



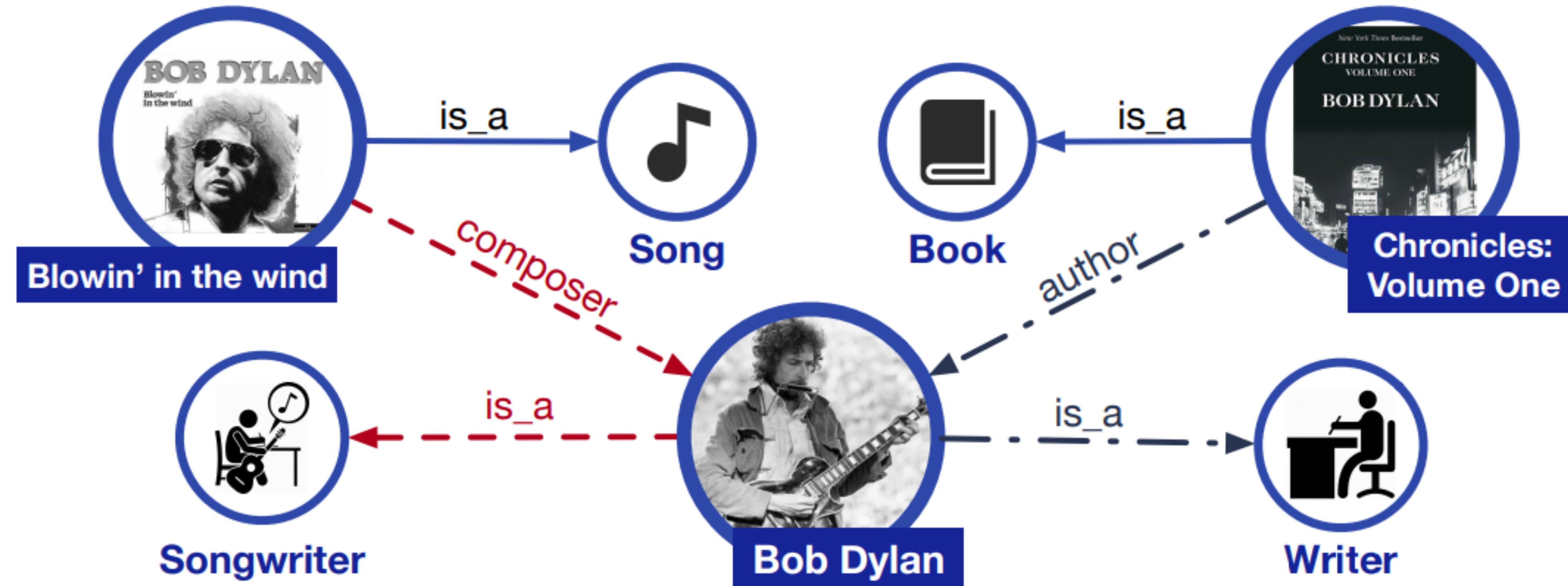
ERNIE: Enhanced Language Representation with Informative Entities

석사과정 송지수

Introduction

- Word2Vec, FastText, BERT, ...
- 널리 알려진 Embedding (Language Model) 알고리즘들은 모두 large-scale corpus를 학습하여 semantic / syntactic 정보 및 패턴들을 이용해 단어들을 Vectorize
- 그런데 여기에 추가적인 정보를 Injecting 해 주면 더 성능이 좋은 Language Model을 만들 수 있지 않을까?!

Introduction



Bob Dylan wrote **Blowin' in the Wind** in 1962, and wrote **Chronicles: Volume One** in 2004.

