1) 판단한다.

M_{CNN}

이 모델은 인공신경망 모델에 문장들의 자질을 입력하고 모델에서 예측한 문장들간의 관계를 통해 주제 분리를 결정한다. 이 모델의 인공신경망 구조를 Fig. 5에 나타내었다.

이 모델은 4개의 문장($s_{p-1}, s_p, s_{p+1} s_{p+2}$)에 대한 자질을 입력으로 받는다. 입력된 문장들은 2개의 합성 곱 신경망(Convolutional Neural Network, CNN)과 Max-pooling 층을 거치게 된다.

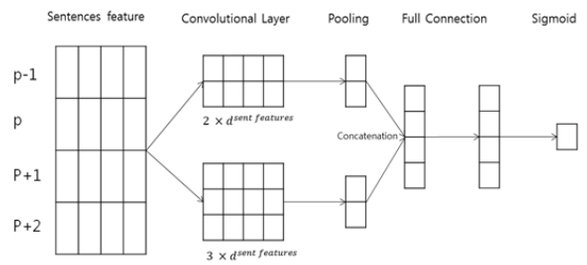


Fig. 5. Model of M_{CNN}

모델에 입력된 문장 표현들에 크기가 2와 3인 1-dimension convolutional layer 필터를 적용시킨다. 이 필터들은 인접한 문장들 또는 이어지는 3개의 문장들의 관계를 모델에 반영하기 위해 설정되었다. 각각의 필터들은 300차원의 출력을 만들어 낸다(각 단어들은 300개의 차원을 가지고 단어들의 임베딩 합으로 생성된 문장의 표현도 300차원을 갖는다. CNN레이어의 출력도 이와 같은 크기를 사용하였다). 필터를 거친 결과들을 Max-pooling 층을 적용하여 각각 300차원을 가진 출력을 생성하고 이들을 연결하여(Concatenation) 하나의 벡터로 만들었다. 이 벡터를 크기가 600인 은닉 층(Hidden Layer)에 입력하고 출력 층에 Sigmoid를 적용하여 0~1 사이의 출력 값을 갖도록 하였다.

이 모델은 4개 문장에 대한 자질 입력을 통해 가운데 두 문장(s_p, s_{p+1}) 사이의 주제 분리 레이블(seg_p)을 학습하고 예측한다. 출력 층을 Sigmoid로 설정하여 관계 점수를 0~1(관계 있음=0, 관계없음=1) 나타낼 수 있도록 하였다. 모델을 통해 관계 점수에 Equation (3)과 같이 threshold를 적용하여 문장 블록간의 레이블을 결정한다.

4. 실험 및 평가

본 논문에서 제안하는 모델의 성능을 검증하기 위하여 2가지 실험을 진행하였다. 첫번째 실험은 On-line 상황의 주제 변화 인식 과정을 가정하고 본 논문에서 제안하는 모델들의 성능을 평가 하였다. 두번째 실험은 M_{CNN}모델을 기존의 C99의 유사도 계산 알고리즘으로 사용하였을 때 주제 분리 성능을 평가하였다. 이를 통해 본 논문에서 제안한 모델이 지역 정보로부터 적절한 수준의 문서간 관계 평가를 할 수 있음을 나타내고자 한다.

4.1 평가데이터

실험에 사용하는 데이터는 TS 분야 연구에서 주로 사용되는 choi Text segmentation data를 활용하였다[3]. 이 데이터는 총 700개의 문서(document)로 구성되며 각 문서는 각기 다른 출처에서 추출된 10개의 주제 세그먼트로 구성된다. 문서는 각기 다른 최소, 최대 세그먼트 길이를 가지는 4개의 그룹으로 나누어진다. 각 그룹의 세그먼트 최소, 최대 길이에 따른 데이터 개수는 Table 1과 같다.

