Visual News: Benchmark and Challenges in News Image Captioning

Fuxiao Liu, Yinghan Wang, Tianlu Wang, Vicente Ordonez

2021/EMNLP

2024.02.20
이상민

목차

1. Introduction

2. Visual News dataset

3. Method

4. Experimental Results

5. Conclusion

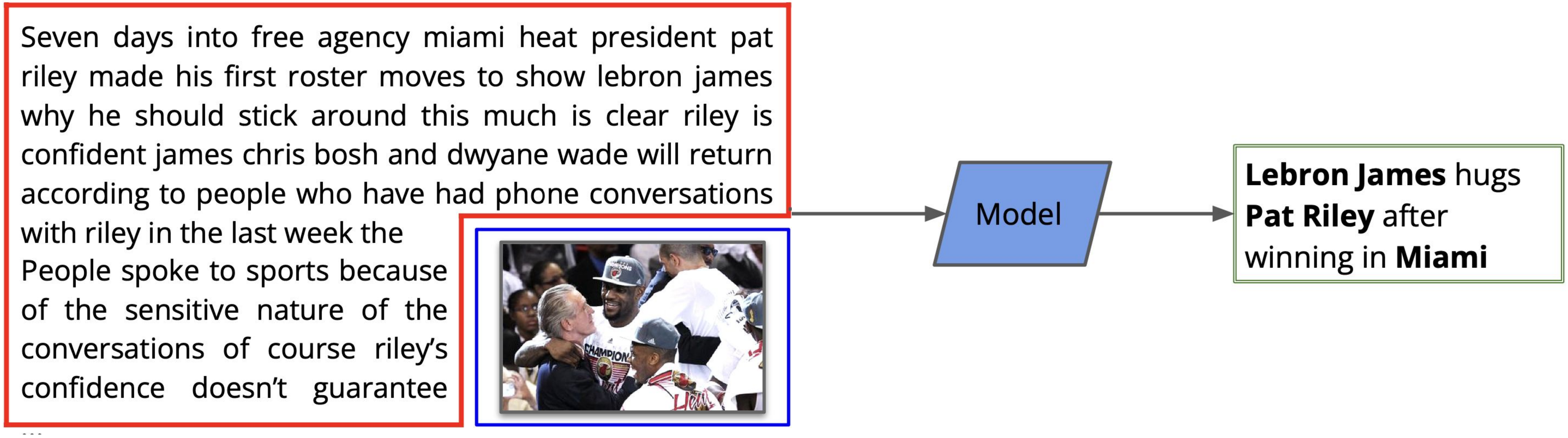
7. Question

1. Introduction

- **News Image Caption**

  - 이미지와 기사로부터 기사의 caption을 예측하는 Task

  - 이미지에 등장하는 entity와 구체적인 상황을 포함하는 caption을 생성해야하기 때문에 단순 Image Caption 보다 어려운 Task



Seven days into free agency miami heat president pat riley made his first roster moves to show lebron james why he should stick around this much is clear riley is confident james chris bosh and dwyane wade will return according to people who have had phone conversations with riley in the last week the People spoke to sports because of the sensitive nature of the conversations of course riley's confidence doesn't guarantee …

Model

Lebron James hugs **Pat Riley** after winning in **Miami**

1. Introduction

- COCO dataset
  - 단순한 상황을 묘사하는 caption

A bunch of people who are holding red umbrellas.