Introduction

- **Knowledge Graph**

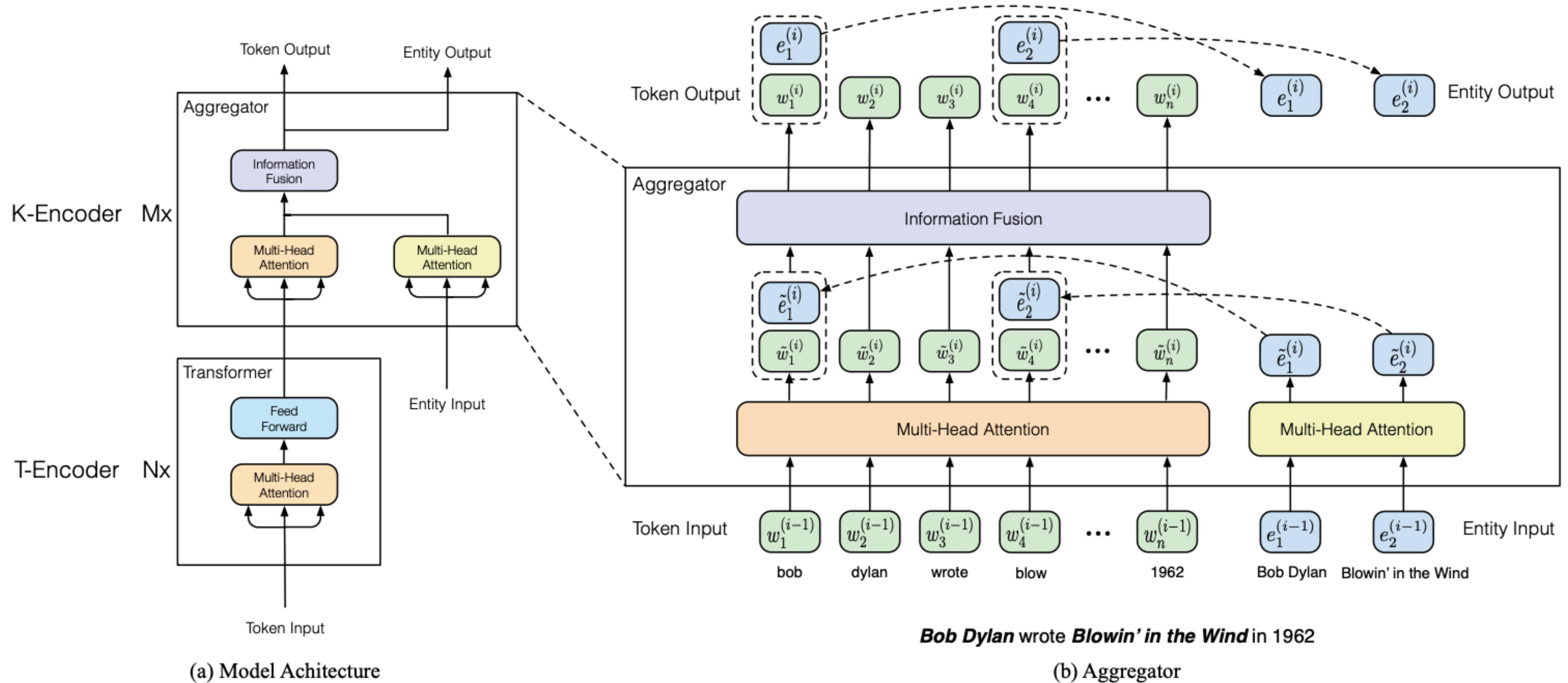
- 다양한 소스로부터 축적한 Knowledge Base를 Graph 형태로 나타낸 것으로, Google이 검색 결과를 향상시키기 위해 이용 중
- Bob Dylan이라는 Entity가 포함된 서로 다른 문장 두 개에서, writer에 관련된 것과 songwriter에 관련된 것 두 가지 facts를 extract 할 수 있음
- 이 두 facts를 graph 형태로 정보를 가지고 있고 그 정보를 Language Representation에 이용한다면 조금 더 좋은 성능 기대 가능!

*Bob Dylan wrote **Blowin' in the Wind** in 1962, and wrote **Chronicles: Volume One** in 2004.*

