위키피디아 링크 데이터를 이용한 Neural Network Model 기반 한국어 개체명 연결

Young-Hoon Lee, Seung-Hoon Na

Cognitive Computing Lab.

Chonbuk National University

목차

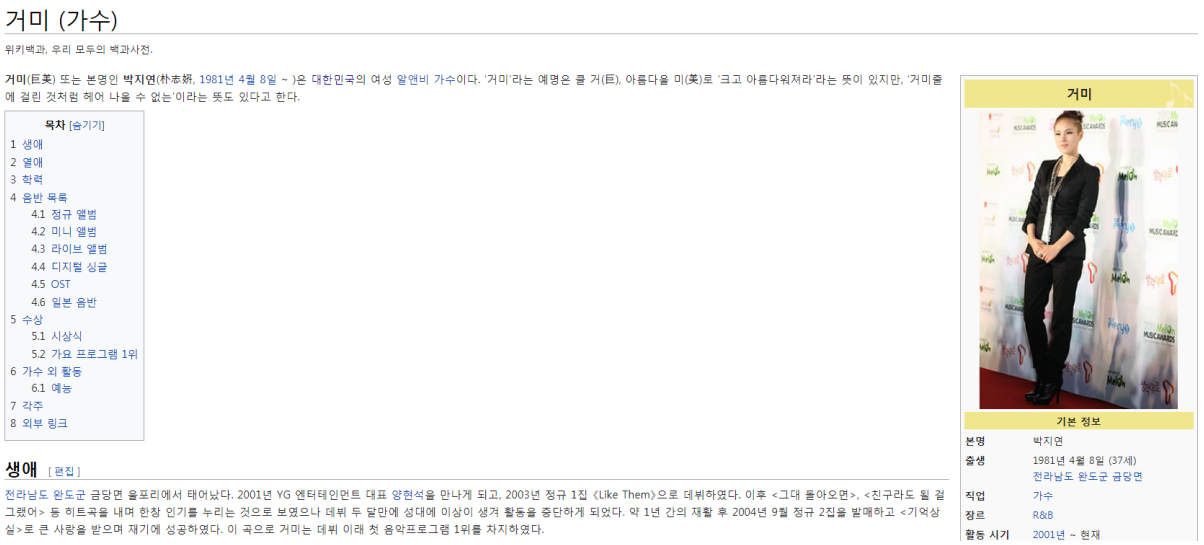
- 서론
- 관련연구
- 모델 구조
- 학습/평가 데이터
- 실험결과

개체명 연결 (Entity Linking)

주어진 문장에 출현한 단어를 지식 기반(Knowledge base)상의 개체와 연결하여 특정 개체가 무엇인지 식별하는 작업

개체명 연결 (Entity Linking)

- **Entity : 거미 (가수) (KB 상의 하나의 Entity)** https://ko.wikipedia.org/wiki/거미_(가수)
- **Mention : 거미 (Entity의 표현)**
- **Context : SM 아카데미 대표인 이솔림 씨의 추천으로 참가해 거미&휘성의 곡인《Do It》을 부르고 이수만 대표에게 직접 발탁된 후, 2년의 연습생기간을 걸쳐 2008년 5월 25일 샤이니의 멤버로 정식 데뷔하였다. (Mention이 포함된 문장)**



거미 (가수)

위키백과, 우리 모두의 백과사전.

거미(巨美) 또는 본명인 **박지연**(朴志娟, 1981년 4월 8일 ~)은 대한민국의 여성 알앤비 가수이다. '거미'라는 예명은 클 거(巨), 아름다울 미(美)로 '크고 아름다워져라'라는 뜻이 있지만, '거미줄에 걸린 것처럼 헤어 나올 수 없는'이라는 뜻도 있다고 한다.

목차 [숨기기]

1 생애
2 열애
3 학력
4 음반 목록
4.1 정규 앨범
4.2 미니 앨범
4.3 라이브 앨범
4.4 디지털 싱글
4.5 OST
4.6 일본 음반
5 수상
5.1 시상식
5.2 가요 프로그램 1위
6 가수 외 활동
6.1 예능
7 각주
8 외부 링크

생애 [편집]

전라남도 완도군 금당면 울포리에서 태어났다. 2001년 YG 엔터테인먼트 대표 양현석을 만나게 되고, 2003년 정규 1집 《Like Them》으로 데뷔하였다. 이후 <그대 들어오면>, <진구라도 될 걸 그랬어> 등 히트곡을 내며 한창 인기를 누리는 것으로 보였으나 데뷔 두 달만에 성대에 이상이 생겨 활동을 중단하게 되었다. 약 1년 간의 재활 후 2004년 9월 정규 2집을 발매하고 <기억상실>로 큰 사랑을 받으며 재기에 성공하였다. 이 곡으로 거미는 데뷔 이래 첫 음악프로그램 1위를 차지하였다.