Table 1. Choi Data Structure

Segment Length	3-11	3-5	6-8	9-11
Count of Document	400	100	100	100

4.2 학습데이터

본 논문의 M_{CNN}모델을 학습하기 위해 학습 데이터가 필요하다. M_{CNN}모델은 위치 p의 좌우 4문장($s_{p-1}, s_p, s_{p+1} s_{p+2}$)을 이용하여 주제 분리 여부를 결정한다. 실제 상황에서 이들 4문장 사이에 나타날 수 있는 주제 분리 레이블($seg_{p-1}, seg_p, seg_{p+1}$) 관계는 총 8가지이다((0,0,0), (0,0,1), (0,1,0), (0,1,1), (1,0,0), (1,0,1), (1,1,0), (1,1,1), 0=같은 주제, 1=주제 분리). 영문판 위키피디아 문서에서 문장들을 수집하여 8개의 관계 중 하나의 관계를 만족하는 문장 집합을 생성하였다. 이러한 문장 집합을 천만개 수집하고 이를 학습에 사용하였다. 수집된 문장에는 소문자화, 토큰화(Tokenization), 어근화(Lemmatization)를 적용하여 학습에 사용하였다. 수집된 데이터의 분포와 예는 Table 2와 3과 같다.

Table 2. Distributions of Training Data

Type	0,0,0	0,0,1	0,1,0	0,1,1
Distribution	13.8%	12.8%	12.7%	12.1%
Type	1,0,0	1,0,1	1,1,0	1,1,1
Distribution	12.8%	12.1%	12.1%	11.5%

4.3 학습 방법

모델을 학습하기 위해서 back-propagation 알고리즘을 이용하였고 모델 학습을 위한 Batch-size를 128로 설정하였다. 파라미터 최적화를 위해서 초기 학습률(learning rate)를 0.025로 설정한 Adagrad알고리즘을 이용하였으며 모델이 예측한 문장 간의 관계에 대한 출력(sigmoid, 0~1)과 주제 레이블 값(0 또는 1)의 차이를 최소화하기 위한 목적함수로 binary cross-entropy를 사용하였다.

4.4 실험 결과

1) On-Line 주제 변화 실험

본 장에서는 On-line상황을 가정하였을 때, 본 논문에서 제안한 모델들의 성능을 평가하였다. 하나의 문서에서 각

Table 3. Examples of Training Data

Type	Example
0,0,0	- After studying at the Brighton School of Art, she had her first solo exhibition in 1953 at the Kensington Art Gallery. - She married Lawrence Alloway, a curator and art critic, before moving to the United States in 1961. - The following year, Alloway became a curator at the Solomon R. Guggenheim Museum. - Around 1970, from feminist principles, she painted a series of works reversing stereotypical artistic themes by featuring nude men in poses that were traditionally associated with women.
0,1,0	- After they finish the upper Secondary program, students may choose to attend a Tertiary school or continue their apprenticeship. - During the 2010-11 school year, there were a total of 18 students attending one class in Surpierre. - Later that month he scored the winning penalty as Rovers put Premiership Everton out of the League Cup. - His time with Rovers was disrupted by injuries, and he left in September 2002 when his contract was terminated for 'gross misconduct' involving an 'incident at a gym'.
1,1,1	- L'Isle-Adam is a commune and town in north central Val-d'Oise. - Prior to entering Parliament, he was the National Secretary of the Australian Workers' Union from 2001 to 2007. - The "Tangerines" were relegated out of the Second Division at the end of the 1977 - 78 season. - Desaster are a black/thrash metal band formed in Koblenz, Germany in 1988.

문장 사이의 세그먼트 후보 위치($seg_1 \sim seg_{n-1}$)에 대해 seg_1부터 시작하여 까지 seg_{n-1}이동하면서 좌우 각각 2개의 문장을 이용하여 해당 위치의 레이블을 결정하는 실험이다. Pk점수[11]는 TS분야 연구에서 주로 사용되는 평가 도구로 segment가 적절히 나누어 졌는지를 평가한다. 숫자가 낮을수록 더 좋은 TS 결과임을 나타낸다. Fig. 6~8은 각 모델에서 threshold에 따른 Pk 점수의 평균을 나타낸 표이다.