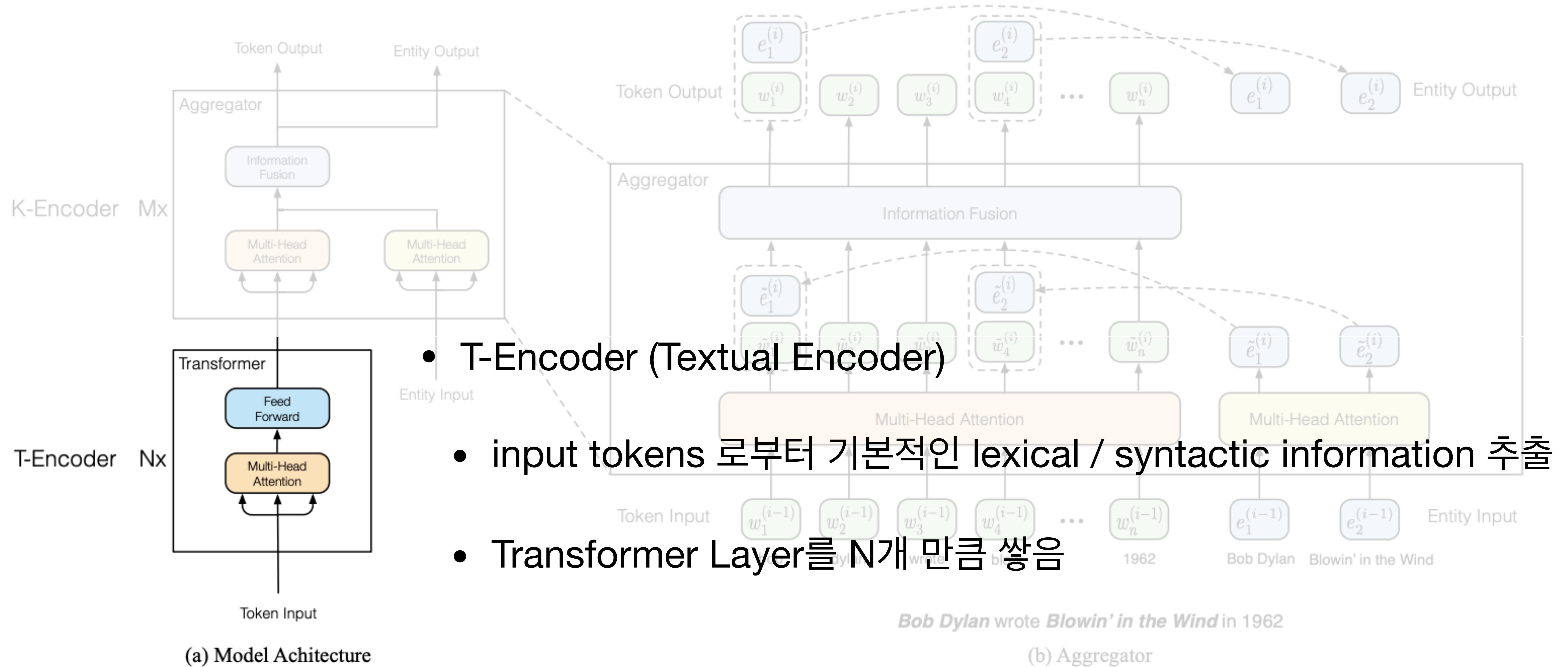
Introduction

- 이 논문에서는 KG 활용을 위해 필요한 다음과 같은 두 가지의 challenge 를 해결
 - **(1) Structured Knowledge Encoding**
 - Graph 형태의 structured knowledge를 어떻게 encoding 하여 사용할 것인가
 - **(2) Heterogeneous Information Fusion**
 - encoding 된 structured knowledge 와 각 token에 대응하는 representation 정보를 어떻게 섞을 것인가

