한국어 임베딩

컴퓨터는 자연어의 의미를 어디까지 이해할 수 있을까

NAVER Chatbot Model
이기창

목차

임베딩이란

- 단어나 문장을 벡터로 바꾼 것 혹은 그 과정

임베딩이란

- 단어나 문장을 벡터로 바꾼 것 혹은 그 과정

구분	메밀꽃 필 무렵	운수 좋은 날	사랑 손님과 어머니	삼포 가는 길
기차	0	2	10	7
막걸리	0	1	0	0
선술집	0	1	0	0

임베딩이란

- 단어나 문장을 벡터로 바꾼 것 혹은 그 과정

구분	메밀꽃 필 무렵	운수 좋은 날	사랑 손님과 어머니	삼포 가는 길
기차	0	2	10	7
막걸리	0	1	0	0
선술집	0	1	0	0

단어 임베딩

문서 임베딩

임베딩이란

- ‘희망’이라는 단어의 Word2Vec 임베딩

$$[-0.00209\ -0.03918\ 0.02419\ \ldots\ 0.01715\ -0.04975\ 0.09300]$$

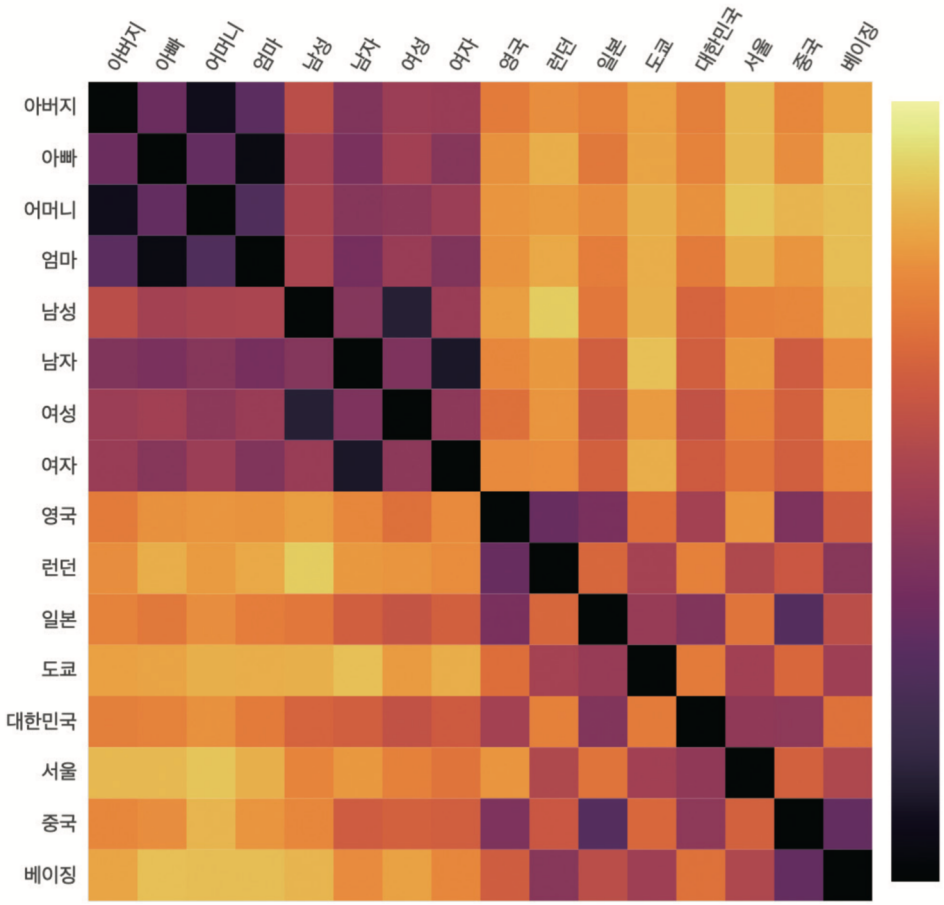
임베딩이란

- 임베딩으로 할 수 있는 것 : 관련도/유사도 계산

희망	절망	학교	학생	가족	자동차
소망	체념	초등	대학생	아이	승용차
행복	고뇌	중학교	대학원생	부모	상용차
희망찬	절망감	고등학교	고학생	편부모	트럭
꿈	상실감	야학교	교직원	고달픈	대형트럭
열망	번민	중학	학부모	사랑	모터사이클