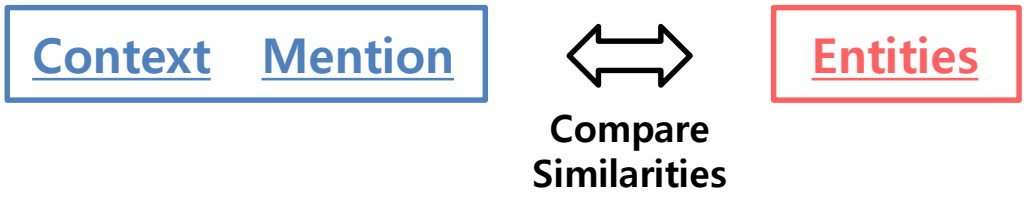
거미

기본 정보

본명	박지연
출생	1981년 4월 8일 (37세) 전라남도 완도군 금당면
직업	가수
장르	R&B
활동 시기	2001년 ~ 현재

개체명 연결 (Entity Linking)

- Input pair(Context, Mention)과 Candidate Entities의 유사도를 비교하여



Entity Disambiguation → Ranking Task

가수 이승철은 존박의 1집 앨범 '이너차일드(Inner Child)'에 대해 ...

Candidate Entities

이승철_(가수)	0.65
이승철_(배우)	0.05
이승철_(기업인)	0.01
...	

Joint Learning of the Embedding of Words and Entities for Named Entity Disambiguation, I. Yamada, 2016

Skip-gram Model + (KB graph model, anchor context model)

- **Average the embedding vectors**

$$\overrightarrow{v_{c_w}} = \frac{1}{|W_{c_m}|} \sum_{w \in W_{c_m}} \overrightarrow{v_w} \qquad \overrightarrow{v_{c_e}} = \frac{1}{|E_{c_m}|} \sum_{e^* \in E_{c_m}} \overrightarrow{v_{e^*}}$$

Textual context **Context entities**

Embedding vector의 평균을 이용해 유사도를 비교하여

Similarity (Candidate entity, Textual context) $sim(\overrightarrow{v_e}, \overrightarrow{v_{c_w}})$

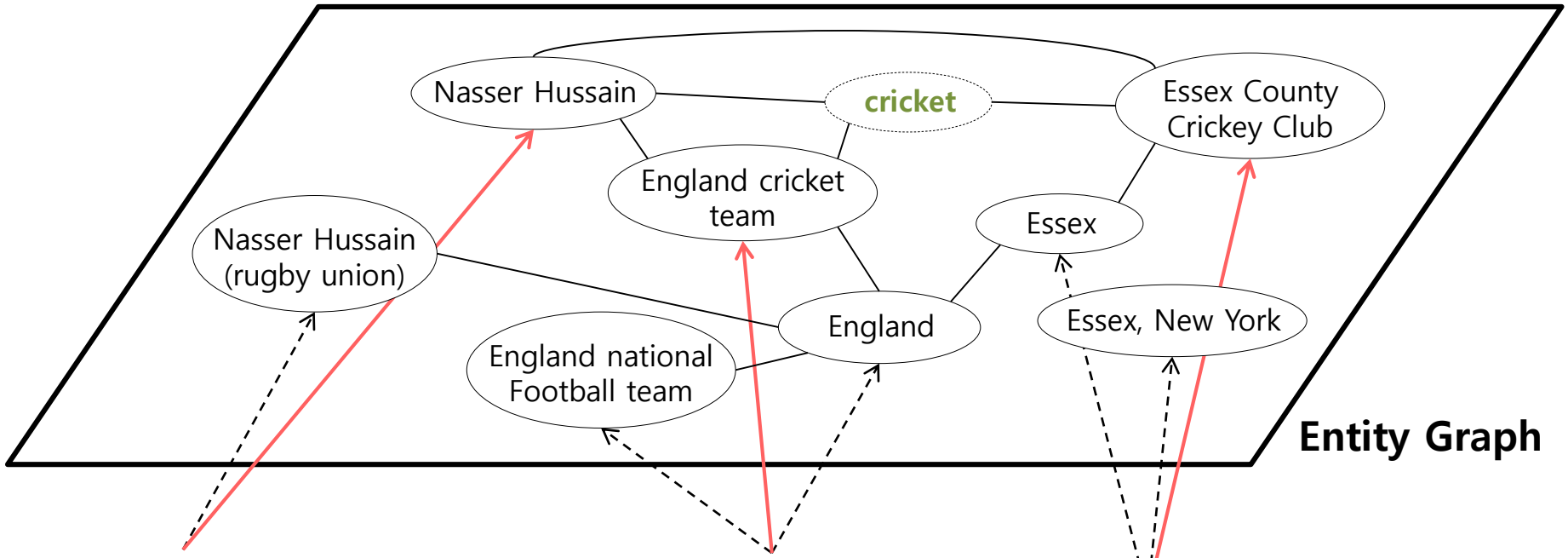
Similarity (Entity, Contextual entities) $sim(\overrightarrow{v_e}, \overrightarrow{v_{c_e}})$

