RoBERTa catSeqE: 개체 연결을 이용한 RoBERTa기반 키워드 추출

Cognitive Computing Lab.
이정두, 나승훈
전북대학교
bo0od12@naver.com, nash@jbnu.ac.kr

Ⅰ. 서론

키워드 추출(Keyphrase Extraction)이란 각 문서에서 내용과 주제를 포괄하는 핵심 단어 또는 구문을 추출하는 것을 말한다. 이는 문서의 핵심 내용을 짧은 구나 단어로 유추할 수 있기 때문에 매우 중요하다.

최근 뉴스, 블로그 등 실시간으로 생성되는 대량의 데이터를 이용하여 정보를 추출하는 기술이 큰 주목을 받고 있으며, 키워드 추출은 뉴스에서 중요한 정보를 추출하는 데 매우 중요한 역할을 한다.

문서

*프랑스 남부*에서 14일 현지시간 *스쿨버스와 열차가 충돌*하는 사고로 *어린이 4명이 숨졌다*고 르 피가로 등 프랑스 언론들이 전했다. 경찰은 이날 오후 4시께 프랑스 남부 페르피냥에서 서쪽으로 18㎞ 떨어진 소도시 미야의 한 철도 건널목에서 TER 열차와 통학버스가 충돌했다고 밝혔다. 이 사고로 버스에 타고 있던 어린이 4명이 숨지고 20여 명이 다쳤다. 다친 어린이 중 7명은 중상이다. 총리가 현장으로 가고 있다고 르 피가로는 전했다.

→

키워드

프랑스 남부
스쿨버스와 열차 충돌
어린이 4명이 숨졌다

Ⅱ. 제안 방법

Dataset

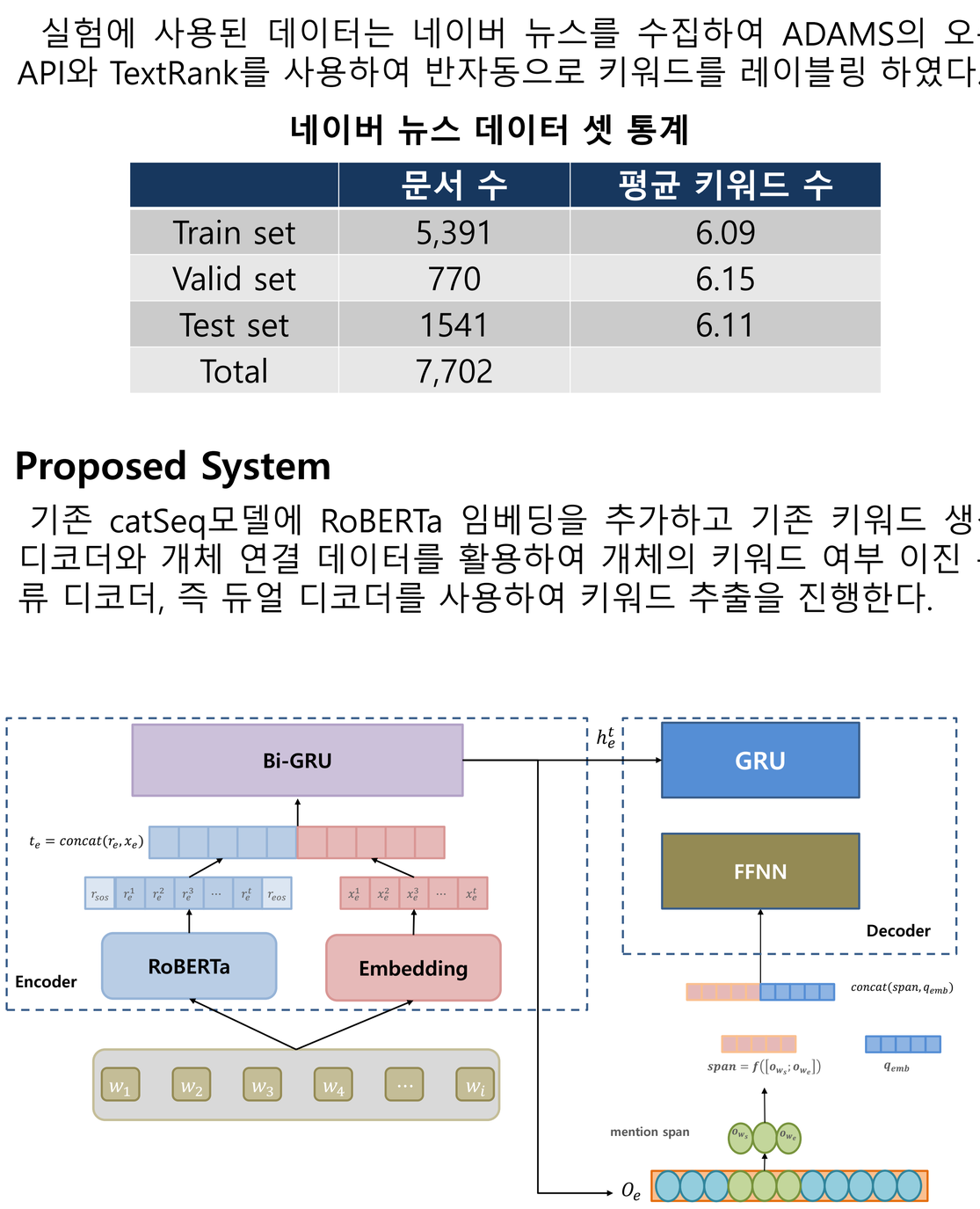
실험에 사용된 데이터는 네이버 뉴스를 수집하여 ADAMS의 오픈 API와 TextRank를 사용하여 반자동으로 키워드를 레이블링 하였다.