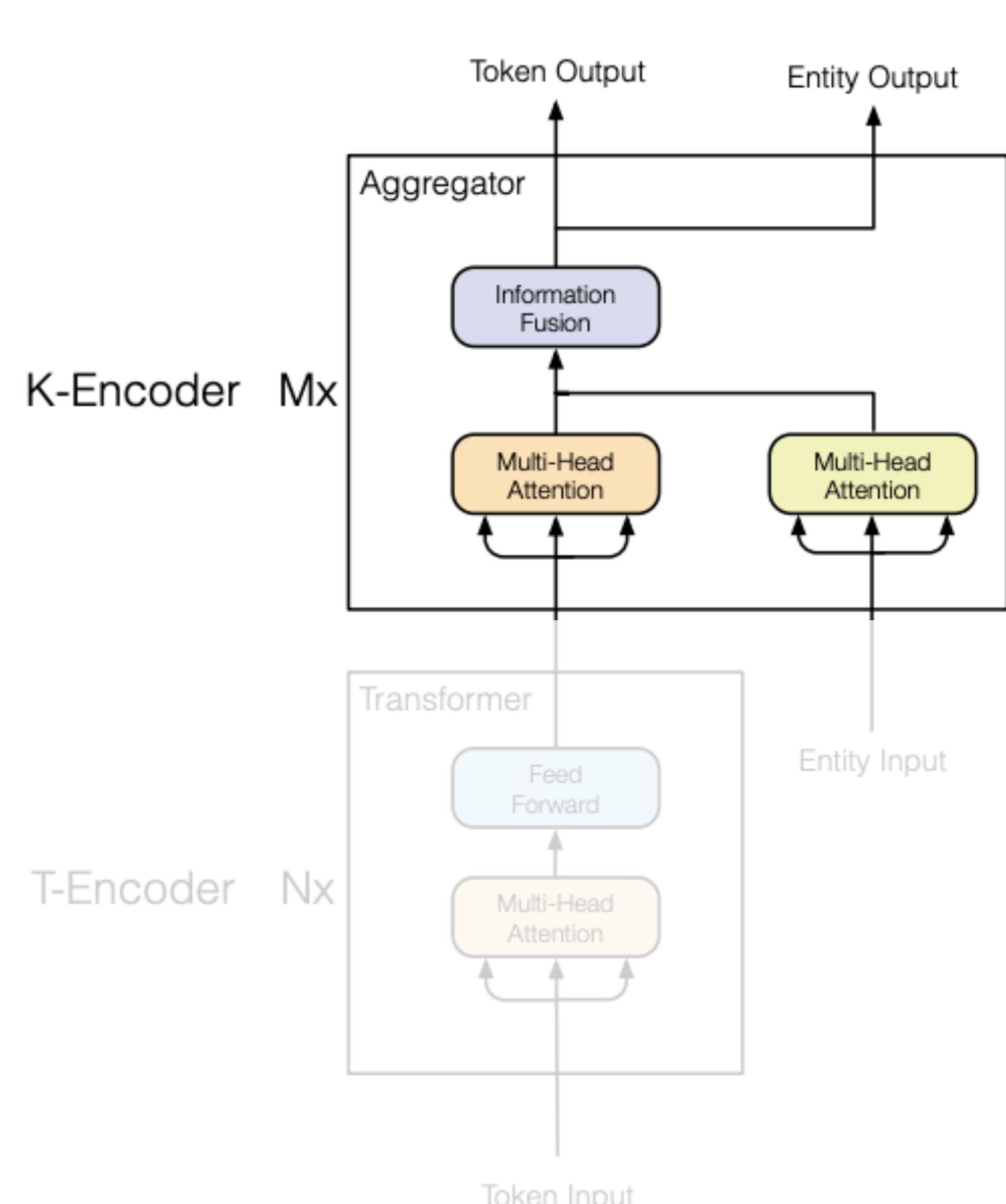
Model & Methodology



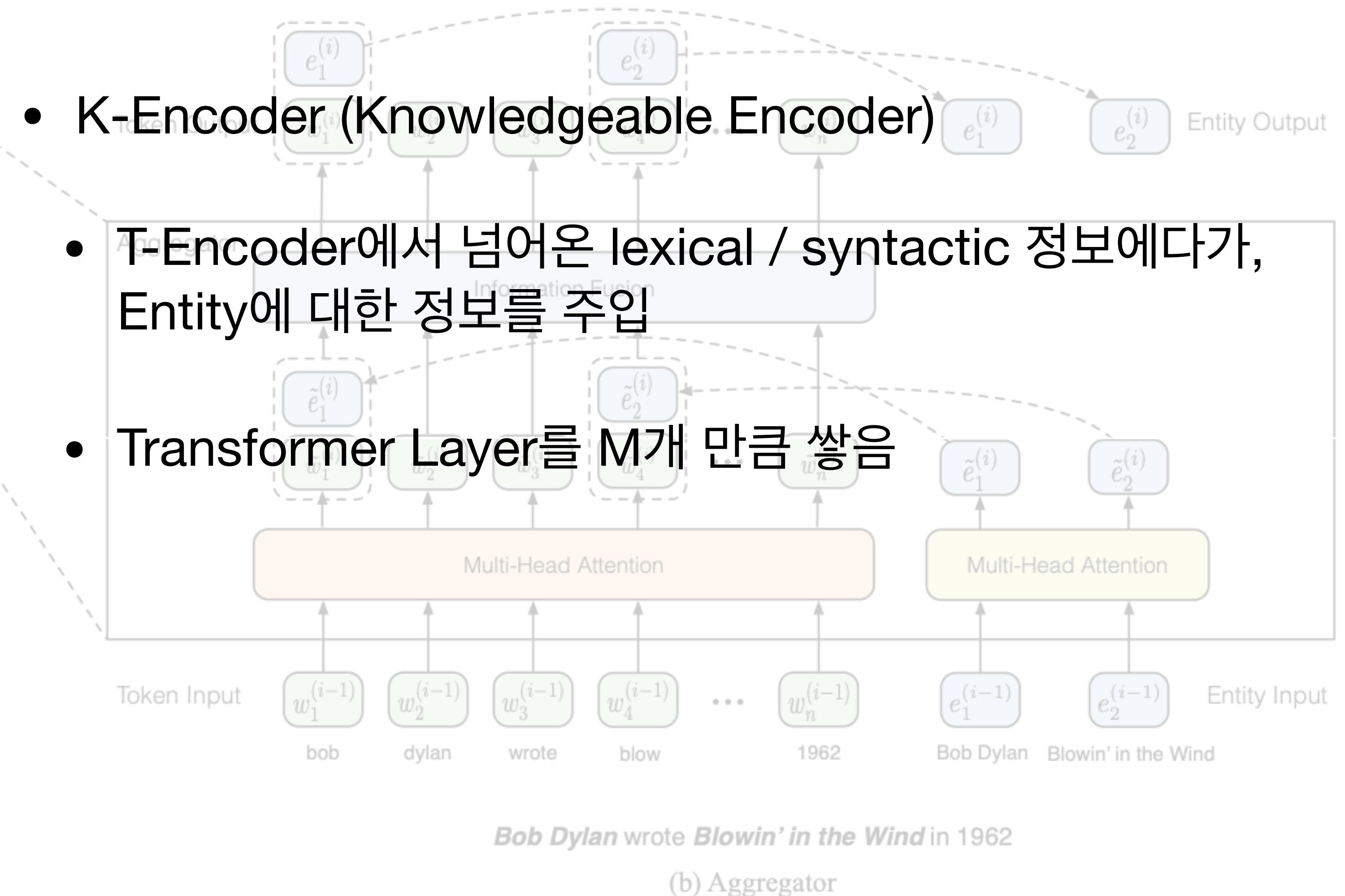
Model & Methodology



Model & Methodology



(a) Model Architecture



(b) Aggregator

- K-Encoder (Knowledgeable Encoder)
- T-Encoder에서 넘어온 lexical / syntactic 정보에다가, Entity에 대한 정보를 주입
- Transformer Layer를 M 개 만큼 쌓음