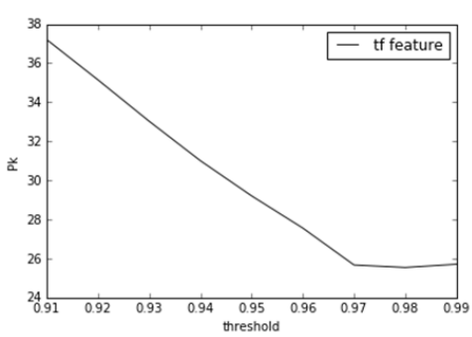


Fig. 6. Pk Score of $M_{\text{base}_{\text{tf}}}$ by Threshold

Fig. 6은 M_{base}모델에서 문장 표현 자질을 tf로 사용한 $M_{\text{base}_{\text{tf}}}$모델을 실험의 결과이다. $M_{\text{base}_{\text{tf}}}$는 문장 표현 자질로 tf 점수를 이용한다. 그렇기 때문에 Cosine Distance 값이 0~1사이로 표현된다. 전체적으로 문장들간의 거리가 멀리 분포하였다. 그래서 다른 모델들과 다르게 threshold를 0.9~0.99 사이로 하였을 때 유의미한 분리 결과가 나타났다. threshold를 0.98로 설정하였을 때, (즉 인접한 두 문장 블록간의 Cosine Distance가 0.98 이상 일 때를 두 블록간의 관계가 없다고 판단하였을 때) Pk 점수가 25.54로 가장 높게 나타났다.

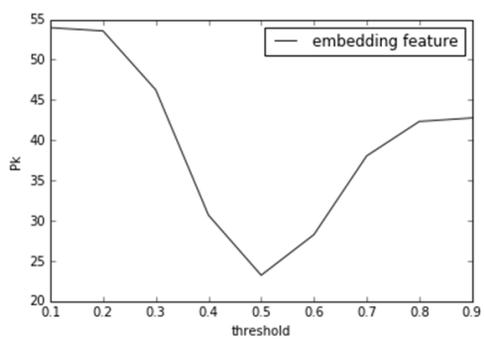


Fig. 7. Pk Score of $M_{\text{base}_{\text{emb}}}$ by Threshold

Fig. 7은 M_{base}모델의 문장 표현 자질로 각 단어의 임베딩[12] 벡터 평균을 사용한 모델($M_{\text{base}_{\text{emb}}}$)의 결과이다. 이 모델은 threshold를 0.5로 설정하였을 때 가장 좋은 23.22의 Pk 점수를 나타내었다. Threshold를 더 낮추면 주제 분리가 아닌 위치에서도 주제가 분리되었다고 설정하게 되어 오류율이 높아지고, 높이면 주제가 분리되어야 하는 곳을 분리하지 못하여 오류율이 높아진 것으로 분석된다.

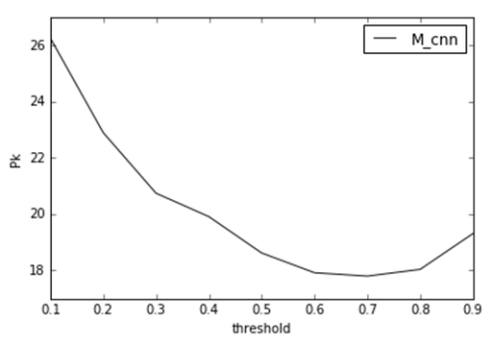


Fig. 8. Pk Score of M_{CNN} by Threshold

Fig. 8은 M_{CNN}모델을 적용하였을 때 실험 결과이다. 이 모델은 threshold를 0.7로 설정하였을 때 가장 높은 17.8의 Pk점수를 나타내었다.

