# 트위터 해시 태그를 이용한 End-to-end 뉴럴 모델 기반 키워드 추출

Young-Hoon Lee, Seung-Hoon Na

Cognitive Computing Lab.

Chonbuk National University

# 목차

- 서론
- 관련연구
- 모델 구조
- 학습/평가 데이터
- 실험결과

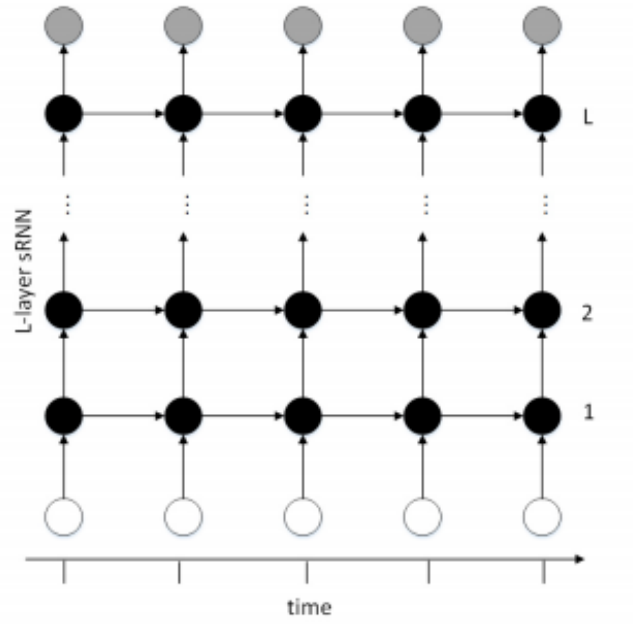
# 키워드 추출

- 문서 또는 문장의 내용을 함축하여 몇 개의 단어로 이루어지는 표현
- 트위터(Twitter)의 해시 태그는 주로 그 트윗(Tweet)의 키워드가 됨

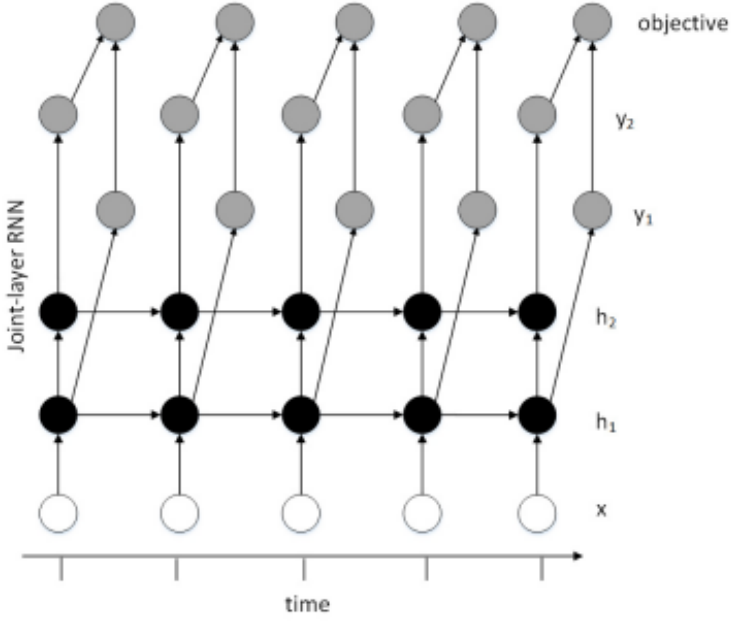
그룹 #방탄소년단 이 #한국 그룹 #최초 로 #미국 캘리포니아주 로스앤젤레스에서 열린 '2018 #아메리칸뮤직어워즈'에서 #수상 했습니다.

# Keyphrase Extraction Using Deep Recurent Neural Networks on Twitter, [Q. Zhang, 2016]

- Twitter의 트윗 정보를 이용하여 KeyPhrase를 시도하였음.
- 키워드 추출 작업을 수행하기 위해 Keyword와 Context 정보를 결합하는 새로운 joint-layer RNN (그림 b.) 모델을 제안.
- Seq2Seq 모델을 이용하여 Keyword Ranking Task와 Keyphrase Generation Task를 공동으로 처리




(a)




(b)

