# RoBERTa-catseqKG
# 지식그래프를 이용한 RoBERTa 기반 키워드 추출

*이정두° • 나승훈(전북대)*

# 목 차



# 서 론

- **키워드 추출(Keyphrase Extract)**
  - 기계가 문서를 이해하고 문서의 핵심 단어를 추출하는 태스크

**트럼프 "북한은 인류의 문제...시진핑과 북한 문제 논의할 것"**

도널드 트럼프 미국 대통령이 4일(현지시간) 미·중 정상회담에서 북한의 핵·미사일 문제 해결을 주요 의제로 삼아 대화하겠다고 밝혔다.
트럼프 대통령이 오는 6~7일 미국 플로리다에서 열리는 시진핑(習近平) 중국 국가 주석과의 정상회담 테이블에 북한 문제를 올리겠다고 공식으로 언급한 것은 이번이 처음이다.
트럼프 대통령은 이날 백악관에서 열린 미 최고경영자(CEO) 대상 타운홀 미팅에 참석해 "시 주석과 저는 당연히 북한을 포함해 여러 현안에 대해 논의할 것"이라고 말했다. 그는 특히 "북한은 문제이다. 정말 인류의 문제이다. 그 점에 대해 논의할 것"이라고 강조했다.
이에 따라 트럼프 대통령은 북한의 6차 핵실험 가능성이 커지는 등 안보 위협이 최고조에 달한 가운데 열리는 정상회담에서 북한의 핵·미사일 문제 해결을 위한 중국의 적극적인 역할을 강하게 요구할 것으로 보인다. 트럼프 대통령은 지난 2일 영국 일간 파이낸셜타임스(FT) 인터뷰에서도 "중국은 북한에 엄청난 영향력을 가졌고 우리를 도와 북한 문제를 다룰지 말지 결정할 것"이라며 "만약 중국이 그렇게 한다면 중국에 좋을 것이고, 그렇게 하지 않는다면 누구에게도 좋지 않을 것"이라고 압박을 가한 바 있다. ...