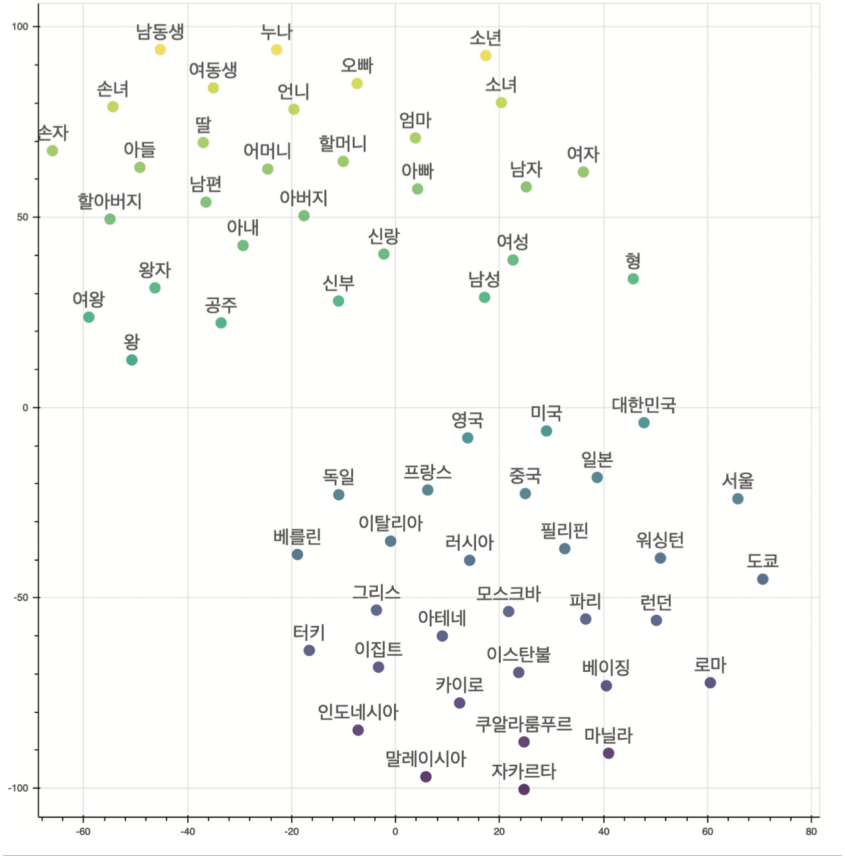
임베딩이란

- 임베딩으로 할 수 있는 것 : 관련도/유사도 계산



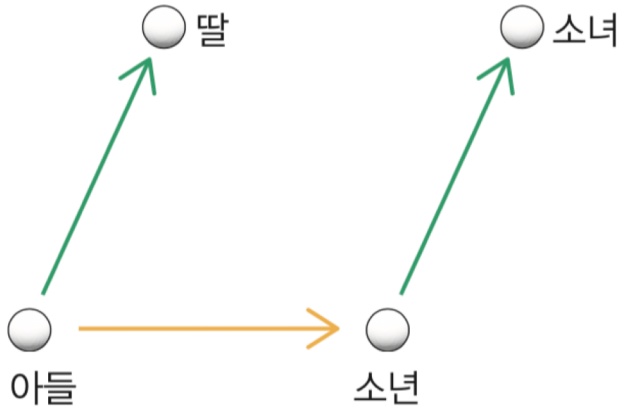
임베딩이란

- 임베딩으로 할 수 있는 것 : 시각화



임베딩이란

- **벡터 연산(유추 평가)** : 아들 - 딸 + 소녀 = 소년



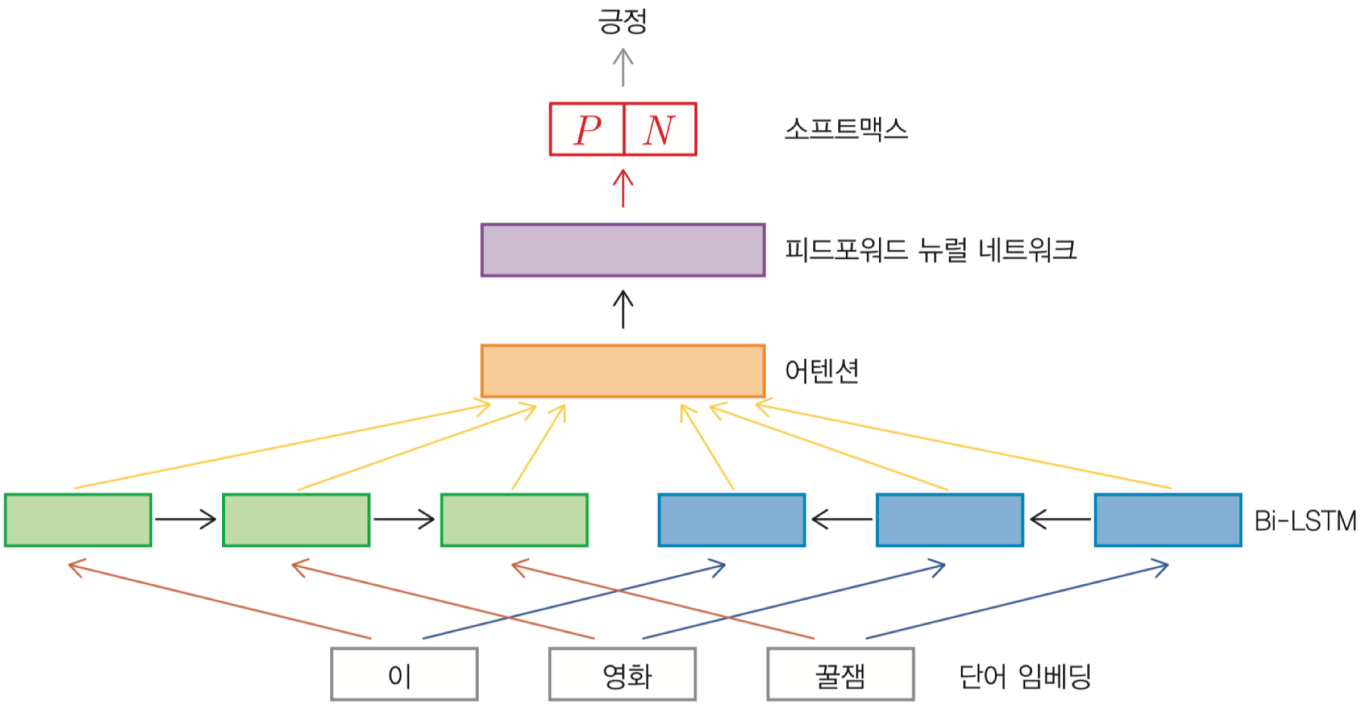
임베딩이란

- **벡터 연산(유추 평가)** : 아들 - 딸 + 소녀 = 소년

단어1	단어2	단어3	결과
아들	딸	소년	**소녀**
아들	딸	아빠	**엄마**
아들	딸	남성	**여성**
남동생	여동생	소년	**소녀**
남동생	여동생	아빠	**엄마**
남동생	여동생	남성	**여성**
신랑	신부	왕	**여왕**
신랑	신부	손자	**손녀**
신랑	신부	아빠	**엄마**