전반적으로, M_{CNN} 모델을 사용하였을 때, 다른 모델들보다 더 나은 성능을 나타내었다. tf자질을 사용하였을 때, 두 블록간에 같은 단어를 공유하지 않는다면 Cosine Distance 값이 낮아지게 된다. 비슷한 의미의 단어가 사용되었다 하더라도 단어가 다르면 그 의미를 반영할 수 없다. 반면, 단어 임베딩을 사용 할 경우, 같은 단어가 아니더라도, 비슷한 의미의 단어가 사용되었다면 그 영향력이 어느 정도 발휘 될 수 있기 때문에 동일한 단어가 사용되지 않았다 하더라도 비슷한 단어가 사용되었다면 의미상의 거리 계산에 영향을 줄 수 있다. 그렇기 때문에 단어 임베딩 벡터를 사용한 모델들이 tf자질을 사용한 모델보다 On-line 상황을 가정한 실험에서 더 나은 성능을 나타낸 것으로 판단된다.

단어 임베딩을 단순히 거리를 비교하는 것 보다, 문장들 사이의 관계를 표현할 수 있는 모델을 학습시키고 이를 이용하여 주제 변화 여부를 판별하는 것이 좀더 나은 주제 변화를 예측 할 수 있는 것으로 나타났다. 이는 Cosine Distance 계산으로는 나타낼 수 없는 문장들 간의 관계를 모델이 학습 할 수 있기 때문에 가능한 것으로 생각된다. 아래 Table 3은 각 모델 별 최적화된 threshold에 대한 test data 각 그룹별 Pk 점수와 평균을 나타낸 표이다. Choi 데이터의 모든 유형에서 모델이 가장 좋은 성능을 나타내었다.

Table 4. Best Pk Score of Each Model

Type	$M_{\text{base}_{\text{tf}}}$ (0.98)	$M_{\text{base}_{\text{emb}}}$ (0.5)	M_{CNN} (0.7)
3-11	26.5	22.6	17.22
3-5	20.47	22.49	17.62
6-8	25.4	24.09	17.62
9-11	29.8	23.72	18.73
Avg	25.54	23.8	17.8

2) C99 with M_{CNN}

앞선 실험에서 M_{CNN}이 M_{base}모델보다 더 나은 Pk 점수를 나타내었다. 하지만 기존의 전역 최적화 알고리즘과 비교하여 보면 특별히 더 나은 성능으로 보기 어려울 수 있다. On-line 주제 분리를 위하여 지역 정보만을 이용할 수 밖에 없는 한계 때문에, 모델이 가지고 있는 문장간의 관계 계산 능력이 희생된 것으로 판단된다.

이 장에서는 M_{CNN}모델을 C99알고리즘의 문장 유사도 계산에 사용한 실험을 수행하였다. 이를 통해 M_{CNN}모델이 문장간 관계 계산에 충분한 능력이 있음을 보이고, On-line 상황에서뿐만 아니라 기존의 전역 최적화 알고리즘과 함께 활용될 수 있음을 보이고자 한다.

이 실험을 위해 구현한 C99알고리즘 결과(Table 5의 C99 with tf)와 C99알고리즘의 유사도 매트릭스 계산을 위한 자질을 단어 빈도수 에서 단어 임베딩 벡터로 바꾼 모델(Table 5의 C99 with emb) 그리고 유사도 매트릭스 값 계산에 모델의 관계점수를 이용하여 사용한 모델(Table 5의 C99 with M_{CNN})의 Pk 점수를 나타내었다.

M_{CNN}모델의 관계 점수는 관계가 있는 경우 0, 없는 경우 1을 나타낸다. C99알고리즘에서 사용하는 Cosine 유사도는 관계가 있을 경우 1, 없을 경우 0의 값을 갖는다. 그렇기 때문에 이 실험에서는 1 - $Score_p$를 사용하여 M_{CNN}을 기존의 C99알고리즘에 적용하였다. C99의 Rank matrix 계산을 위한 필터 크기는 11을 사용하였고, 세그먼트의 개수는 특정하지 않고 C99의 알고리즘을 통해 구하도록 하였다.