# Related Work

언어학적 특징 : A language-independent approach to keyphrase extraction and evaluation, [M. Paukkeri, 2008]
순위기반 : Automatic Construction and Ranking of Topical Keyphrases on Collections of Short Documents, [M. Danilevsky, 2014] , A ranking approach to keyphrase extraction, [X. Jiang, 2009]

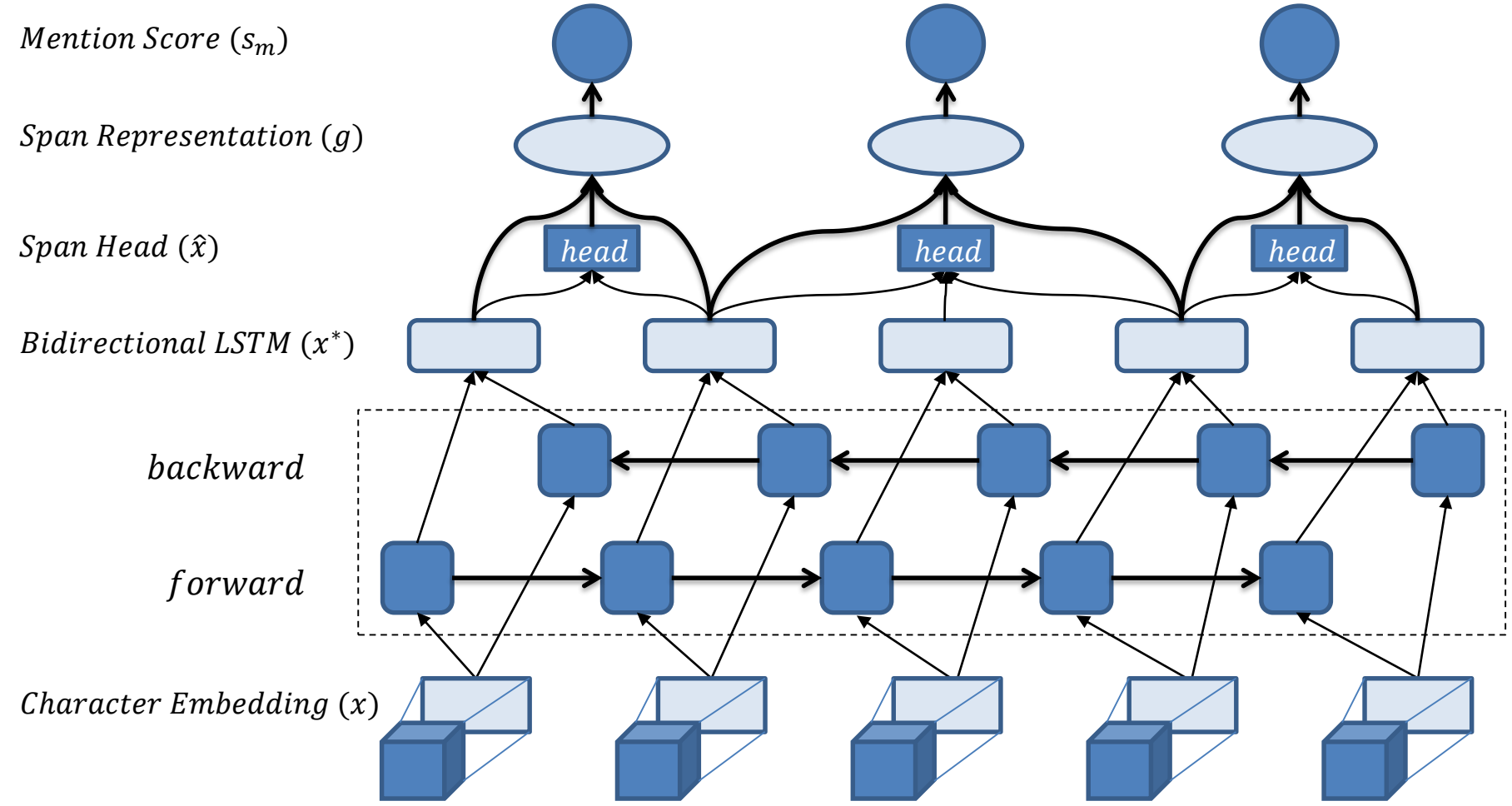