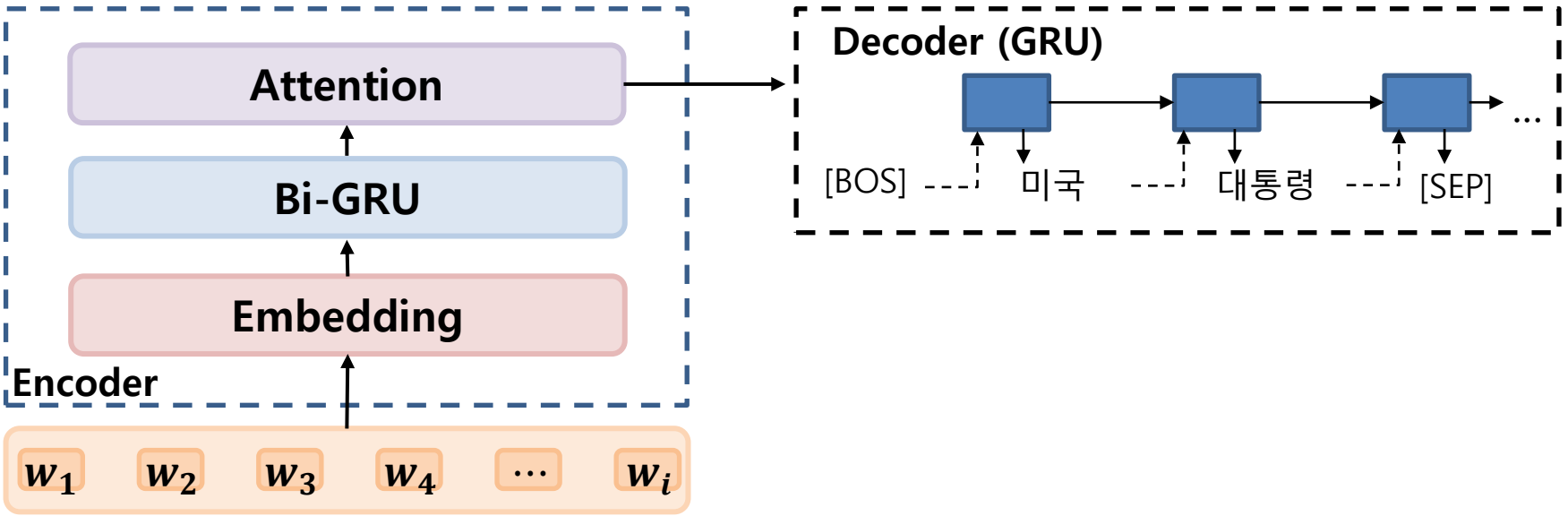
⬇

트럼프 대통령/북한의 핵 미사일/시진핑/북한 문제/논의

뉴스 출처: https://www.yna.co.kr/view/AKR20170405007900071?input=1195m

# 관련 연구

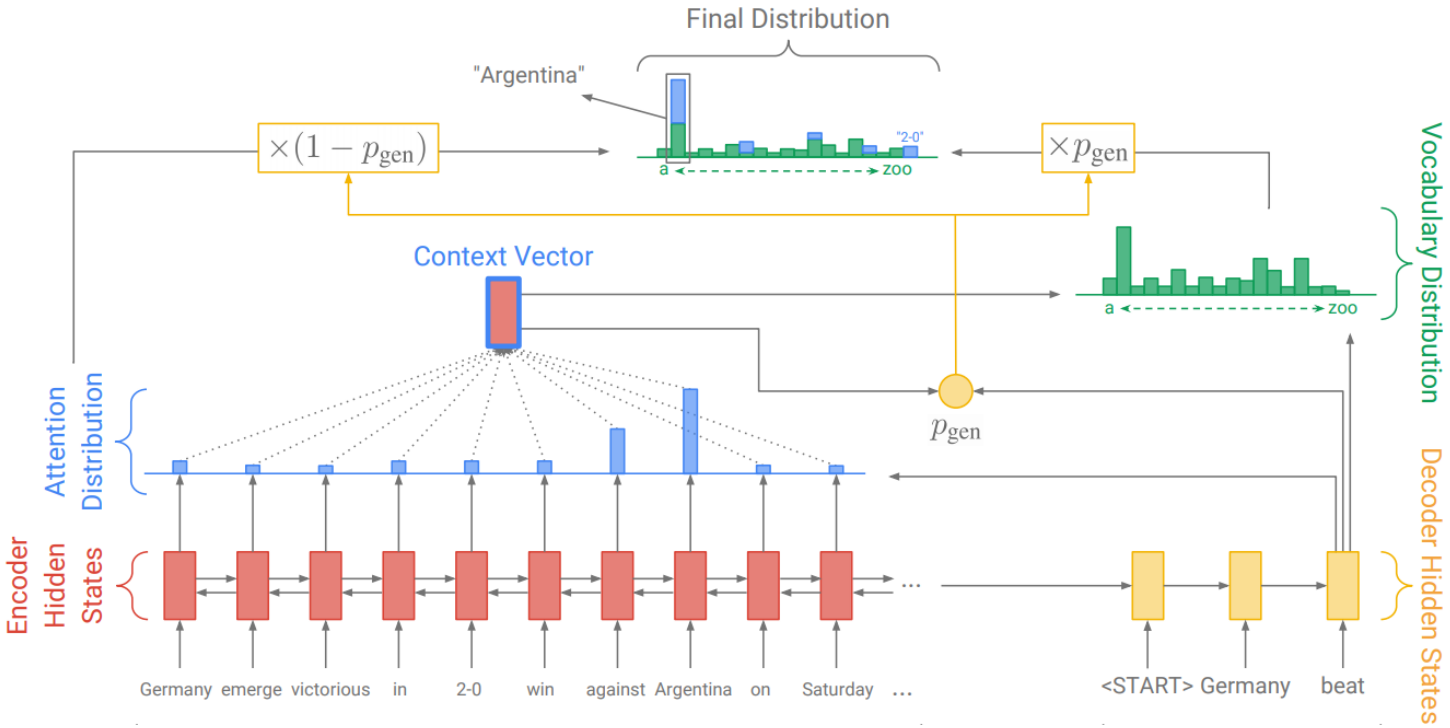
- **One Size Does Not Fit All: Generating and Evaluating Variable Number of Keyphrases[Yuan, Xingdi, et al, 2018]**
  - 한번에 여러 키워드를 추출하기 위해 구분자를 사용한 모델




**catSeq 모델 구조**

# 관련 연구

- **Copy mechanism**
  - 문서에만 나타나는 새로운 단어도 생성 가능




출처: See, Abigail, Peter J. Liu, and Christopher D. Manning.
"Get to the point: Summarization with pointer-generator networks." *arXiv preprint arXiv:1704.04368* (2017).