} **Sort**

⇒ Ranking로 변환하여 Entity Disambiguation

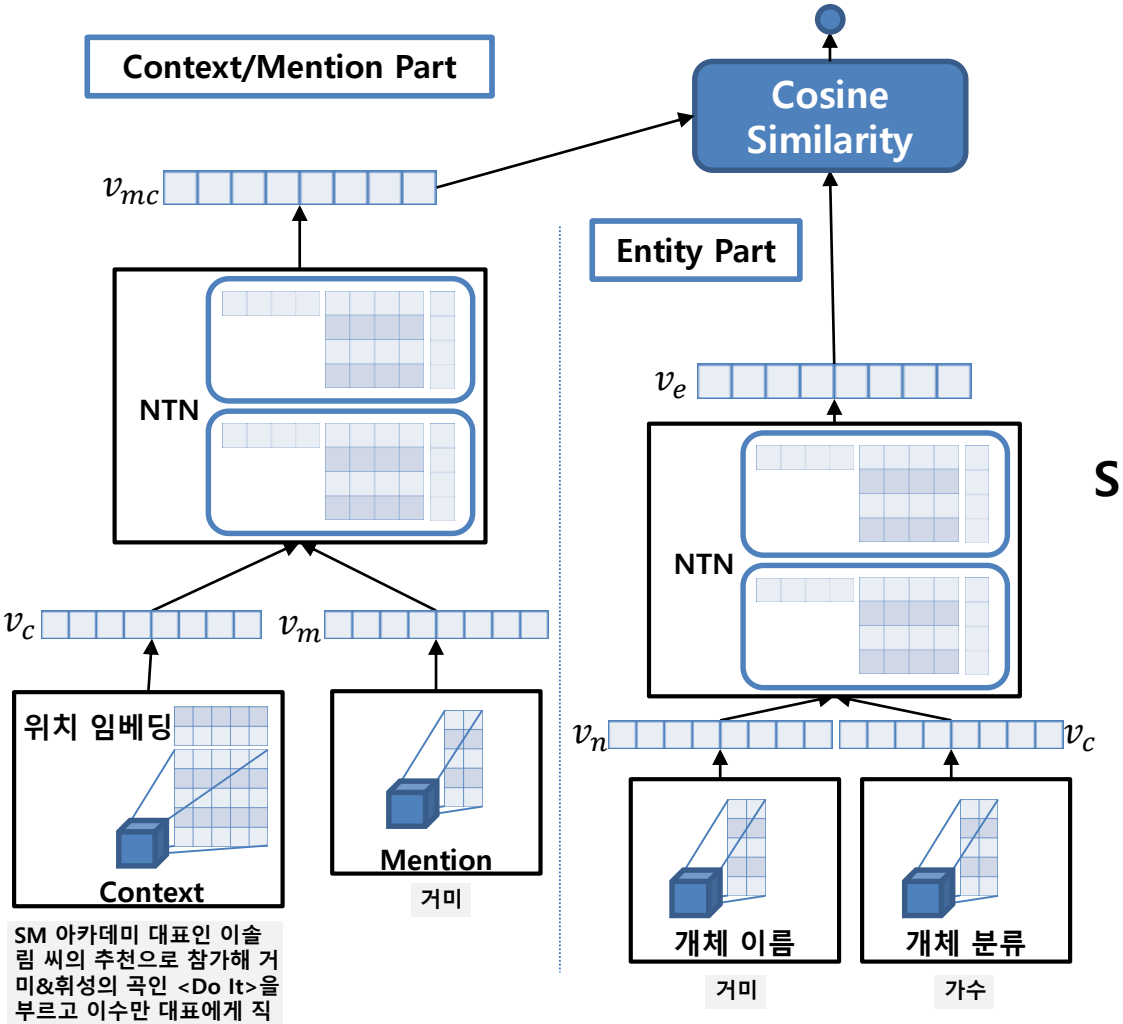
Neural Collective Entity Linking, Y. Cao, 2018

Local Feature + Global Feature



Hussain, considered surplus to **England**'s one-day requirements, Struck 158, his first championship century of the season, as **Essex** reached 372 And took a first innings lead of 82

제안 모델



Modeling Mention, Context and Entity with Neural Networks for Entity Disambiguation,
Y. Sun, IJCAI, 2015

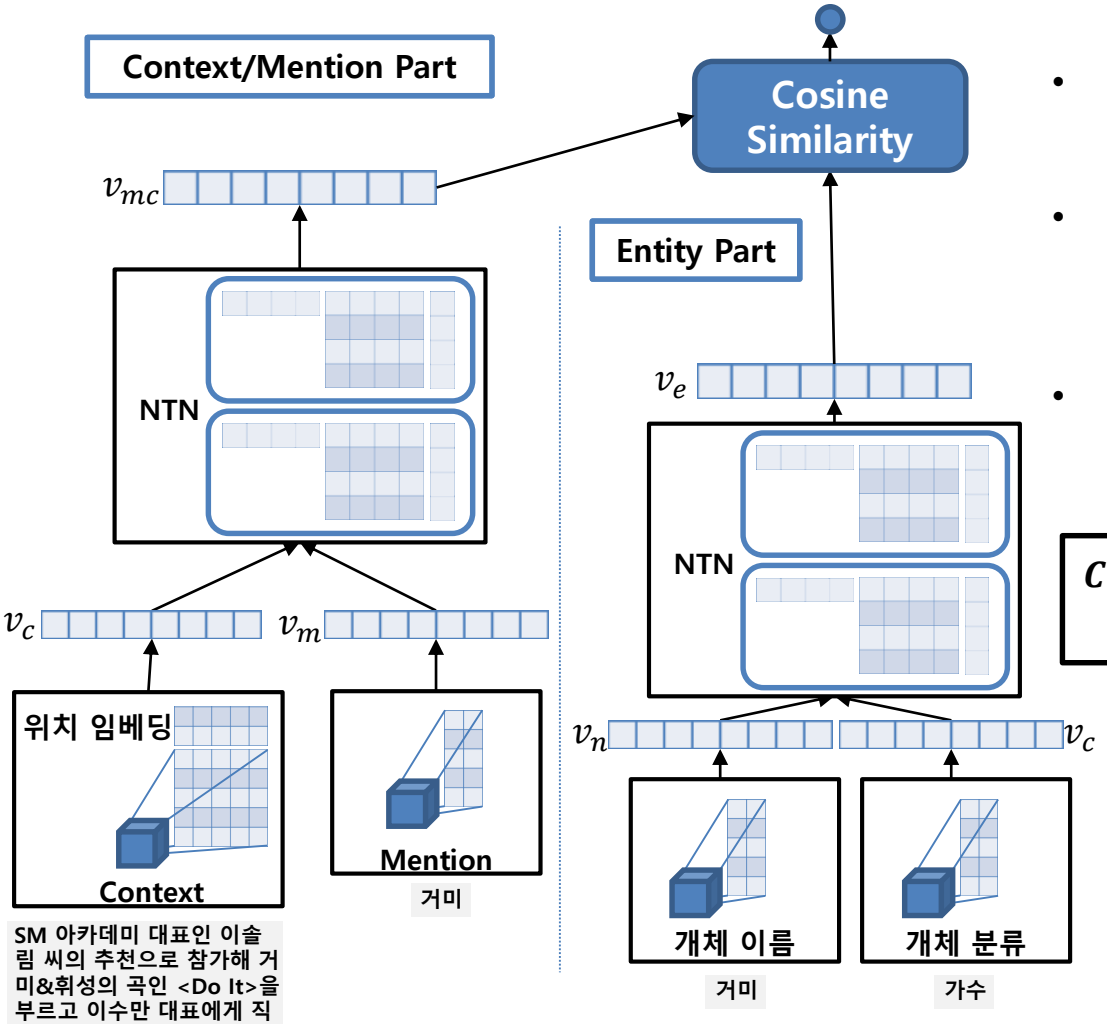
Similarity(entity, mention/context)