Model & Methodology

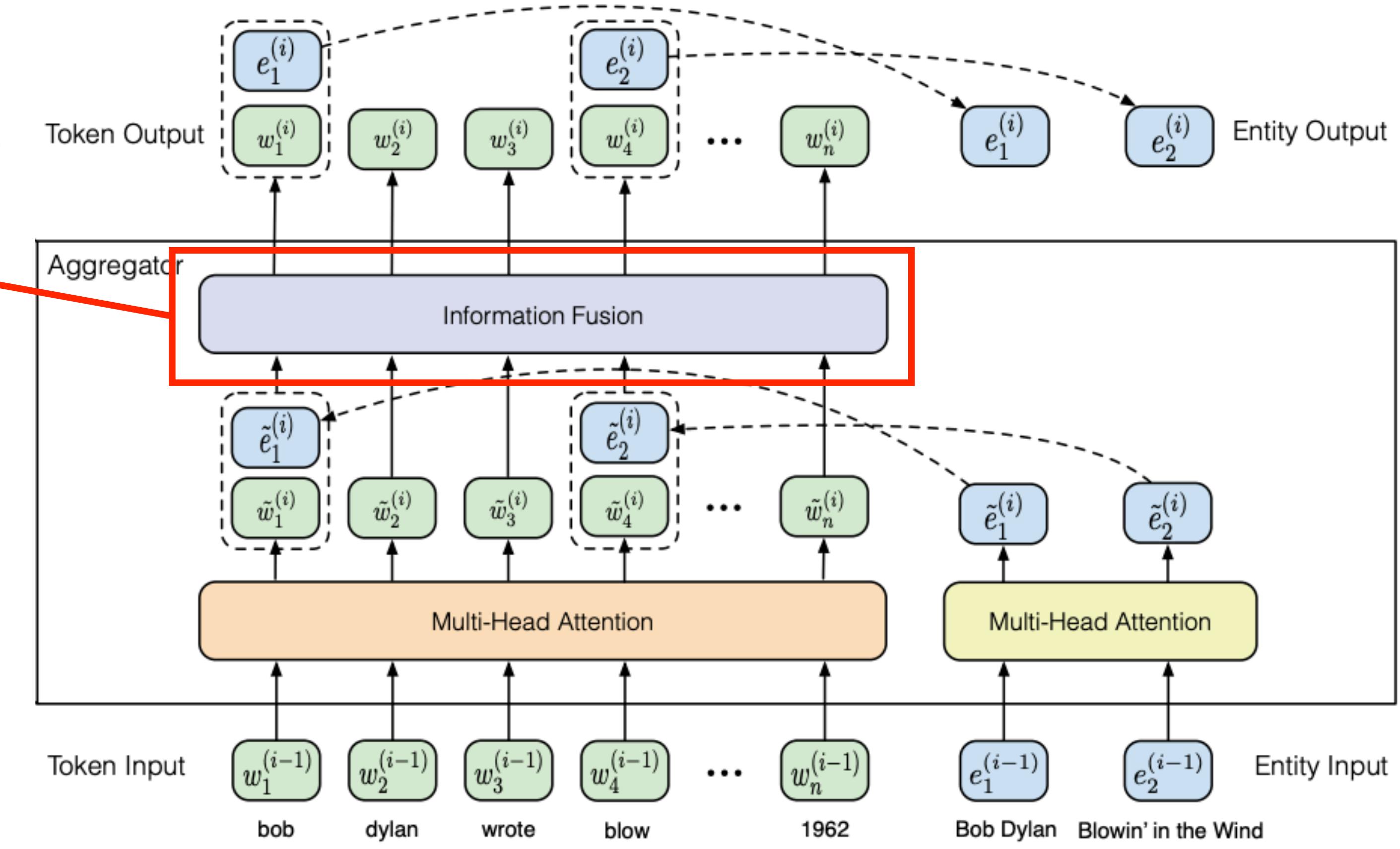
$$\mathbf{h}_j = \sigma(\tilde{\mathbf{W}}_t^{(i)} \tilde{\mathbf{w}}_j^{(i)} + \tilde{\mathbf{W}}_e^{(i)} \tilde{\mathbf{e}}_k^{(i)} + \tilde{\mathbf{b}}^{(i)}),$$

$$\mathbf{w}_j^{(i)} = \sigma(\mathbf{W}_t^{(i)} \mathbf{h}_j + \mathbf{b}_t^{(i)}),$$

$$\mathbf{e}_k^{(i)} = \sigma(\mathbf{W}_e^{(i)} \mathbf{h}_j + \mathbf{b}_e^{(i)}).$$

sigma: GELU Function
 (ReLU 보다 조금 더 부드러운 곡선의 함수)