# 제안 모델

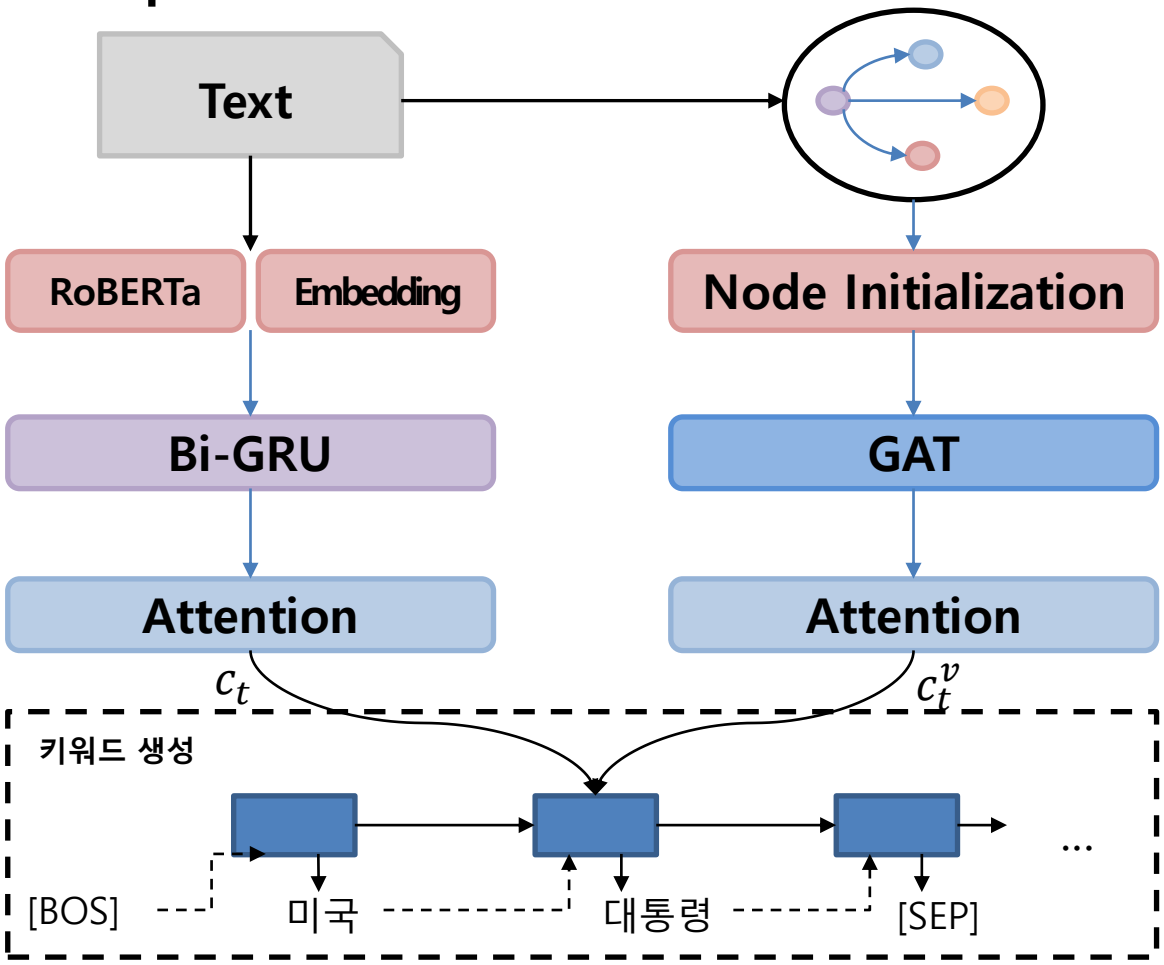
- **RoBERTa-catSeqKG: 지식그래프를 이용한 RoBERTa 기반 키워드 추출**
  **- KG를 활용하여 동일 개체에 대해 문서 여러 곳에 나타나는 복잡한 이벤트를 반영**