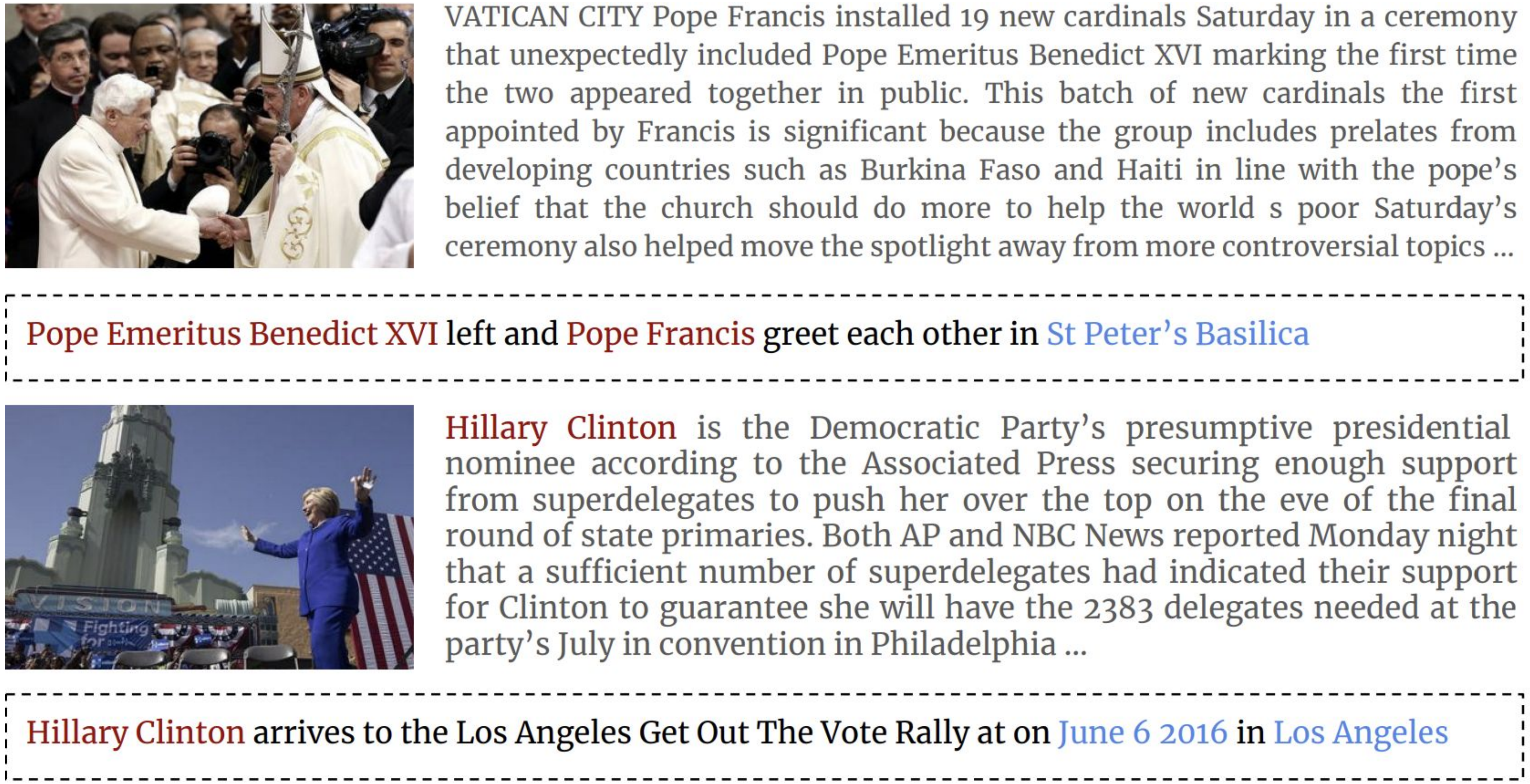
A baseball player hitting the ball during the game.

2. Visual News Dataset

- Visual News

- 기사, 이미지, 캡션, 메타데이터로 구성

- Gaurdian, BBC, USA Today, Washington Post에서 추출됨



VATICAN CITY Pope Francis installed 19 new cardinals Saturday in a ceremony that unexpectedly included Pope Emeritus Benedict XVI marking the first time the two appeared together in public. This batch of new cardinals the first appointed by Francis is significant because the group includes prelates from developing countries such as Burkina Faso and Haiti in line with the pope's belief that the church should do more to help the world s poor Saturday's ceremony also helped move the spotlight away from more controversial topics ...

Pope Emeritus Benedict XVI left and Pope Francis greet each other in St Peter's Basilica

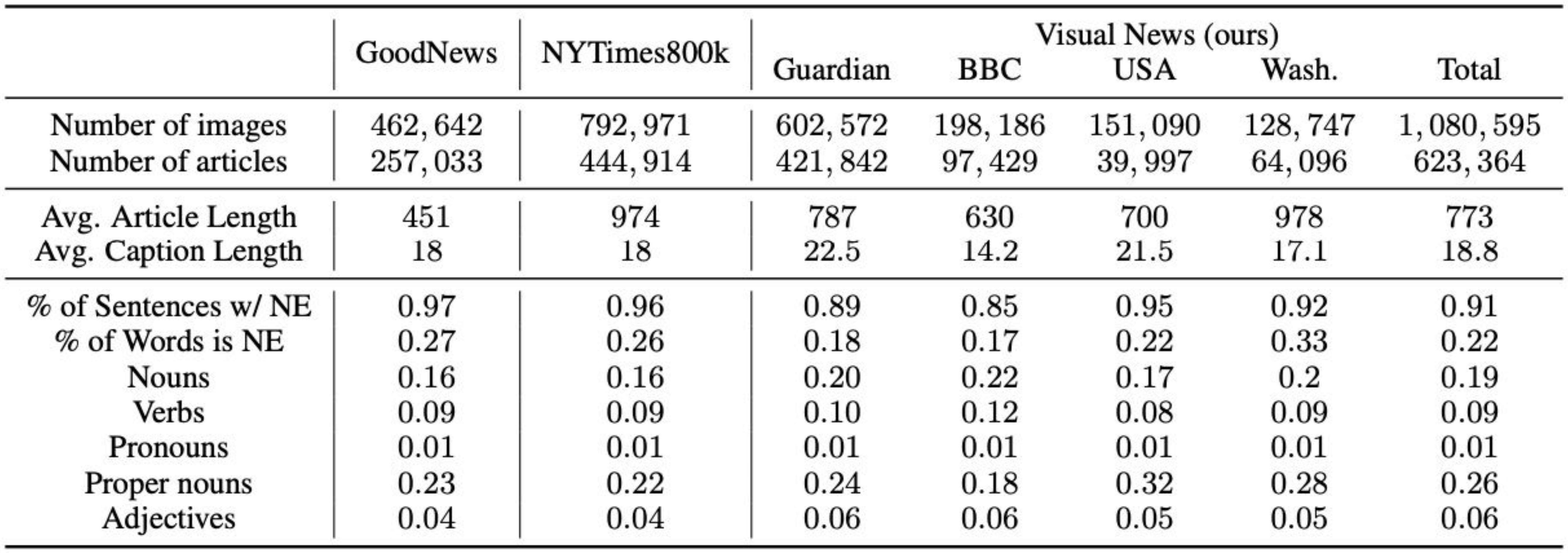
Hillary Clinton is the Democratic Party's presumptive presidential nominee according to the Associated Press securing enough support from superdelegates to push her over the top on the eve of the final round of state primaries. Both AP and NBC News reported Monday night that a sufficient number of superdelegates had indicated their support for Clinton to guarantee she will have the 2383 delegates needed at the party's July in convention in Philadelphia ...