v_e = NTN(entity name, entity class)

v_{mc} = NTN(mention, context)

$$sim(e, mc) = cosine(v_e, v_{mc})$$

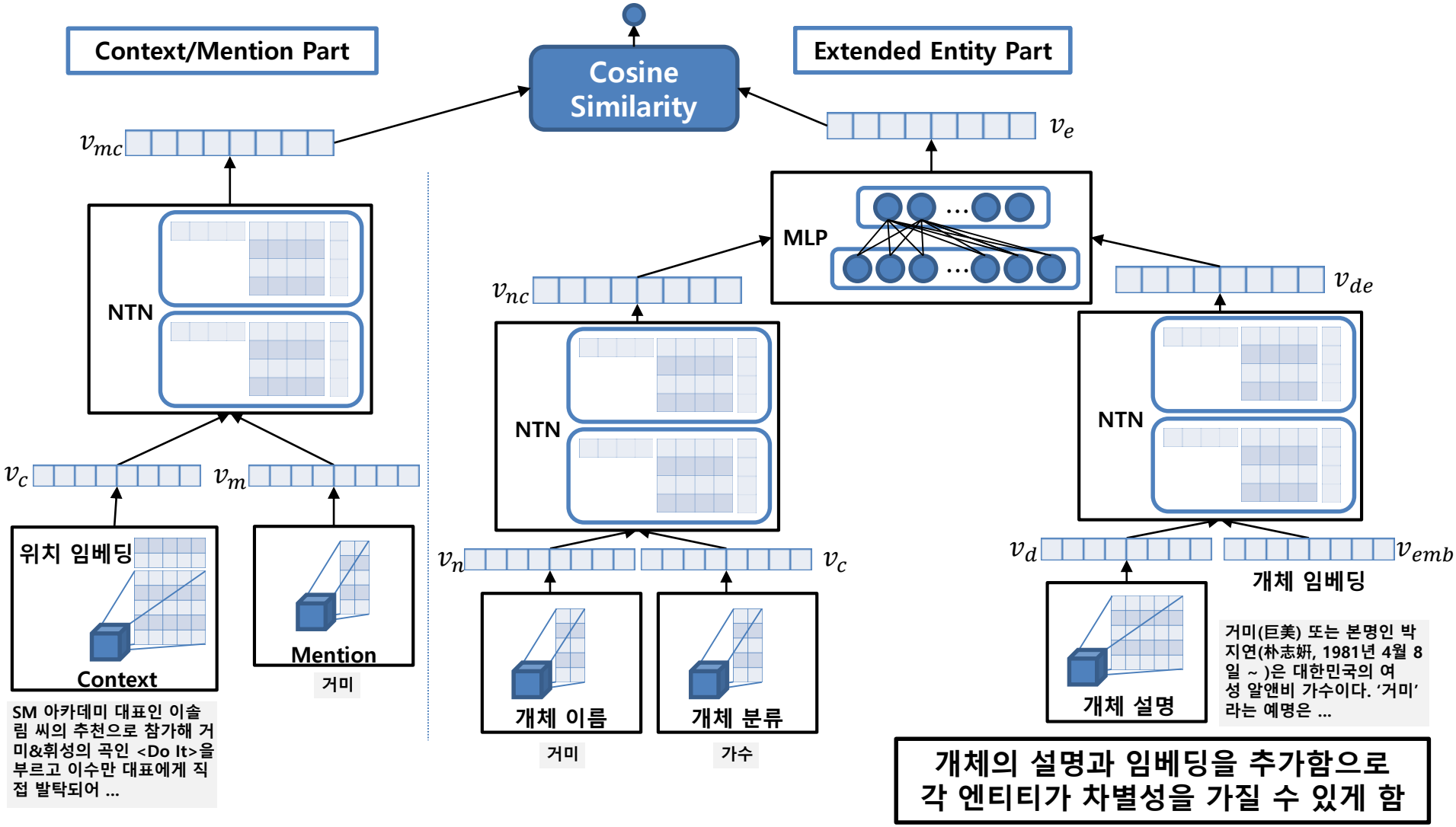
제안 모델



- 하지만, Entity의 이름과 클래스 표현만으로 Entity를 분류하기에는 차별성이 부족함
- 따라서 Entity의 Description과 Embedding을 통하여 Extend Entity Representation을 얻고 Entity들간의 차별성을 부여
- 또한 위 모델을 한국어에 적용하기 위해 Convolution 연산과 Pooling 연산을 통해 문장 크기의 Vector를 구함