**트럼프 "북한은 인류의 문제...시진핑과 북한 문제 논의할 것"**

도널드 트럼프_1 미국 대통령이 4일(현지시간) 미·중 정상회담에서 북한_1의 핵·미사일 문제 해결을 주요 의제로 삼아 대화하겠다고 밝혔다.
트럼프_2 대통령이 오는 6~7일 미국 플로리다에서 열리는 시진핑(習近平) 중국 국가 주석과의 정상회담 테이블에 북한_2 문제를 올리겠다고 공식으로 언급한 것은 이번이 처음이다.
트럼프_3 대통령은 이날 백악관에서 열린 미 최고경영자(CEO) 대상 타운홀 미팅에 참석해 "시 주석과 저는 당연히 북한_3을 포함해 여러 현안에 대해 논의할 것"이라고 말했다. 그는 특히 "북한_4은 문제이다. 정말 인류의 문제이다. 그 점에 대해 논의할 것"이라고 강조했다.
이에 따라 트럼프_4 대통령은 북한_5의 6차 핵실험 가능성이 커지는 등 안보 위협이 최고조에 달한 가운데 열리는 정상회담에서 북한_6의 핵·미사일 문제 해결을 위한 중국의 적극적인 역할을 강하게 요구할 것으로 보인다. 트럼프_5 대통령은 지난 2일 영국 일간 파이낸셜타임스(FT) 인터뷰에서도 "중국은 북한_7에 엄청난 영향력을 가졌고 우리를 도와 북한_8 문제를 다룰지 말지 결정할 것"이라며 "만약 중국이 그렇게 한다면 중국에 좋을 것이고, 그렇게 하지 않는다면 누구에게도 좋지 않을 것"이라고 압박을 가한 바 있다. ...