- token과 entity의 정보를 동시에 갖는 weight matrix \mathbf{h}_j 를 통해 두 가지 정보 aggregate



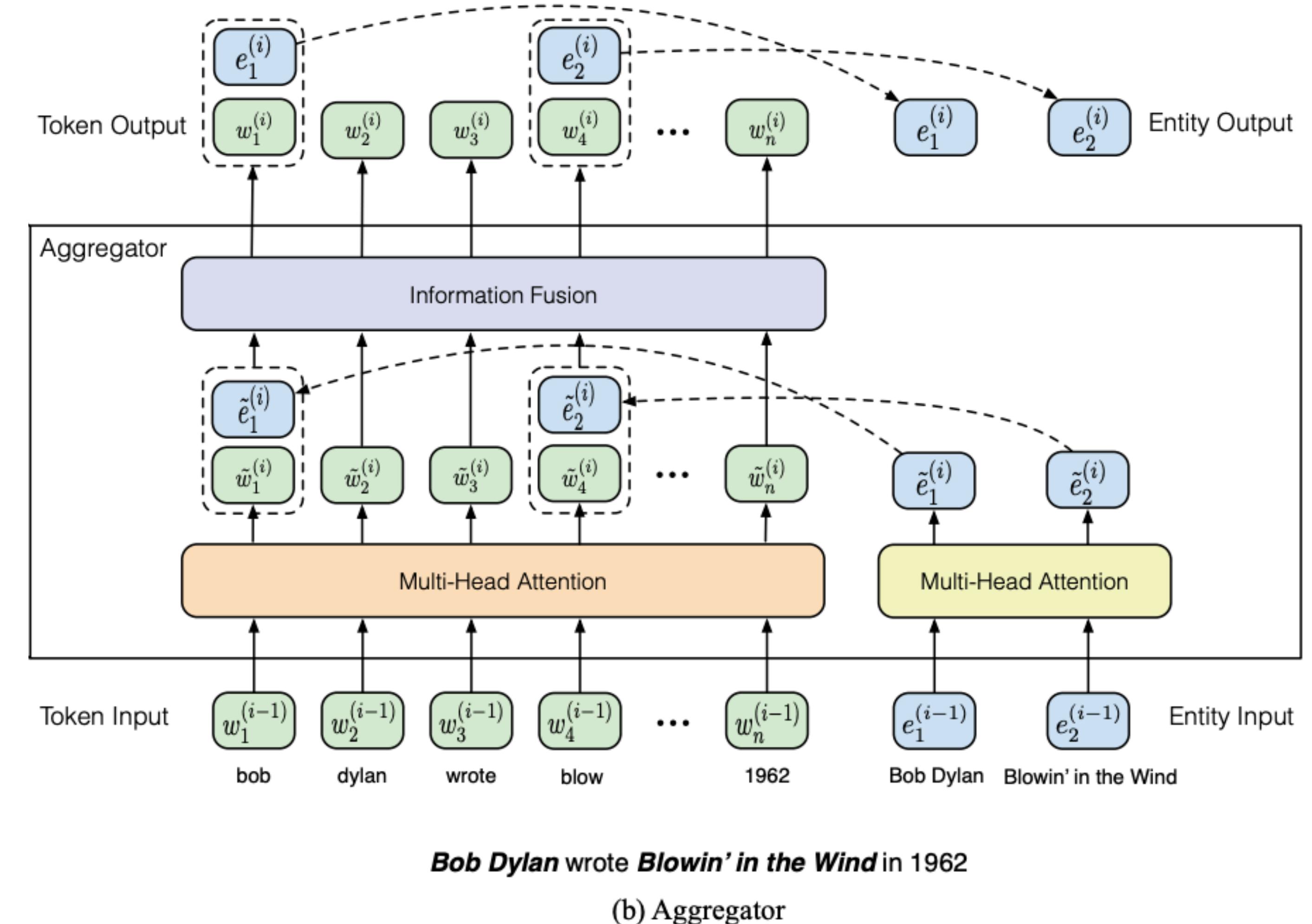
Bob Dylan wrote Blowin' in the Wind in 1962

(b) Aggregator

Model & Methodology

$$p(e_j | w_i) = \frac{\exp(\text{linear}(\mathbf{w}_i^o) \cdot \mathbf{e}_j)}{\sum_{k=1}^m \exp(\text{linear}(\mathbf{w}_i^o) \cdot \mathbf{e}_k)},$$

- token-entity alignment 중 일부를 랜덤하게 mask 시키고, 그것을 잘 복원해내도록 cross-entropy loss function 이용해 학습
- like denoising AutoEncoder