Hillary Clinton arrives to the Los Angeles Get Out The Vote Rally at on June 6 2016 in Los Angeles

2. Visual News Dataset

•Visual News VS Other News Image dataset

- Visual News는 총 100만개 이상의 뉴스 이미지와, 60만개 이상의 뉴스 기사로 구성



	GoodNews	NYTimes800k	Visual News (ours)				
			Guardian	BBC	USA	Wash.	Total
Number of images	462, 642	792, 971	602, 572	198, 186	151, 090	128, 747	1, 080, 595
Number of articles	257, 033	444, 914	421, 842	97, 429	39, 997	64, 096	623, 364
Avg. Article Length	451	974	787	630	700	978	773
Avg. Caption Length	18	18	22.5	14.2	21.5	17.1	18.8
% of Sentences w/ NE	0.97	0.96	0.89	0.85	0.95	0.92	0.91
% of Words is NE	0.27	0.26	0.18	0.17	0.22	0.33	0.22
Nouns	0.16	0.16	0.20	0.22	0.17	0.2	0.19
Verbs	0.09	0.09	0.10	0.12	0.08	0.09	0.09
Pronouns	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Proper nouns	0.23	0.22	0.24	0.18	0.32	0.28	0.26
Adjectives	0.04	0.04	0.06	0.06	0.05	0.05	0.06

2. Visual News Dataset

- Visual News VS Other News Image dataset

- 기사의 정치적인 성향, 등장하는 인물, 주제가 다양한 데이터 셋

PERSON:
Donald Trump, Hillary Clinton, Barack Obama, Lebron James, Pope Francis, Michelle Obama, Novak Djokovic...

GPE:
New York, Florida, Washington, US, Los Angeles, Chicago...

ORG:
NBA, Apple, The White House, Capitol Hill, Boeing, United States Senate...

USA TODAY

PERSON:
Donald Trump, Bill Clinton, Michelle Obama, Barack Obama, Hillary Clinton, Jeb Bush, Pope Francis...

GPE:
Washington, US, New York, Maryland, Virginia...

ORG:
The White House, United States Senate, GOP, Apple, Capitol Hill...

The Washington Post

PERSON:
David Cameron, Nicola Sturgeon, Ed Miliband, Angela Merkel, Carwyn Jones, Nigel Farage, Nick Clegg, Prince Charles...

GPE:
Scotland, China, UK, London, US, India, Syria...

ORG:
BBC, EU, UN, the White House...

BBC

PERSON:
David Cameron, Nicola Sturgeon, Ed Miliband, Angela Merkel, Nigel Farage, Nick Clegg, vladimir putin, Luis Suarez, George Osborne...

GPE:
London, Australia, US, England, Paris, UK, China, India...

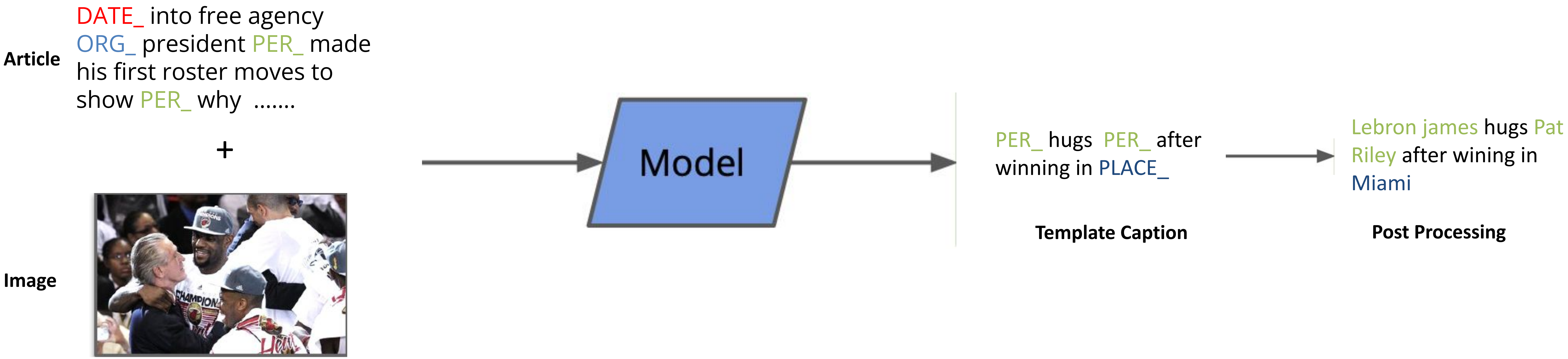
ORG:
UN, EU, BBC, Apple, Chelsea, Tesco...