네이버 뉴스 데이터 셋 통계

	문서 수	평균 키워드 수
Train set	5,391	6.09
Valid set	770	6.15
Test set	1541	6.11
Total	7,702	

Proposed System

기존 catSeq모델에 RoBERTa 임베딩을 추가하고 기존 키워드 생성 디코더와 개체 연결 데이터를 활용하여 개체의 키워드 여부 이진 분류 디코더, 즉 듀얼 디코더를 사용하여 키워드 추출을 진행한다.



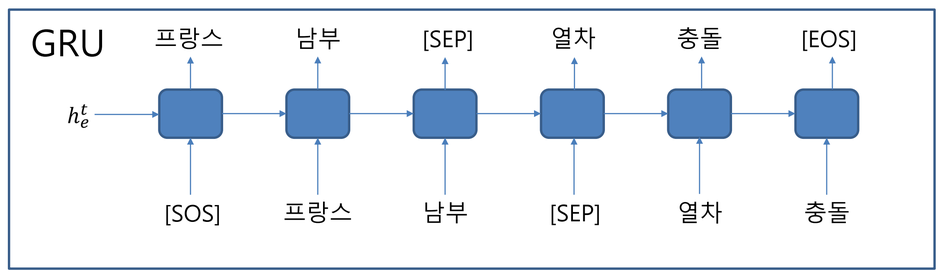
RoBERTa catSeqE 모델 구조

i. 인코더

주어진 문서를 형태소 분석한 후 단어(W_i)를 RoBERTa와 단어 임베딩의 입력값으로 넣어준다. RoBERTa의 첫번째와 마지막 토큰을 제외한 임베딩 값(r_e)과 단어 임베딩 값(x_e)을 결합한다.
결합한 값(t_e)을 양방향 GRU의 입력값으로 넣어 문서를 표현하는 h_e^t(양방향 GRU의 마지막 히든 상태)와 각 단어를 표현하는 O_e(각 타임스탭 마다의 GRU 아웃풋) = ($o_e^1, o_e^2, \dots, o_e^t$) 을 각각의 디코더로 전달한다.

ii. 디코더

a. 키워드 구문 생성 디코더



키워드 구분 생성 디코더는 기존 catSeq 모델과 같이 구분자를 사용하여 여러 키워드를 한번에 생성한다. 이때 단방향 GRU를 사용하고 $h_d^0 = h_e^t$ 이다.

b. 개체 이진 분류 디코더

개체의 단어 표상 E_i는 다음과 같다.

$$E_i = concat(span, q_{emb})$$
$$span = concat(o_{w_s}, o_{w_e})$$

여기서 o_{w_s}는 개체의 mention span의 시작 단어에 해당하는 인코더의 양방향 GRU 아웃풋이고, o_{w_e}는 개체의 mention span의 마지막 단어에 해당하는 인코더의 양방향 GRU 아웃풋이다.
그리고 q_{emb}는 개체를 설명하는 문서(description)의 임베딩 값이다. 이는 문서에 대해 개체 연결 시에 얻는 값이다.

Ⅲ. 실험 결과

실험 평가는 Top-k개의 키워드가 아닌 생성된 모든 키워드를 두고 평가한다. 이때 중복 생성된 키워드는 제거한다.

모델	Precision	Recall	F1
catSeq	41.22%	34.69%	37.67%
RoBERTa catSeq	59.65%	49.89%	54.34%
RoBERTa catSeq +Entity classification	59.60%	54.09%	56.71%
RoBERTa catSeqE	60.06%	53.94%	56.83%

각 모델의 실험 방법을 요약하면 다음과 같다.

- catSeq: CopyRNN 모델을 베이스로 구분자를 사용하여 여러 키워드를 한번에 생성한다. 그 외 직교 정규화와 시멘틱 보전은 사용하지 않는다.
- RoBERTa catSeq: 언어 모델 RoBERTa를 사용하여 임베딩을 추가적으로 적용하고 키워드 추출에 알맞게 미세 조정한다.
- RoBERTa catSeq + Entity classification: 문서 내 개체의 키워드 여부를 분류하여 키워드라고 예측된 개체와 기존 RoBERTa catSeq를 통해 생성된 키워드를 결합한다.
- RoBERTa catSeqE: RoBERTa catSeq 모델과 Entity classification 모델을 통합한 것으로 이는 Ⅱ에 나와 있다.

평가 결과 베이스라인의 catSeq보다 RoBERTa를 추가했을 경우 큰 성능향상이 있었다. 이후 개체 이진 분류 결과를 후처리로 더해준 결과 2.37%가 향상 되었고 두 모델을 통합한 결과 앞선 후처리 결과보다 0.12% 성능향상이 있었다.

향후 개체 연결 정보를 이용해 각 문장의 트리플 구조를 얻고 지식 그래프를 형성한 다음 이를 그래프 인코더와 문서 인코더, 즉 듀얼 인코더를 사용해 더욱 정교한 키워드 추출 모델을 연구할 것이다.