Table 5. Results of C99 with Model

Type	C99 with tf	C99 with emb	C99 with M_{CNN}
3-11	14.23	17.36	13.52
3-5	14.95	15.82	13.16
6-8	11.13	13.33	9.09
9-11	11.8	14.45	12.05
Avg	13.03	15.24	11.95

Table 5의 결과를 살펴보면, C99에 M_{CNN}모델을 적용한 결과가 가장 좋은 Pk 점수를 나타내었다. 4.4.1의 실험에서 On-line 상황에서 M_{CNN}모델의 Pk점수는 17.8 이었다. C99 알고리즘의 유사도 매트릭스 계산에 M_{CNN}을 적용하였을 때 11.95로 약 6정도의 Pk 점수 상승이 있다. On-line에서는 단순히 좌우의 정보만을 이용하지만, C99의 경우 다른 위치의 문장과의 관계도 함께 고려한다. 이때 M_{CNN}이 문장들 간의 관계를 더 정확하게 측정하였기 때문에, 다른 알고리즘 보다 더 높은 성능을 나타내었다고 볼 수 있다.

Table 6은 C99을 확장한 [8]의 Choi 데이터에 대한 Pk 점수이다. 문장에 적용한 전처리 과정과 적용한 단어 임베딩이 달라 직접적인 비교는 어렵지만, C99을 기반으로 하는 연구들의 대략적인 성능을 알 수 있다.

Table 6. Results of Other Researches

Type	oC99	oC99tf	oC99tfidf
3-11	15.56	14.91	14.78
3-5	14.22	12.14	10.27
6-8	12.20	13.17	12.23
9-11	11.59	14.60	15.87
Avg	13.40	13.7	13.29

Table 6의 oC99은 실험을 위해 구현한 [8]의 C99알고리즘 결과이다. oC99tf는 문장을 표현하는 벡터를 단어들의 임베딩 벡터 합으로 나타내어 C99에 적용한 결과이다. oC99tfidf는 각 단어의 벡터 결합 가중치를 조정하여 실험한 결과이다. 이 연구에서는 단어 임베딩으로 Glove[9] 알고리즘을 사용하였다. 본 연구와 비교하였을 때, 기본 C99 구현의 성능은 비슷하나 본 연구의 구현이 좀더 높게 나타났다. 이는 C99알고리즘이 문서의 전처리에 민감하게 반응하기 때문에 같은 알고리즘에서도 다른 성능을 나타낼 수 있기 때문이다[8]. 단어 임베딩을 사용한 결과에서는 본 논문의 결과(C99 with emb)가 [8]의 연구보다 낮은 성능을 나타내었다. 전반적으로, M_{CNN}을 적용한 모델이 기존의 모델과 비교할 만한 결과를 나타냄을 알 수 있다.

5. 결 론

본 논문에서는 문장이나 대화가 진행 중인 상황에서 주제 변화를 인식할 수 있는 모델에 대해 제안하였다. 이를 위해 문장이나 대화의 진행 상황에서 알 수 있는 가장 최근의 문장 또는 발화 정보를 이용하여 주제 분류를 수행하였다. 이들 문장을 표현하는 방법으로 단어 빈도수 또는 단어 임베딩 벡터를 활용하고, 문장 사이의 거리를 통해 주제 분리를 결정하는 모델을 제안하였다. 또한 합성곱 신경망 모델을 이용하여 주제 분리 양쪽의 정보를 학습하고 이를 토대로 주제 분리 여부를 결정하는 모델에 대해 제안하였다. 그 결과, On-line 상황을 가정한 실험에서 합성곱 신경망을 이용한 모델이 Pk 점수 17.8을 나타냈다. 또한 전역 최적화를 수행하는 C99 알고리즘의 문장간 유사도 계산에 모델을 적용하여 11.95의 Pk점수를 나타냈다. 이를 통해 On-line 상황에서의 주제 분리 문제에 본 논문에서 제안 하는 모델의 적용 가능성을 확인하였다.