The Guardian

3. Method

- Previous Work
  - Template 기반 methods

1. 입력으로 사용할 기사에 존재하는 모든 named entity들을 entity type tag들로 교체한 뒤 모델에 입력해 학습
2. 이후 모델이 template caption을 생성하면 해당 caption에 존재하는 entity tag를 기사에 등장하는 entity로 교체



Article DATE_ into free agency ORG_ president PER_ made his first roster moves to show PER_ why ....

+

Image

PER_ hugs PER_ after winning in PLACE_

Template Caption

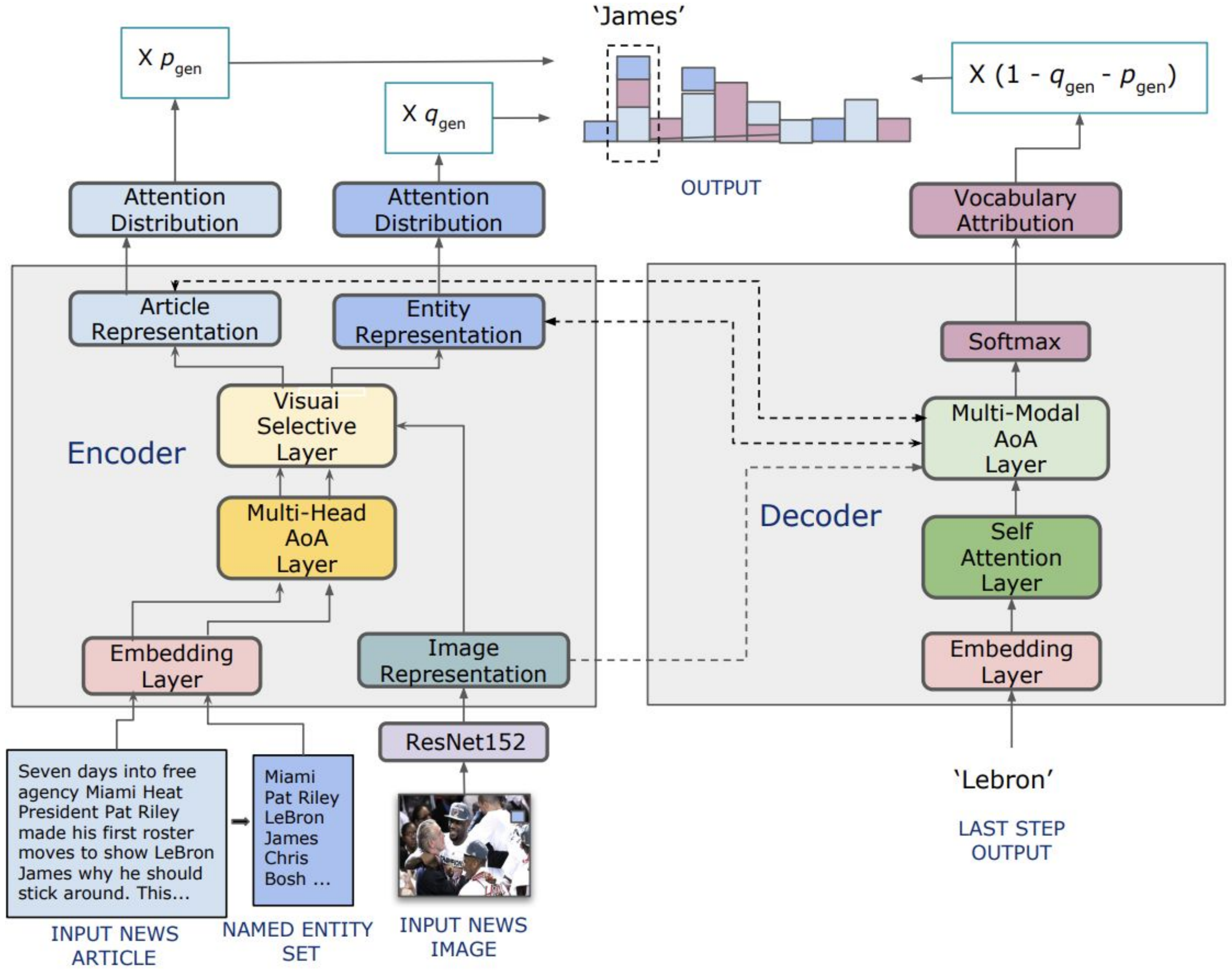
Lebron james hugs Pat Riley after wining in Miami