Model & Methodology

- Knowledge Graph는 Wikidata 라는 곳에서 가져옴
(5,040,986 entities, 24,267,796 fact triplets)
- ERNIE model의 초기 parameter는 Google이 학습시켜 놓은 BERT에서 가져옴
- Knowledge Graph의 Embedding은 TransE (Bordes et al., 2013) 알고리즘 사용

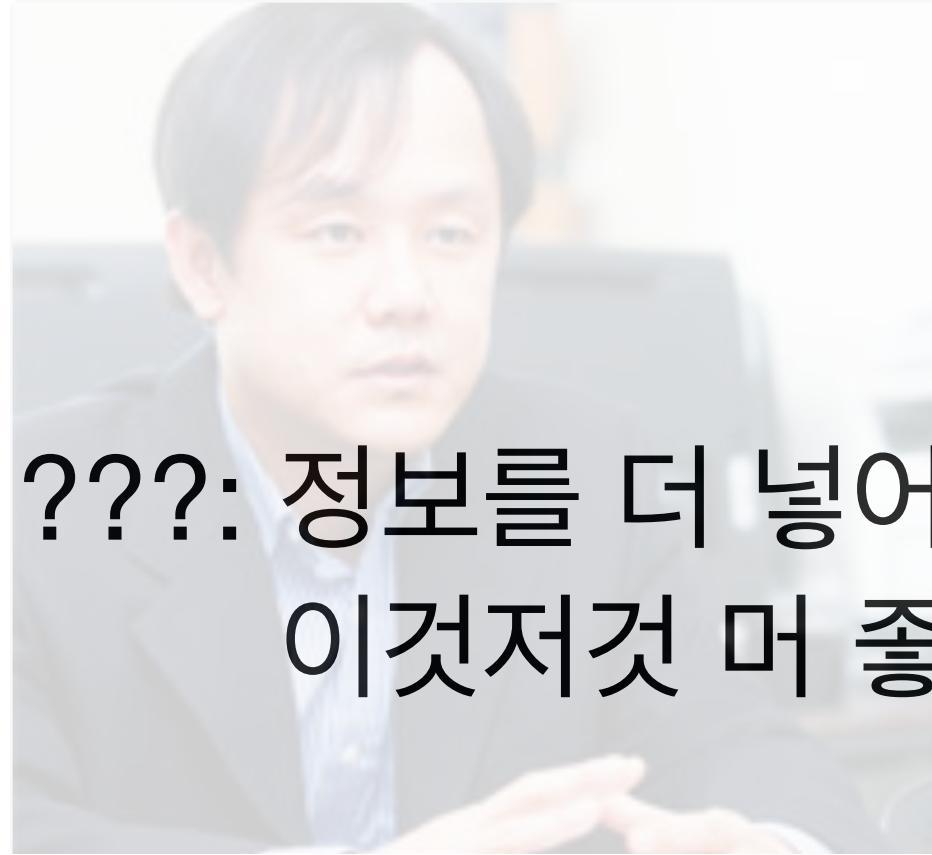
Experiments

Model	FewRel			TACRED		
	P	R	F1	P	R	F1
CNN	69.51	69.64	69.35	70.30	54.20	61.20
PA-LSTM	-	-	-	65.70	64.50	65.10
C-GCN	-	-	-	69.90	63.30	66.40
BERT	85.05	85.11	84.89	67.23	64.81	66.00
ERNIE	88.49	88.44	88.32	69.97	66.08	67.97

Table 5: Results of various models on FewRel and TACRED.

Relation Classification / Extraction Dataset에서
기존 state-of-the-art 모델인 BERT보다 좋은 성능을 보임

Conclusion



- ????: 정보를 더 넣어주면 성능은 좋아질 수 밖에 없지
이것저것 머 좋다는 거 다 때리박아가 했단 말이가
- 의 아주 좋은 표본!
- 다만 개인적으로, 사람들의 일반적인 언어 학습과 사용 방식을 오로지 확률에만 기반한 Language Model로 완벽히 모사를 할 수 있을까 의문이 들었었는데 그에 대한 힌트를 조금이나마 주는 논문이었음

Thank you!