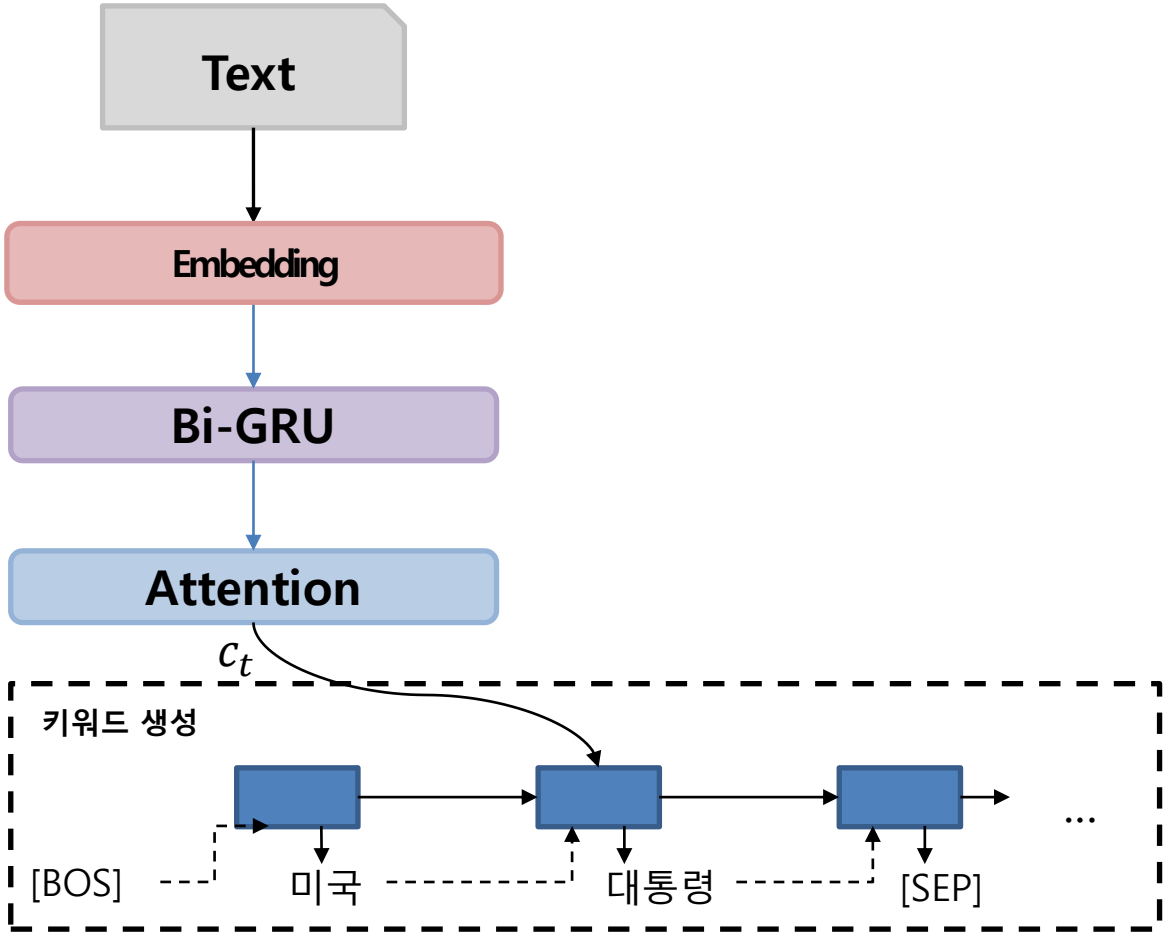
# 제안 모델

- **RoBERTa-catSeqKG 모델 구조**



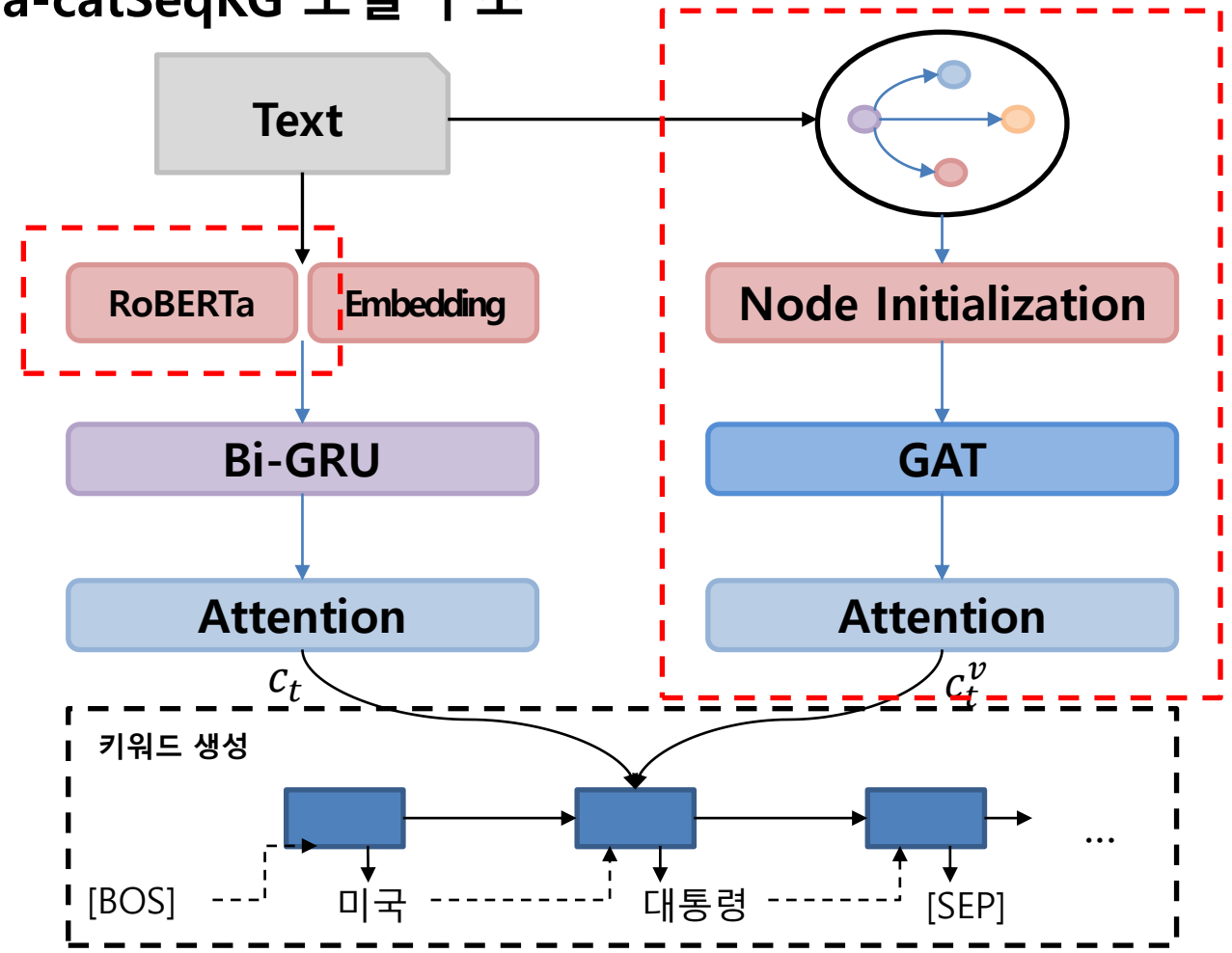

# 제안 모델

- **RoBERTa-catSeqKG 모델 구조**