Post Processing

3. Method

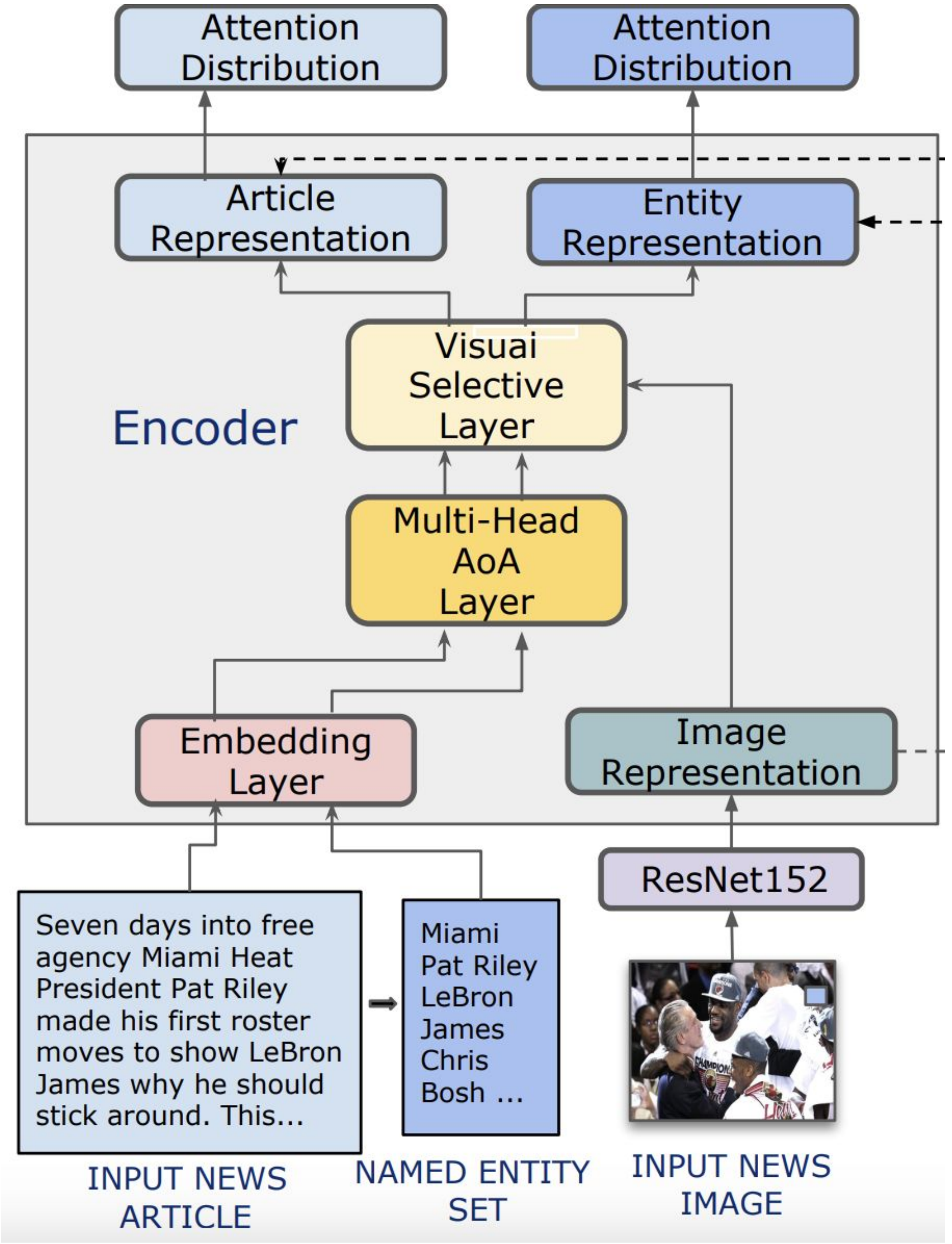
- Visual News Captioner

  - 논문에서 제안한 새로운 News Image caption model의 architecture



3. Method

- Visual News Captioner
  - Encoder



1. 기사에서 entity를 추출해 Named Entity set생성, 기사와 entity set을 Encoder에 입력

2. **Embedding Layer**

- 기사와 named entity set는 각각의 임베딩 표현으로 변환

3. **Multi-Head AoA Layer**

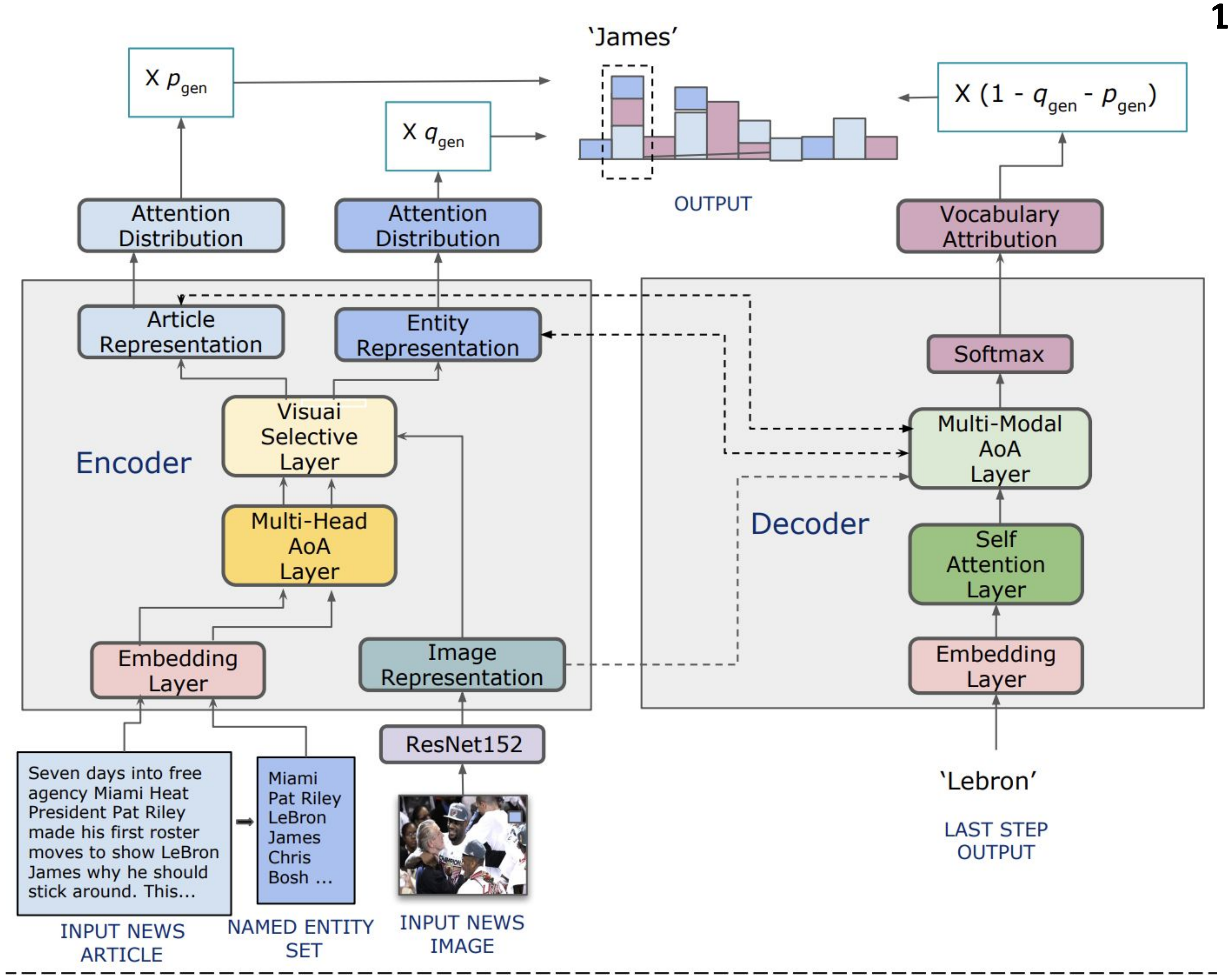
- 기사와 entity의 임베딩 벡터에 self-attention을 진행

4. Visual Selective Layers

- cross-attention을 이용해 택스트의 정보와 ResNet152에서 추출된 이미지의 정보를 함께 인코딩해 기사와 entity의 인코딩 벡터를 생성

3. Method

- Visual News Captioner
  - Decoder



1. **Self-Attention Layer**:

- 이전 time step에서 예측한 시퀀스 Self-attention

2. **Multi-Modal Attention:**

- 입력 시퀀스를 Query로, Encoder에서 전달 받은 정보를 Key로 cross-attention을 진행

$$V'_t = \mathrm{MHAtt}_{\mathrm{AoA}}(x^a_t, \tilde{V}, \tilde{V}),$$
$$A'_t = \mathrm{MHAtt}_{\mathrm{AoA}}(x^a_t, A, A),$$
$$E'_t = \mathrm{MHAtt}_{\mathrm{AoA}}(x^a_t, E, E).$$

x^a_t : 이전 time-step의 출력 시퀀스

$\tilde{V}$: Image의 feature vector

A : article의 인코딩된 정보

E : Entity의 인코딩된 정보

- cross-attention으로 생성된 결과를 이용해 다음 step의 token 확률(P_{S_t}) 예측

$$C_t = V'_t + A'_t + E'_t,$$
$$x'_t = \mathrm{LayerNorm}(x^a_t + C_t),$$
$$x^*_t = \mathrm{LayerNorm}(x'_t + \mathrm{FFN}(x'_t)),$$
$$P_{s_t} = \mathrm{softmax}(x^*_t).$$