본 논문의 연구성과를 활용하면 음성인식이나 대화형 시스템, 작문 시스템 등에서 사용자의 주제 변화를 추적하고 그에 맞게 대응하는 반응형 시스템이 가능하다. 이를 통해 좀더 지능적인 시스템 개발을 기대할 수 있다. 이러한 연구를 위해서는 On-line 상황에서의 주제 변화 인식 성능을 좀 더 향상시킬 필요가 있다. 이러한 연구를 향후 계속 진행해 나갈 계획이다.

References

[1] Jeffrey C. Reynar, "Topic segmentation: Algorithms and applications," IRCS Technical Reports Series, p.66, 1998.

[2] Marti A. Hearst, "TextTiling: Segmenting text into multi-paragraph subtopic passages," *Computational Linguistics*, Vol.23, No.1, pp.33-64, 1997.

[3] Freddy Y. Y. Choi, "Advances in domain independent linear text segmentation," *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, Association for Computational Linguistics, 2000.

[4] Stanley F. Chen, Kristie Seymore, and Ronald Rosenfeld, "Topic adaptation for language modeling using unnormalized exponential models," *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on, IEEE*, Vol.2. 1998.

[5] M. Galley, K. McKeown, E. Fosler-Lussier, and H. Jing, "Discourse segmentation of multi-party conversation," in *Proc 41st ACL '03*, Vol.1, pp.562-569, 2003.

[6] Martin Riedl, and Chris Biemann, "TopicTiling: a text segmentation algorithm based on LDA," *Proceedings of ACL 2012 Student Research Workshop*, Association for Computational Linguistics, 2012.

[7] Martin Riedl, and Chris Biemann, "Text segmentation with topic models," *Journal for Language Technology and Computational Linguistics*, Vol.27, No.1, pp.47-69, 2012.

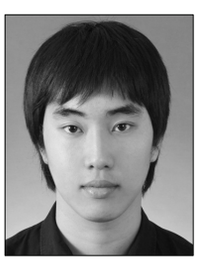
[8] Alexander A. Alemi and Paul Ginsparg, "Text Segmentation based on Semantic Word Embeddings," *arXiv preprint arXiv:1503.05543,* 2015.

[9] Jeffrey Pennington, Richard Socher, and Christopher D. Manning, "Glove: Global Vectors for Word Representation," *EMNLP*, Vol.14. 2014.

[10] Masao Utiyama and Hitoshi Isahara, "A statistical model for domain-independent text segmentation," *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, 2001.

[11] Doug Beeferman, Adam Berger, and John Lafferty, "Statistical models for text segmentation," *Machine Learning*, Vol.34, No.1-3, pp.177-210, 1999.

[12] Tomas Mikolov, et al., "Efficient estimation of word representations in vector space," *arXiv Preprint arXiv: 1301.3781*, 2013.



이 경 호

e-mail : gyholee@gmail.com

2011년 충남대학교 정보통신공학과(학사)

2013년 충남대학교 정보통신공학과(석사)

2013년～현 재 충남대학교 전자전파 정보통신공학과 박사과정

관심분야 : 자연언어처리, 기계학습, 인공지능

이 공 주

e-mail : kjoolee@cnu.ac.kr

1992년 서강대학교 전자계산학과(학사)
1994년 한국과학기술원 전산학과 (공학석사)
1998년 한국과학기술원 전산학과 (공학박사)
1998년~2003년 한국마이크로소프트(유) 연구원
2003년 이화여자대학교 컴퓨터학과 대우전임강사
2004년 경인여자대학 전산정보과 전임강사
2005년~현 재 충남대학교 전파정보통신공학과 교수
관심분야 : 자연언어처리, 기계번역, 정보검색, 정보추출