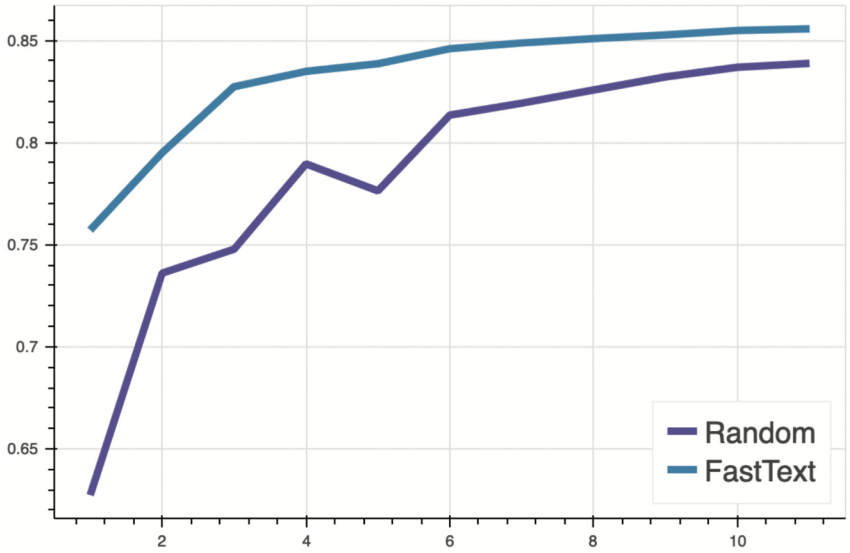
임베딩이란

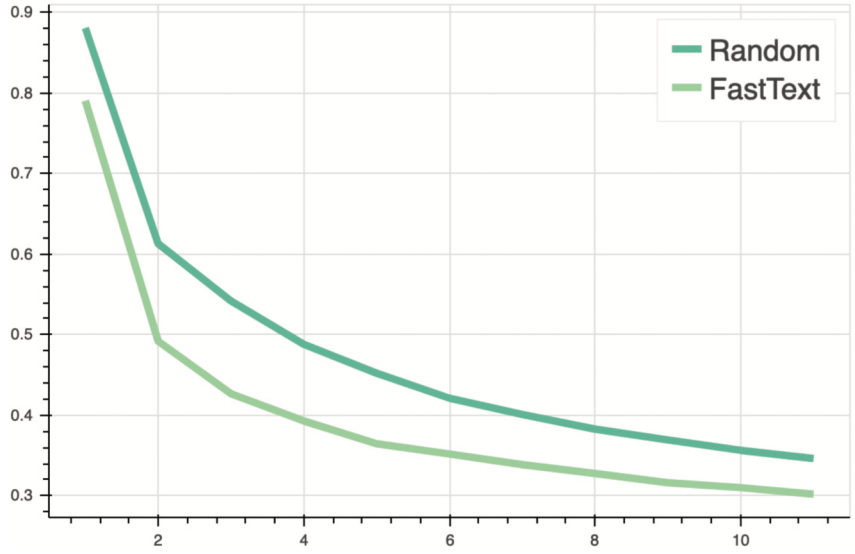
- **전이학습(transfer learning)** : 다른 딥러닝 모델의 입력값



임베딩이란

- **전이학습(transfer learning)** : 다른 딥러닝 모델의 입력값





임베딩에 어떻게 의미를 함축하는가

- 임베딩과 관련한 세 가지 철학

구분	백오브워즈 가정	언어 모델	분포 가정
내용	어떤 단어가 (많이) 쓰였는가	단어가 어떤 순서로 쓰였는가	어떤 단어가 같이 쓰였는가
대표 통계량	TF-IDF	–	PMI
대표 모델	Deep Averaging Network	ELMo, GPT	Word2Vec