$$CharEmbed = MaxPooling(Conv1d(Embed(X)))$$

개체명 연결 Neural Network 모델 구조



개체명 연결 Neural Network 모델 구조

1. Mention/Entity Name, Class Representation

Mention이나 Entity의 이름, 분류 정보는 대체로 짧음

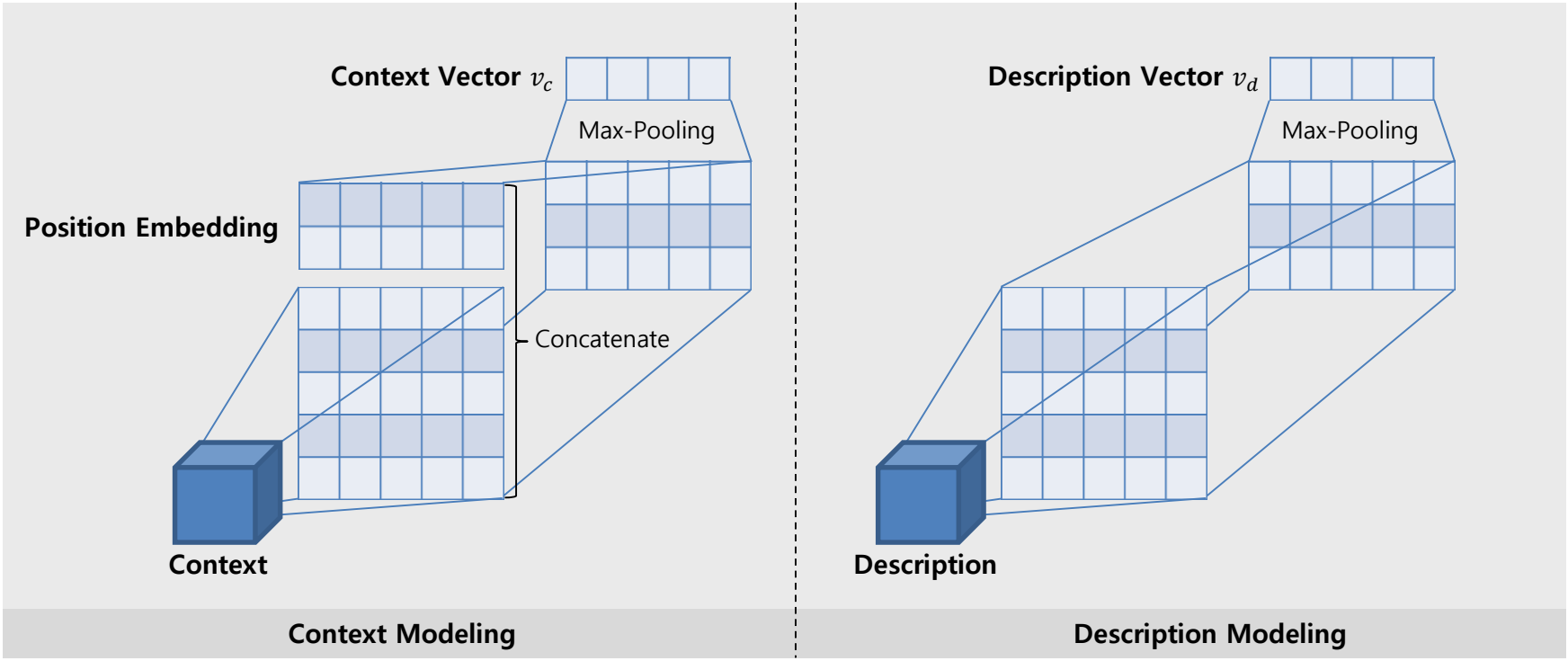
⇩

Embedding Vector의 평균

$$v_T = \frac{1}{|T|} \sum_{t \in T} v_t$$

개체명 연결 Neural Network 모델 구조

2. Context/Entity Description Representation



개체명 연결 Neural Network 모델 구조

3. Neural Tensor Network (NTN)