4. Experimental Results

- BLEU(**bilingual Evaluation Understudy)**
  - Ground truth와 prediction 결과가 얼마나 유사한지 비교하는 방법
  - N-gram에 기반하여 측정

$$p_n = \frac{\sum_{n\text{-}gram \in Candidate} Count_{clip}(n\text{-}gram)}{\sum_{n\text{-}gram \in Candidate} Count(n\text{-}gram)}$$

$$BLEU = exp(\sum_{n=1}^{N} w_n \log p_n)$$

4. Experimental Results

- BLEU(**bilingual Evaluation Understudy)**
  - Ground truth와 prediction 결과가 얼마나 유사한지 비교하는 방법
  - N-gram에 기반하여 precision을 측정

$$p_n = \frac{\sum_{n\text{-}gram \in Candidate} Count_{clip}(n\text{-}gram)}{\sum_{n\text{-}gram \in Candidate} Count(n\text{-}gram)}$$

$$BLEU = exp(\sum_{n=1}^{N} w_n \log p_n)$$

4. Experimental Results

- BLEU(**bilingual Evaluation Understudy)**

Ground Truth: : Lebron james hugs Pat Riley after winning in Miami

Prediction: LeBron James hugs Pat Riley After Miami Win

- unigram

unigram	Lebron	james	hugs	Pat	Riley	After	Miami	win	SUM
count	1	1	1	1	1	1	1	1	8
precision	1	1	1	1	1	1	1	1	8/8

- bigram

bigram	Lebron james	James hugs	hugs Pat	Pat Riley	Riley After	After Miami	Miami win	SUM
count	1	1	1	1	1	1	1	7
precision	1	1	1	1	1	0	0	5/7

- trigram

trigram	Lebron james hugs	James hugs Pat	hugs Pat Riley	Pat Riley after	Riley after Miami	After Miami win	SUM
count	1	1	1	1	1	1	6
precision	1	1	1	1	0	0	4/6

4. Experimental Results

- BLEU(**bilingual Evaluation Understudy)**

Ground Truth: : Lebron james hugs Pat Riley after winning in Miami

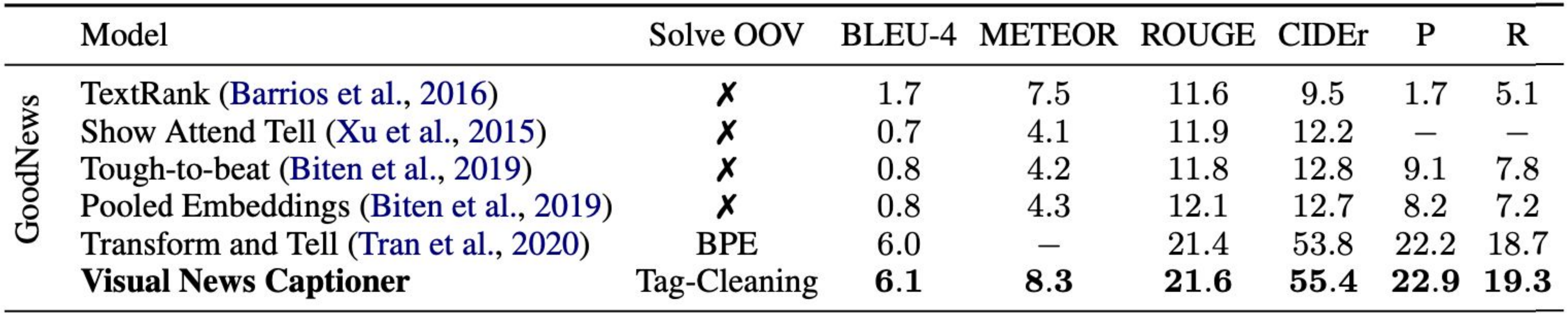
Prediction: LeBron James hugs Pat Riley After Miami Win

N-gram	precision
Uni-gram	8/8
Bi-gram	5/7
Tri-gram	4/6
4-gram	3/5

$$BLEU = exp(\sum_{n=1}^{N} w_n \log p_n)$$

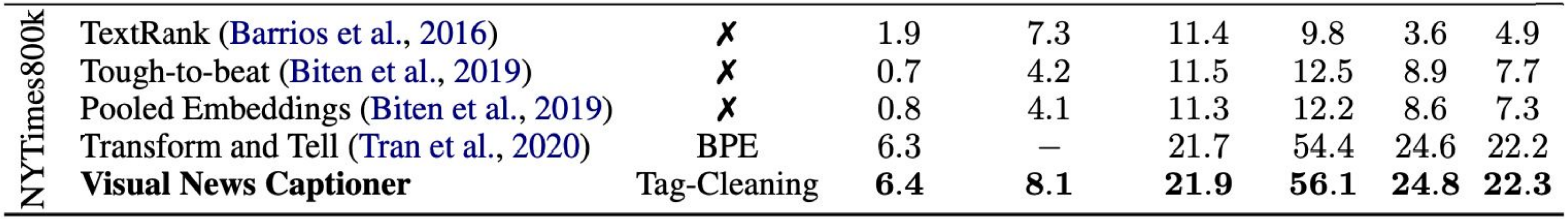
4. Experimental Results

- Evaluation on GoodNews



	Model	Solve OOV	BLEU-4	METEOR	ROUGE	CIDEr	P	R
GoodNews	TextRank (Barrios et al., 2016)	✗	1.7	7.5	11.6	9.5	1.7	5.1
	Show Attend Tell (Xu et al., 2015)	✗	0.7	4.1	11.9	12.2	–	–
	Tough-to-beat (Biten et al., 2019)	✗	0.8	4.2	11.8	12.8	9.1	7.8
	Pooled Embeddings (Biten et al., 2019)	✗	0.8	4.3	12.1	12.7	8.2	7.2
	Transform and Tell (Tran et al., 2020)	BPE	6.0	–	21.4	53.8	22.2	18.7
	Visual News Captioner	Tag-Cleaning	**6.1**	**8.3**	**21.6**	**55.4**	**22.9**	**19.3**