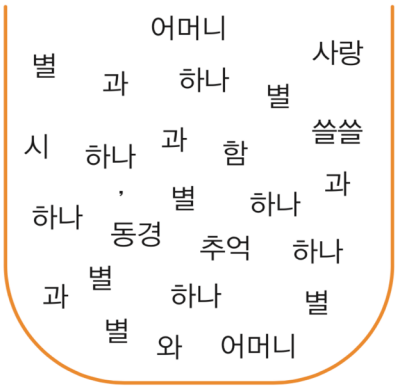
임베딩에 어떻게 의미를 함축하는가

- 백오브워즈 : 어떤 단어가 (많이)쓰였는가

문서의 주제는 문서 내 단어 사용 양상에 드러난다
순서 정보 무시, 어떤 단어가 (많이)쓰였는가가 중요

별 하나 에 추억 과
별 하나 에 사랑 과
별 하나 에 쓸쓸함 과
별 하나 에 동경 과
별 하나 에 시 와
별 하나 에 어머니 , 어머니

⇨



⇩

별	하나	에	추억	과	사랑	쓸쓸	함	동경	시	와	어머니	,
6	6	6	1	4	1	1	1	1	1	1	2	1

임베딩에 어떻게 의미를 함축하는가

- 백오브워즈 : 어떤 단어가 (많이)쓰였는가

문서의 주제는 문서 내 단어 사용 양상에 드러난다
순서 정보 무시, 어떤 단어가 (많이)쓰였는가가 중요

$$\mathrm{TF-IDF}(w) = \mathrm{TF}(w) \times \log(\frac{N}{\mathrm{DF}(w)})$$

구분	메밀꽃 필 무렵	운수 좋은 날	사랑 손님과 어머니	삼포 가는 길
담배	0.2603	0.2875	0.0364	0.2932
를	0.0	0.0034	0.0	0.0