**TF-IDF를 Base로 한 한국어 키워드 추출 연구**

**TF-IDF의 변형을 이용한 전자뉴스에서의 키워드 추출 기법, [이성직, 2009]**

**TF-IDF와 소설 텍스트의 구조를 이용한 주제어 추출 연구 [유은순, 2015]**

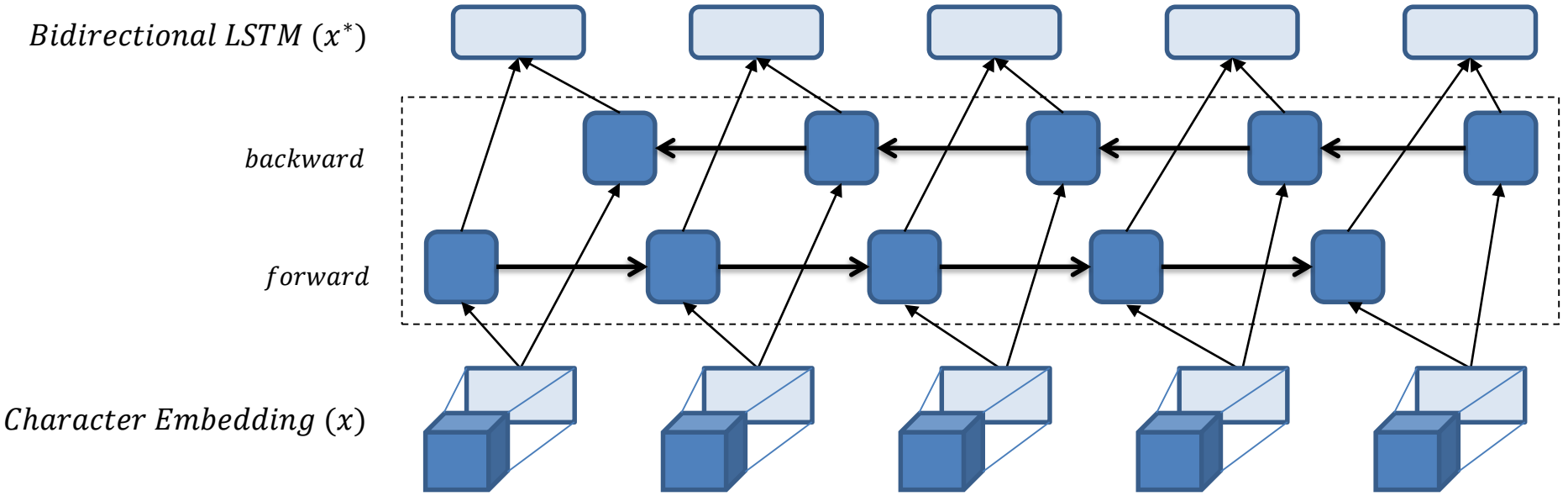
**TF-IDF 기반 키워드 추출에서의 의미적 요소 반영을 위한 결합벡터 제한, [박대서, 2018]**

# End-to-End 뉴럴 모델




End-to-end Neural Network Coreference Resolution K. Lee, 2017, **Mention Scoring Model**