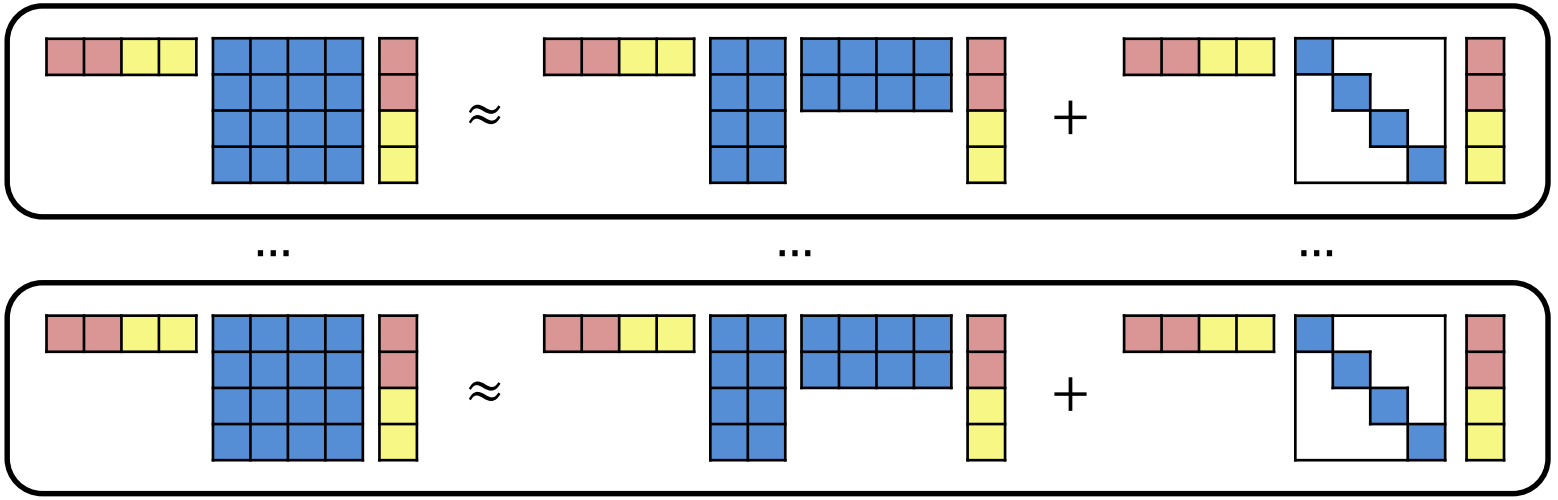
두 개의 Vector 표현을 하나로 통합하고
Parameter의 수를 줄이기 위해 근사화된 Low-Rank NTN 사용

$$v_{mc} = [v_m; v_c]^T \left[M_i^{appr}\right]^{[1:L]} [v_m; v_c]$$

$$M_i^{appr} = M_{i1} \times M_{i2} + diag(m_i)$$

$$M_{i1} \in \mathbb{R}^{N \times r},\ M_{i2} \in \mathbb{R}^{r \times N}, m_i \in \mathbb{R}^N$$

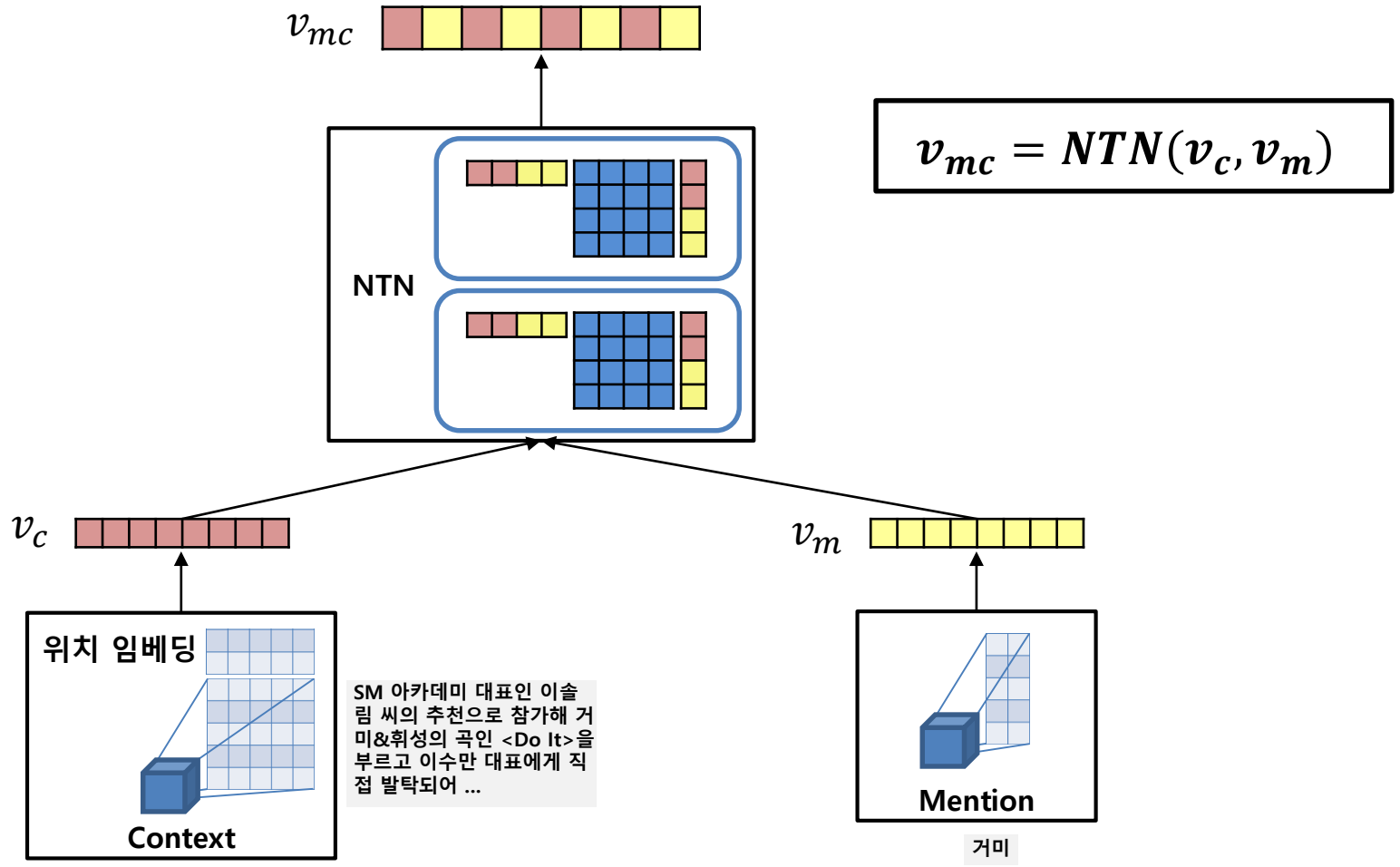
Tensor의 각 Matrix을 반으로 나누어 곱하고, Diagonal Matrix을 더하여 근사화된 NTN을 표현