임베딩에 어떻게 의미를 함축하는가

- 백오브워즈 : 어떤 단어가 (많이)쓰였는가

문서의 주제는 문서 내 단어 사용 양상에 드러난다
순서 정보 무시, 어떤 단어가 (많이)쓰였는가가 중요

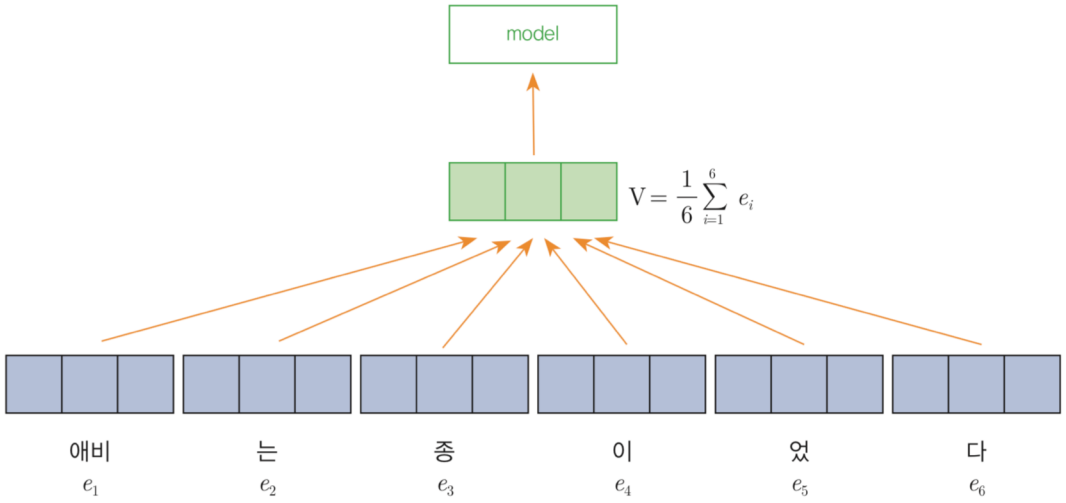


그림 2-2 Deep Averaging Network

임베딩에 어떻게 의미를 함축하는가

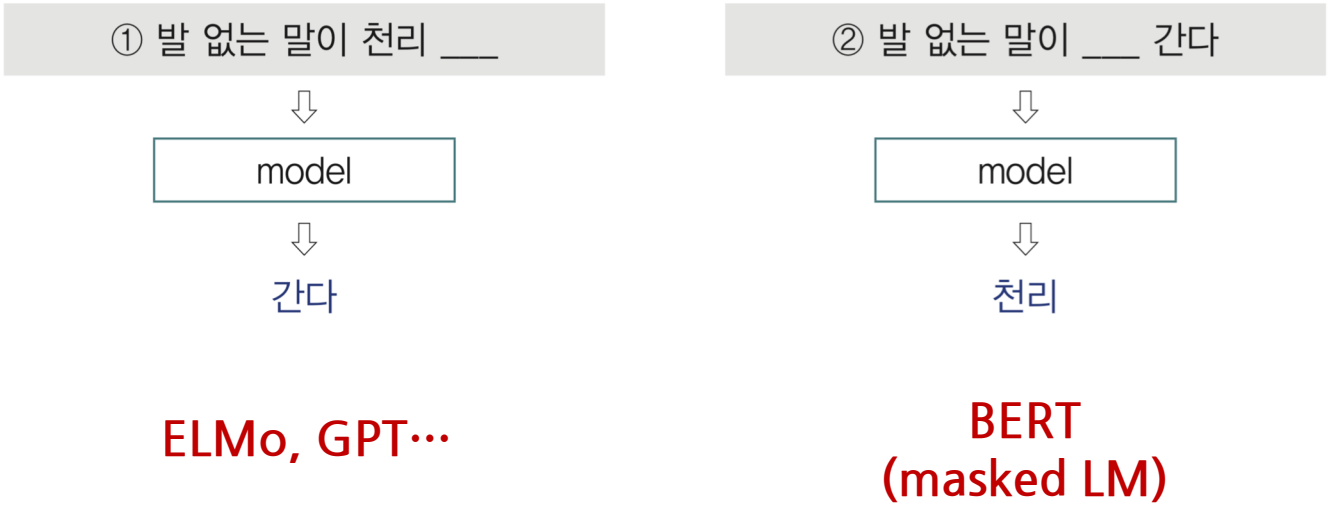
- 언어모델 : 단어가 어떤 순서로 쓰였는가

  단어 등장 순서 정보를 명시적으로 학습

임베딩에 어떻게 의미를 함축하는가

- 언어모델 : 단어가 어떤 순서로 쓰였는가

단어 등장 순서 정보를 명시적으로 학습