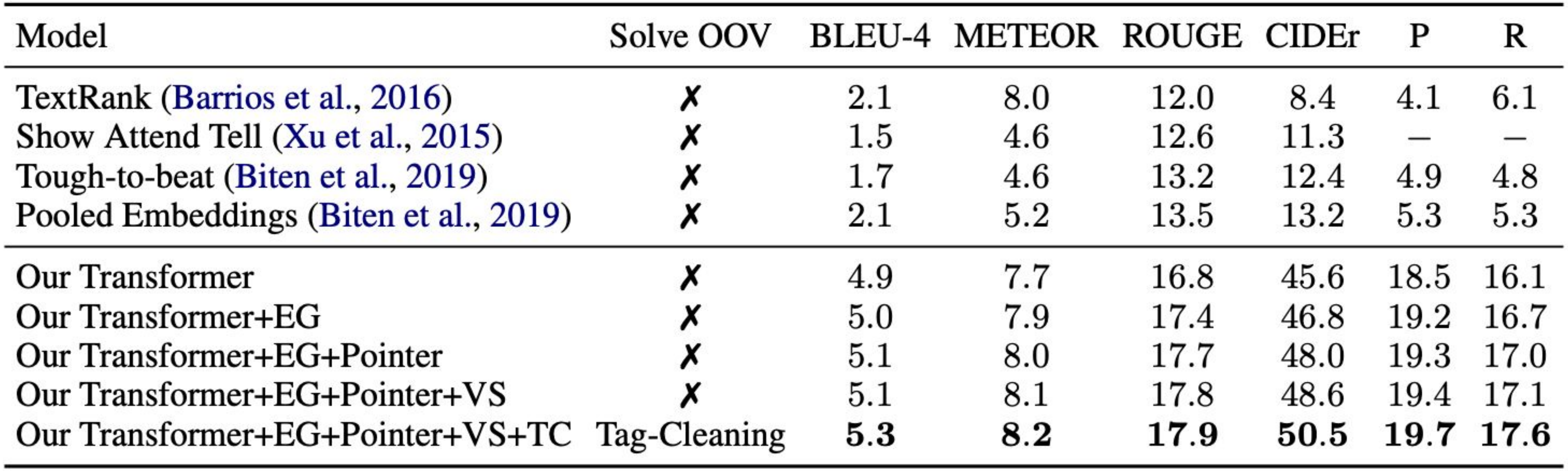
- Evaluation on NYTimes800k



NYTimes800k	TextRank (Barrios et al., 2016)	✗	1.9	7.3	11.4	9.8	3.6	4.9
	Tough-to-beat (Biten et al., 2019)	✗	0.7	4.2	11.5	12.5	8.9	7.7
	Pooled Embeddings (Biten et al., 2019)	✗	0.8	4.1	11.3	12.2	8.6	7.3
	Transform and Tell (Tran et al., 2020)	BPE	6.3	–	21.7	54.4	24.6	22.2
	Visual News Captioner	Tag-Cleaning	**6.4**	**8.1**	**21.9**	**56.1**	**24.8**	**22.3**

4. Experimental Results

- Evaluation on Visual News



Model	Solve OOV	BLEU-4	METEOR	ROUGE	CIDEr	P	R
TextRank (Barrios et al., 2016)	✗	2.1	8.0	12.0	8.4	4.1	6.1
Show Attend Tell (Xu et al., 2015)	✗	1.5	4.6	12.6	11.3	–	–
Tough-to-beat (Biten et al., 2019)	✗	1.7	4.6	13.2	12.4	4.9	4.8
Pooled Embeddings (Biten et al., 2019)	✗	2.1	5.2	13.5	13.2	5.3	5.3
Our Transformer	✗	4.9	7.7	16.8	45.6	18.5	16.1
Our Transformer+EG	✗	5.0	7.9	17.4	46.8	19.2	16.7
Our Transformer+EG+Pointer	✗	5.1	8.0	17.7	48.0	19.3	17.0
Our Transformer+EG+Pointer+VS	✗	5.1	8.1	17.8	48.6	19.4	17.1
Our Transformer+EG+Pointer+VS+TC	Tag-Cleaning	**5.3**	**8.2**	**17.9**	**50.5**	**19.7**	**17.6**

5. Conclusion

- 100만개 이상의 article과 이미지가 동반된 대규모 Visual News dataset 제안
- 시각 정보와 텍스트 정보를 함께 활용하는 Entity-aware 방식의 Visual News Captioner 제안

6. Questions

- 본 논문에서는 이미지 + 텍스트(entity or 기사)의 정보만 함께 인코딩했지만, 텍스트(entity) + 텍스트(기사)을 함께 인코딩을 해보면 어떨까?

- visual news dataset의 다른 활용 방안