# End-to-End 뉴럴 모델



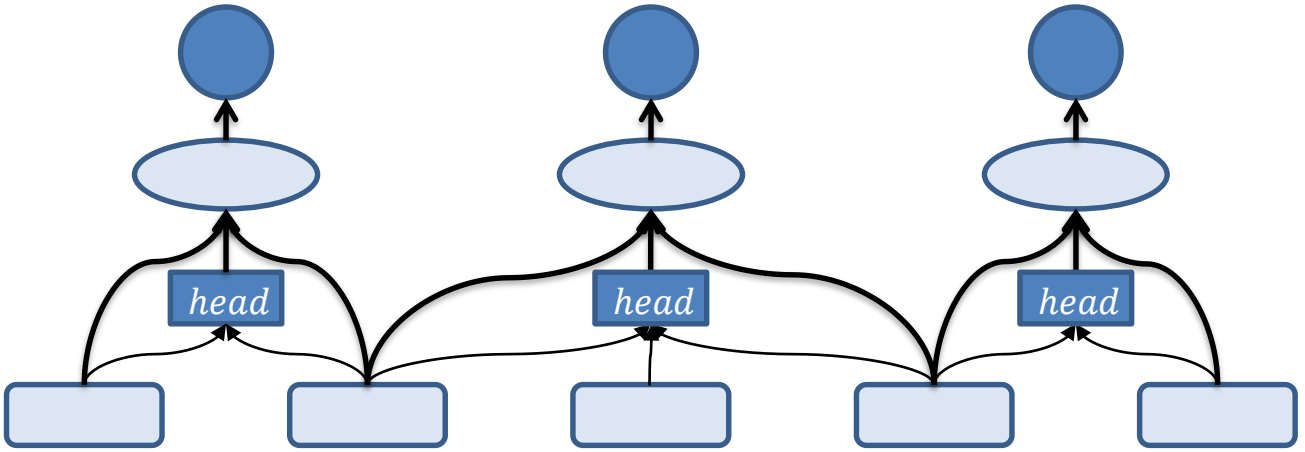

한국어에 적용하기 위해 Convolution과 Pooling 연산을 통해 Vector $x$를 얻어냄

$$\boldsymbol{x = MaxPooling(Conv1d(Embed(X)))}$$

Bi-LSTM을 이용하여 Forward/Backward concatenate하여 단어에 대한 문장 표현인 $x^*$를 얻음

$$\boldsymbol{x^* = \overleftrightarrow{LSTM}(x)}$$

# End-to-End 뉴럴 모델




키워드는 주로 짧은 단어들로 구성 ⇨ Span의 길이는 최대 5로 구성

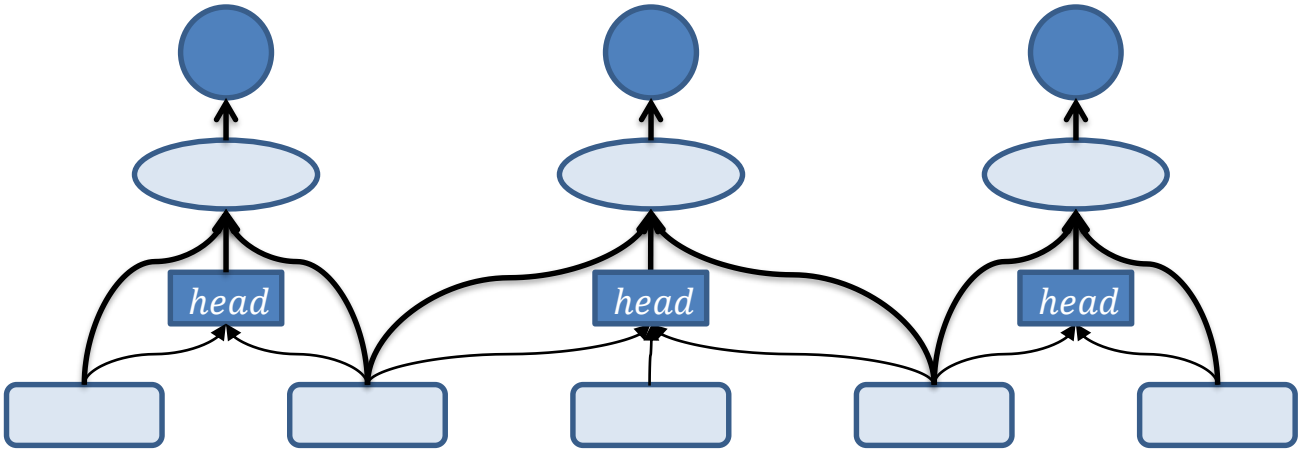
$$span = (START\ INDEX, END\ INDEX)$$

- **Span 표현 ($g$)**

$i$번째 Span표현

$$g_i = [x^*_{START(i)}, x^*_{END(i)}, \hat{x}_i, \varnothing(i)]$$

# End-to-End 뉴럴 모델




- **Span 표현 ($g$)**

$$\boldsymbol{g_i = [x^*_{START(i)}, x^*_{END(i)}, \hat{x}_i, \varnothing(i)]}$$

$x^*_{START(i)}$ : Span index의 $x^*$ 첫 번째 표현
$x^*_{END(i)}$ : Span index의 $x^*$ 마지막 표현
$\varnothing(i)$ : Span 길이 임베딩