임베딩에 어떻게 의미를 함축하는가

- 분포 가정 : 어떤 단어가 같이 쓰였는가

자연어의 의미는 그 주변 문맥을 통해 유추해 볼 수 있다

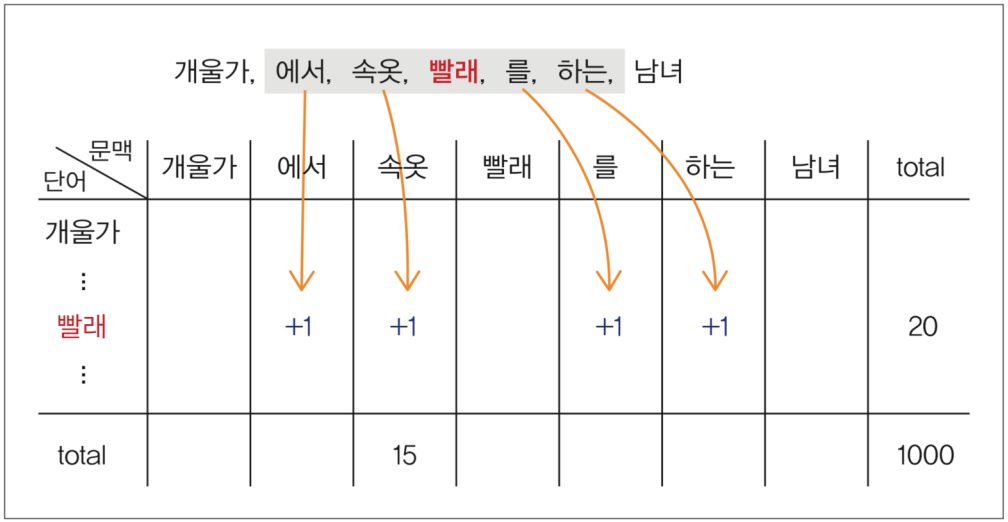
체언(명사): 관형사가 그 앞에 올 수 있고 조사가 그 뒤에 올 수 있음
용언(동사/형용사): 부사가 그 앞에 올 수 있고 선어말어미가 그 뒤에 올 수 있고 어말어미가 그 뒤에 와야 함
관형사: 명사가 그 뒤에 와야 함
부사: 용언, 부사, 절이 그 뒤에 와야 함
조사: 체언 뒤에 와야 함
어미: 용언 뒤에 와야 함
감탄사(간투사): 특별한 결합 제약 없이 즉, 문장 내의 다른 단어와 문법적 관계를 맺지 않고 따로 존재함