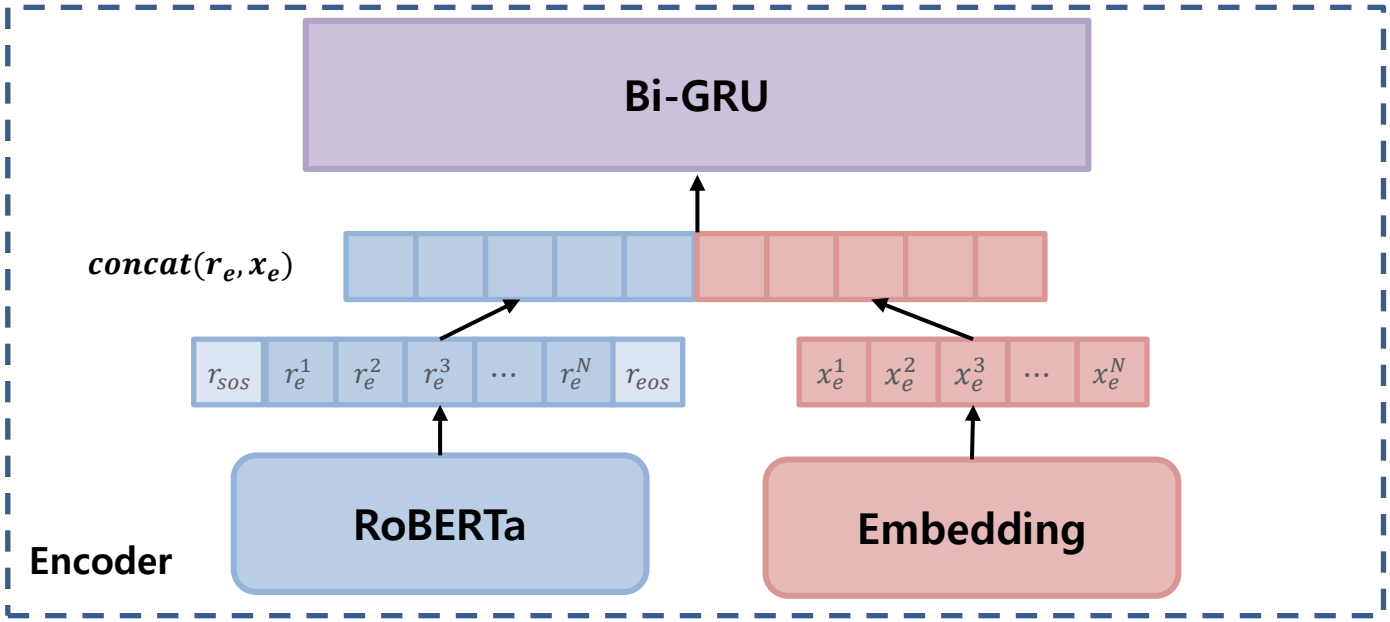

# 제안 모델

- **RoBERTa-catSeqKG 모델 구조**



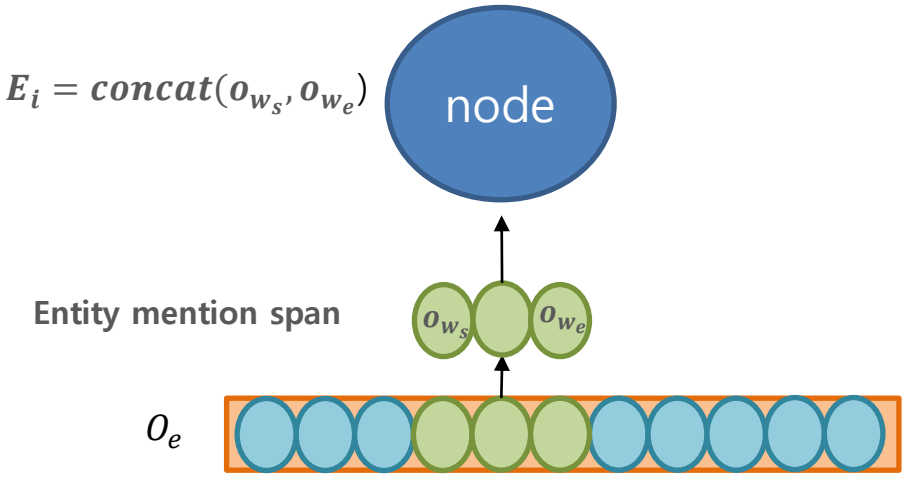

# 제안 모델

- **RoBERTa-catSeqKG**