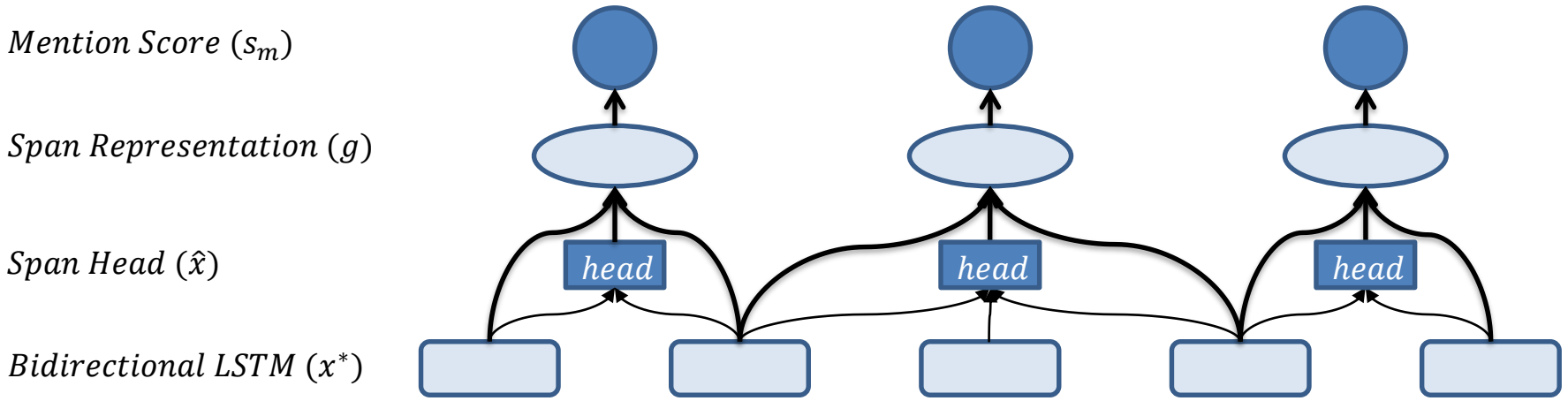
**Span head $\hat{x}_i$**

$$a_t = w_\alpha \cdot FFNN_\alpha(x^*_t)$$

$$a_{i,t} = \frac{\exp(\alpha_t)}{\sum_{k=START(i)}^{END(i)} \exp(x_t)}$$

$$\hat{x}_i = \sum_{k=START(i)}^{END(i)} a_{i,t} \cdot x_t$$

# End-to-End 뉴럴 모델




- **Keyword Score : 출력이 1인 FFNN의 입력으로 각 Span 표현인 $g_i$를 주어 Span들에 대한 점수를 구하고 상위 K개를 해당 문장의 키워드로 예측**

$$\boldsymbol{Score_i = FFNN(g_i)}$$

# 학습 / 평가 데이터 구성

- 해시 태그로 링크 되어 있는 Mention을 Keyword로 사용
- Mention이 포함된 문장을 Context로 사용
- 각 문장에는 하나 이상의 해시 태그가 존재
- 해시 태그의 Mention이 문장의 단어로 사용되는 데이터만을 사용

"#미국프로야구 #LA다저스 의 #류현진 이 뉴욕 메츠전에서 7이닝 무실점을 기록하며 시즌 #4승 을 달성했습니다."

**데이터에 포함 되는 트윗 예제**

피 말리는 '순위 경쟁 키' 쥔 류현진, 다음 상대는 로키스 #MLB #미국프로야구 # LA다저스

**데이터에 포함 되지 않는 트윗 예제**

# 실험 결과

- 랜덤 7:3의 비율, 학습 데이터 셋 : 9857개, 평가 데이터 셋 : 4231개
- 실험에 사용된 데이터 셋의 각 문장이 가지는 태그의 수가 1~10개까지 그 수가 모두 다르기 때문에 실험 평가에 어려움이 있음
- 각 Span의 점수를 비교하여 가장 높은 점수를 가지는 Span이 정답 키워드와 일치하면, 정답으로 평가하였다.

| 정확도 |
| --- |
| **73.29%** |

# 실험 결과

- **Span Score의 Top K에 대한 Precision과 Recall, F1-Score를 측정**

| K | Precision | Recall | F1 |
|---|---|---|---|
| 1 | 0.732 | 0.332 | 0.456 |
| **2** | **0.558** | **0.506** | **0.530** |
| 3 | 0.444 | 0.604 | 0.511 |
| 5 | 0.301 | 0.683 | 0.418 |

**실험 데이터 셋의 문장 당 평균 태그의 개수 : 2.84개**
**K=2일 때, F1-Score 0.53으로 가장 높은 점수를 보였다.**