임베딩에 어떻게 의미를 함축하는가

- 분포 가정 : 어떤 단어가 같이 쓰였는가

> 자연어의 의미는 그 주변 문맥을 통해 유추해 볼 수 있다

window = 2인 단어-문맥 행렬

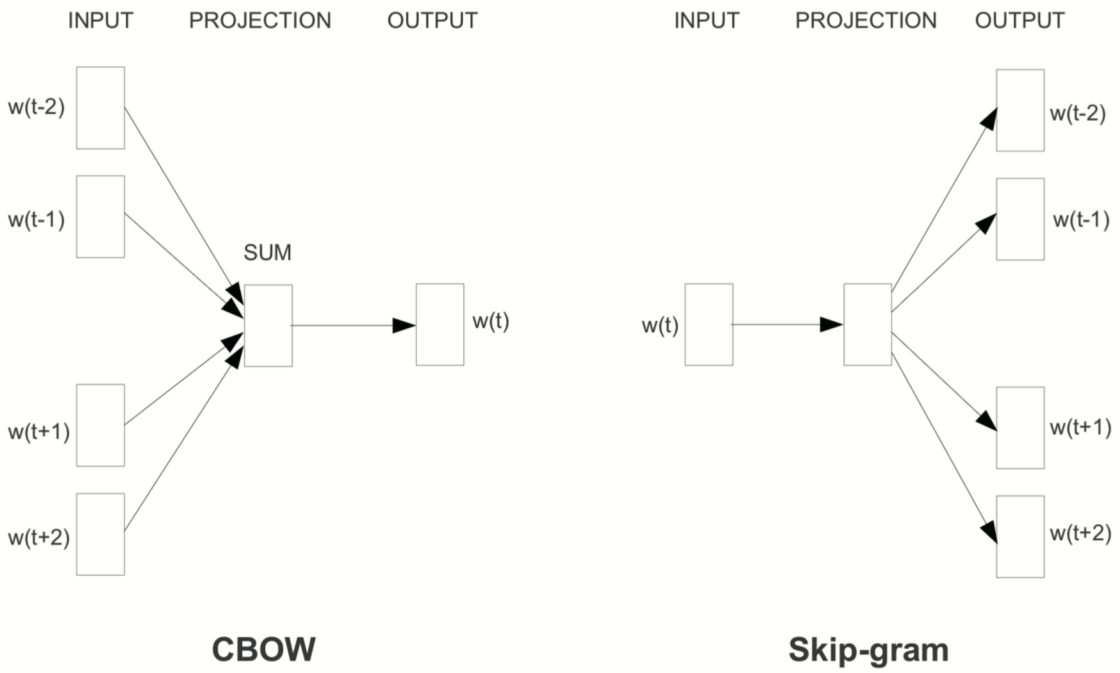


$$\mathrm{PMI}(A,B) = \log \frac{P(A,B)}{P(A) \times P(B)}$$

임베딩에 어떻게 의미를 함축하는가

- 분포 가정 : 어떤 단어가 같이 쓰였는가

> 자연어의 의미는 그 주변 문맥을 통해 유추해 볼 수 있다



임베딩에 어떻게 의미를 함축하는가

- 임베딩과 관련한 세 가지 철학

세 철학은 말뭉치의 통계적 패턴을 서로 다른 각도에서 분석
상호 보완적

구분	백오브워즈 가정	언어 모델	분포 가정
내용	어떤 단어가 (많이) 쓰였는가	단어가 어떤 순서로 쓰였는가	어떤 단어가 같이 쓰였는가
대표 통계량	TF-IDF	–	PMI
대표 모델	Deep Averaging Network	ELMo, GPT	Word2Vec