  - **Text Embedding**



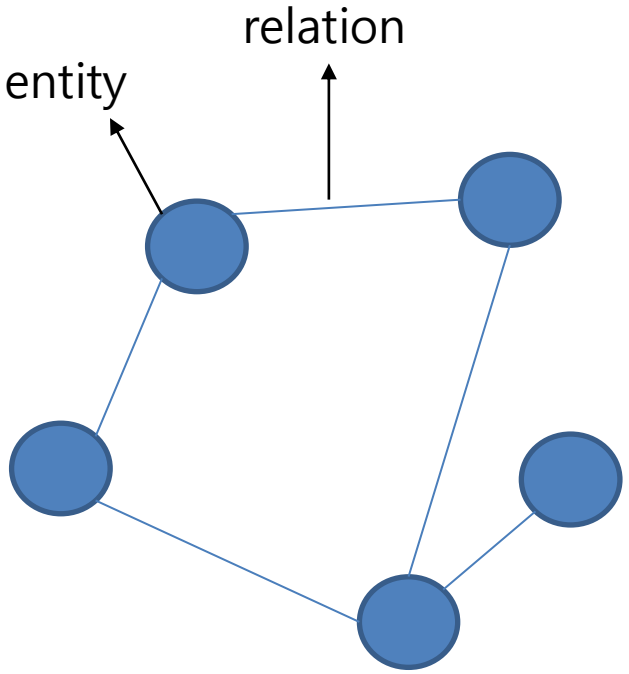

# 제안 모델

- **RoBERTa-catSeqKG**
  - **Node initialization**
    - triple 정보는 사전에 딥러닝 모델을 통해 생성