slice = $1 \sim L$

개체명 연결 Neural Network 모델 구조

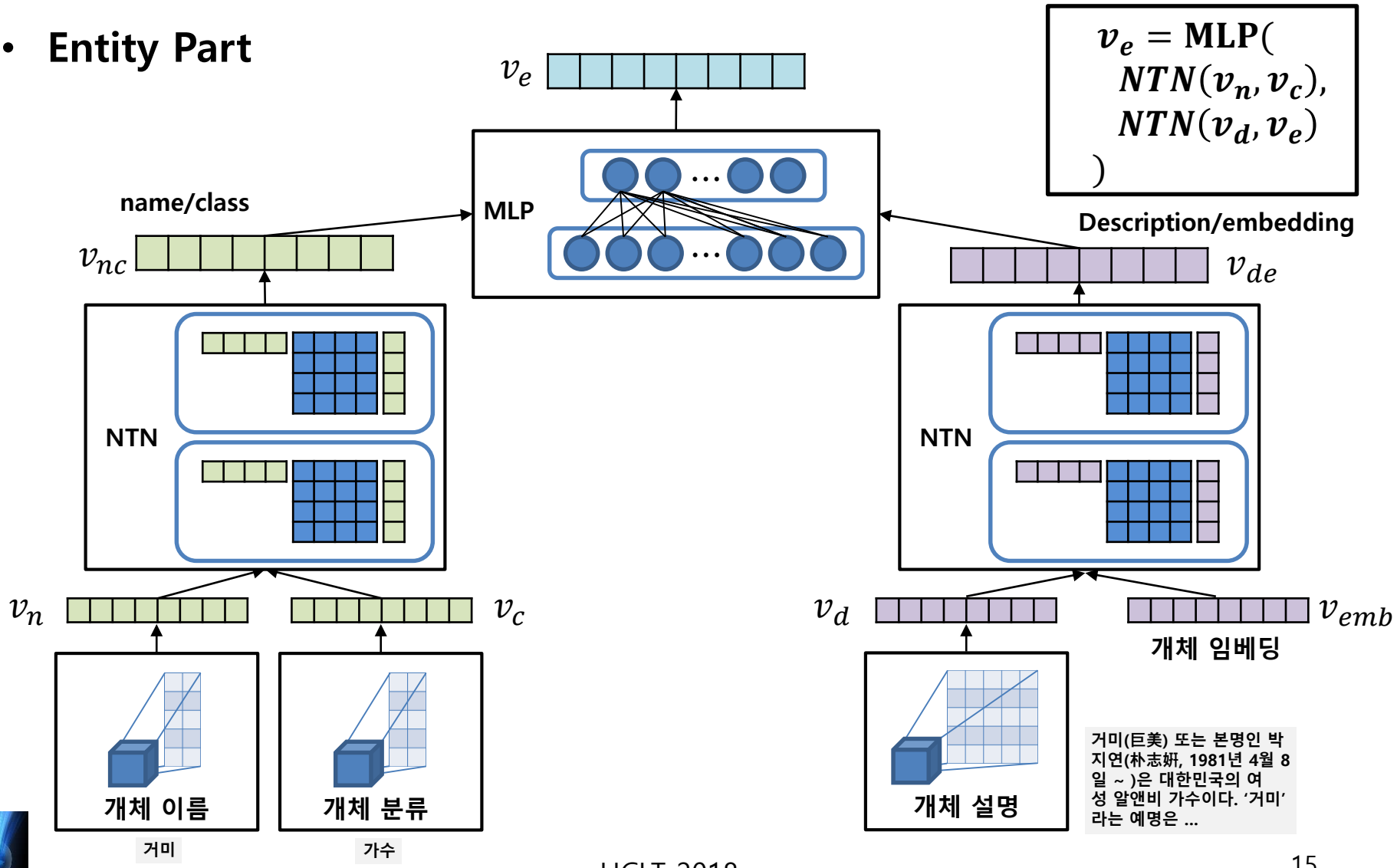
- **Context/Mention Part**



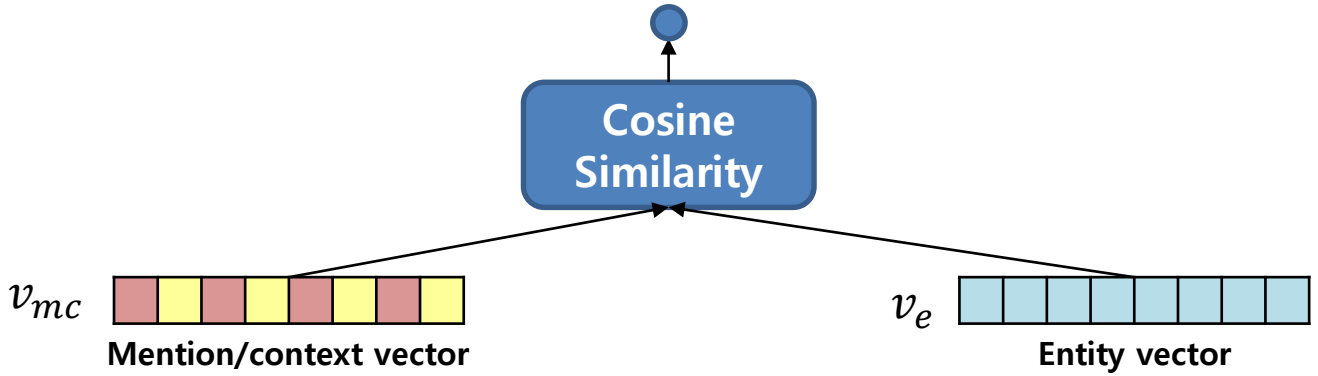
$$v_{mc} = NTN(v_c, v_m)$$

개체명 연결 Neural Network 모델 구조

- **Entity Part**



개체명 연결 Neural Network 모델 구조



- **Cosine Similarity : 개체와 Mention/Context 간의 유사도 비교**

$$sim(e, mc) = cosine(v_e, v_{mc})$$

- **Loss function : 정답 개체의 Score가 임의로 선택된 개체보다 1보다는 크다**

$$loss = \sum_{(m,c)\in T} max(0, 1 - sim(e, mc) + sim(e', mc))$$

e : 정답개체, e': 임의로 선택된 개체

학습 / 평가 데이터 구성

위키피디아 동음이의어 사전 ⇨ Disambiguation 개체 정보

Disambiguation 개체 정보가 위키피디아의 설명에 존재하면, 학습데이터로 사용하였음

… 이솔림 씨의 추천으로 참가해 <a href="…">거미<a>&휘성의 곡인《Do It》을 부르고 이수만 대표에게 직접 발탁된 후, …

https://ko.wikipedia.org/wiki/거미_(가수)

거미 (가수)

위키백과, 우리 모두의 백과사전.

거미(巨美) 또는 본명인 **박지연**(朴志妍, 1981년 4월 8일 ~)은 대한민국의 여성 알앤비 가수이다. '거미'라는 예명은 클 거(巨), 아름다울 미(美)로 '크고 아름다워져라'라는 뜻이 있지만, '거미줄에 걸린 것처럼 헤어 나올 수 없는'이라는 뜻도 있다고 한다.

목차 [숨기기]

1 생애
2 열애
3 학력
4 음반 목록
4.1 정규 앨범
4.2 미니 앨범
4.3 라이브 앨범

Entity Description :
거미(巨美) 또는 본명인 박지연(朴志妍, 1981년 4월 8일 ~)은 대한민국의 여성 알앤비 가수이다. '거미'라는 예명은 …