$O_e$ = *encoder Bi_GRU의 time step마다 output 집합*

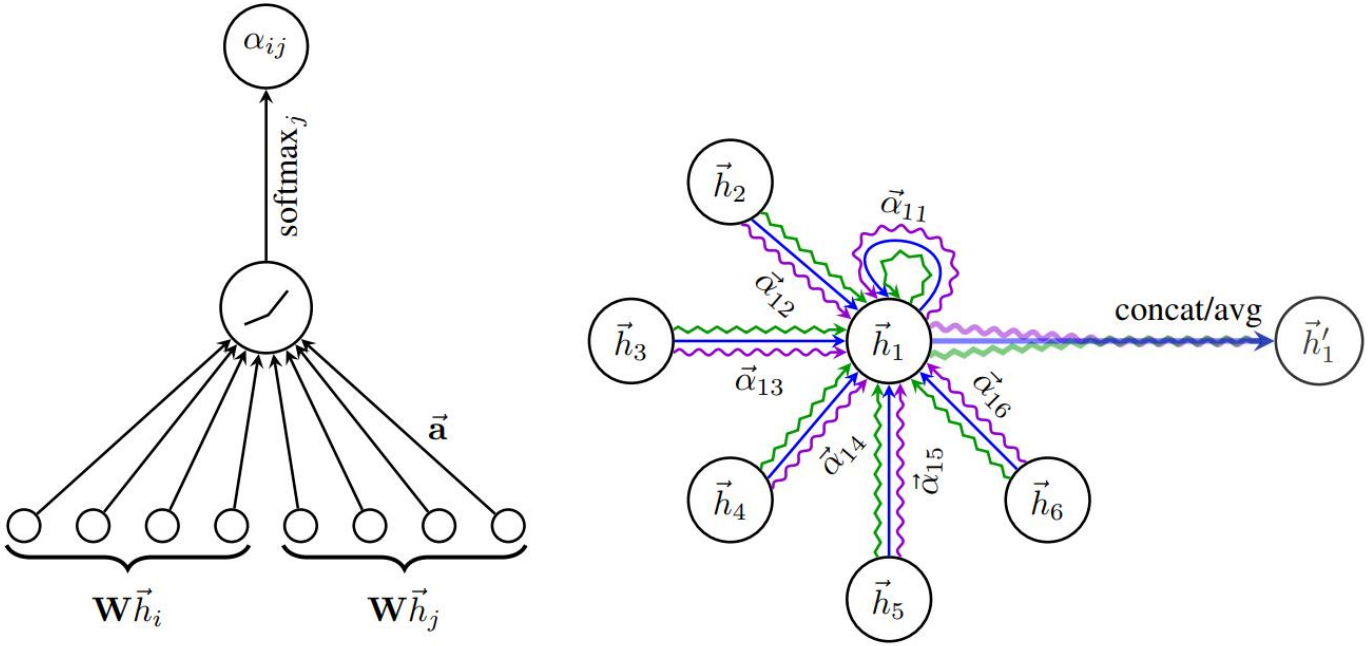
$E_i$ = *entity representation vector*




Knowledge Graph

# 제안 모델

- **RoBERTa-catSeqKG**
  - **Node update (GATs)**




출처: Veličković, Petar, et al. "Graph attention networks." *arXiv preprint arXiv:1710.10903* (2017).

# 제안 모델

- **RoBERTa-catSeqKG**
  - **Node update (GATs)**

$$E_v^l = \text{LayerNorm}\left(\Phi_{\text{k=1}}^{K} elu\left(\sum_{(r,u)\in\mathcal{N}_v} \alpha_{v,r,u}^k W^k E_u^{l-1}\right) + E_v^{l-1}\right)$$

$$\alpha_{v,r,u}^k = \frac{\exp\left(LeakyReLU\left(a^T\left[W^k E_u^{l-1} \Phi W^k\left(E_u^{l-1} + R_r\right)\right]\right)\right)}{\sum_{(r',u')\in\mathcal{N}_v} \exp\left(LeakyReLU\left(a^T\left[W^k E_u^{l-1} \Phi W^k\left(E_{u'}^{l-1} + R_{r'}\right)\right]\right)\right)}$$

$E_v$ = Subject node feature

$E_u$ = Object node feature

$R_r$ = Relation feature

$\Phi$ = concat

$W^k$, $a^T$ = Trainable parameter

# 제안 모델

- **RoBERTa-catSeqKG**

  - **Attention**

    - **Graph**

$$\boldsymbol{c}_t^v = \sum_i a_{i,t}^v \, E_i$$

$$a_{i,t}^v = softmax(u_0^T \tanh(W_1 s_t + W_2 E_i))$$

    - **Text**

$$\boldsymbol{c}_t = \sum_k a_{k,t} \, h_k$$

$$a_{k,t} = softmax(u_1^T \tanh(W_3 s_t + W_4 h_k))$$

  - **Final Distribution**

$$\boldsymbol{c} = \beta \boldsymbol{c}_t^v + (1-\beta)\boldsymbol{c}_t$$

$$\beta = sigmoid(W_5 s_t)$$

$$\boldsymbol{P}_{vocab} = softmax(W_3 \, concat(s_t, \boldsymbol{c}))$$

$$p_{gen} = \sigma\big(W_3 \, concat(s_t, \boldsymbol{c}, y_{t-1})\big)$$

$$\boldsymbol{P} = p_{gen} * \boldsymbol{P}_{vocab} + (1 - p_{gen})\boldsymbol{P}_{copy}$$

$E_i$ = Node feature

$\boldsymbol{c}_t^v$ = Node context vector

$\boldsymbol{c}_t$ = Text context vector

$W_*, u_*^T$ = Trainable parameter

$s_t$ = Decoder hidden state at time step t

$y_{t-1}$ = Word embedding at time step t-1

# 사용된 데이터 셋

- **네이버 뉴스 데이터**

| | 문서 수 | 평균 키워드 수 |
|---|---|---|
| Train set | 5,391 | 6.09 |
| Valid set | 770 | 6.15 |
| Test set | 1,541 | 6.11 |
| Total | 7,702 | - |

- **Entity 키워드 여부 통계**

| | 키워드 중 entity 비율 |
|---|---|
| Train set | 19.78% |
| Valid set | 19.12% |
| Test set | 19.50% |

# 실험 결과 및 향후 계획

- **RoBERTa_catSeqKG**

| Model | Precision | Recall | F1 |
|---|---|---|---|
| catSeq | 40.55% | 35.29% | 37.74% |
| RoBERTa_catSeq | 59.84% | 50.91% | 55.02% |
| RoBERTa_catSeqKG | **60.64%** | **52.17%** | **56.09%** |

- 향후 계획
  - Triple 정보를 얻을 때 사용된 딥러닝 모델 개선(entity linking, relation classification)
  - Absent keyword 생성 모델로 확장

# Q & A

# Thank You!