분류: 1981년 태어남 | 살아있는 사람 | 2003년 데뷔 | 대한민국의 여자 가수 | 골든디스크 본상 수상 음악가 | 한국대중음악상 수상 음악가 | Mnet 아시안 뮤직 어워드 수상 음악가 | Mnet 엠카운트다운 1위 수상자 | SBS 인기가요 뮤티즌송 수상자 | 나는 가수다에 출연한 가수 | 대한민국의 R&B 가수 | 한국어 가수 | 일본어 가수 | 2000년대 가수 | 2010년대 가수 | 에이벡스 그룹 음악가 | 대한민국의 발라드 음악가 | 대한민국의 개신교도 | 세화여자고등학교 동문 | 서울이수초등학교 동문 | 서문여자중학교 동문 | 동덕여자대학교 동문 | 완도군 출신 | YG 엔터테인먼트 소속

Entity Class : {가수 : 7, 음악가 : 5, 동문: 4, 수상자 : 2, …}

확장 모델

- **Model 1 : Entity Name + Entity Class (Y. Sun, 2015)**

$$sim(e, mc) = cosine(v_{nc}, v_{mc})$$

- **Model 2 : Model 1 + Entity Description**

$$sim(e, mc) = cosine(MLP(v_{nc}, v_d), v_{mc})$$

- **Model 3 : Model1 + Entity Description + Entity Embedding**

$$sim(e, mc) = cosine(v_e, v_{mc})$$

실험 결과

- 7:3의 비율로 랜덤하게 나눔, 학습 데이터 셋 : 62657개, 평가 데이터 셋 : 26853개
- 정답 개체 1개와 임의의 개체 19개를 포함한 총 20개의 후보 개체 중에서 정답 개체의 Score가 가장 높은 값을 가지게 되면 정답으로 평가.

모델	정확도
Model 1 (mention, context only)	86.18%
Model 2 (model 1 + entity description)	88.12%
Model 3 (Model 2 + entity embedding)	**89.63%**

- 개체의 설명과 개체 임베딩을 추가한 Model 3은 Base Model보다 약 3.5% 성능을 향상한 89.63%의 성능을 보임
- 특히 Model 2의 성능이 Base Model보다 약 2% 증가율을 보여 Entity Description이 성능 향상에 큰 영향을 가지는 것을 보임