단어 수준 임베딩

- Word2Vec

Skip-Gram with Negative Sampling

... 개울가 (에서 속옷 **빨래** 를 하는) 남녀 ...

c_1 c_2 t c_3 c_4

포지티브 샘플

t	*c*
빨래	에서
빨래	속옷
빨래	를
빨래	하는

네거티브 샘플

t	*c*
빨래	책상
빨래	안녕
빨래	자동차
빨래	숫자

t	*c*
빨래	커피
빨래	떡
빨래	사과
빨래	노트북

단어 수준 임베딩

- Word2Vec

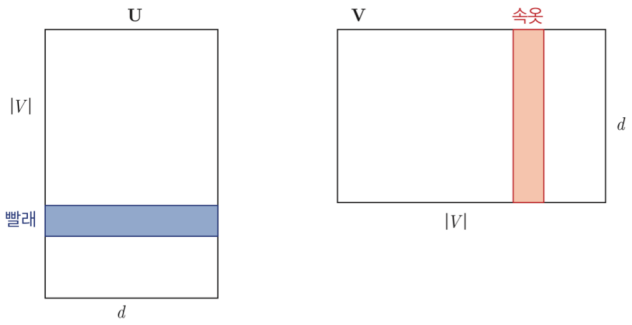
Skip-Gram with Negative Sampling

t, c가 포지티브 샘플($=t$ 주변에 c가 존재)일 확률

$$P(+|t,c) = \frac{1}{1+\exp(-\mathbf{u}_t\mathbf{v}_c)}$$

t, c가 네거티브 샘플(c를 t와 무관하게 말뭉치 전체에서 랜덤 샘플)일 확률

$$P(-|t,c) = 1 - P(+|t,c) = \frac{\exp(-\mathbf{u}_t\mathbf{v}_c)}{1+\exp(-\mathbf{u}_t\mathbf{v}_c)}$$



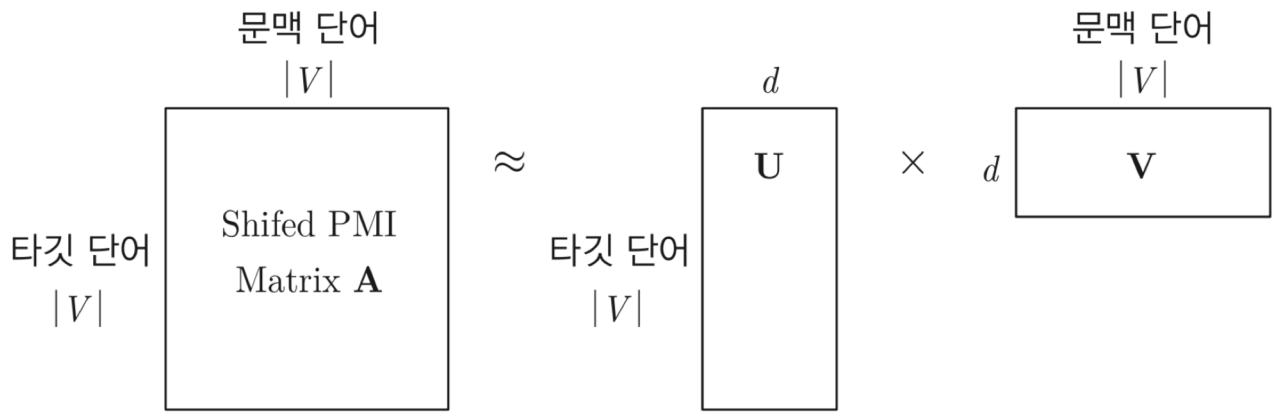
단어 수준 임베딩

- Word2Vec

Skip-Gram with Negative Sampling

행렬 분해 관점에서 이해하는 Word2Vec

$$\mathbf{A}_{ij}^{\mathrm{SGNS}} = \mathbf{U}_i \cdot \mathbf{V}_j = \mathrm{PMI}(i,j) - \log k$$

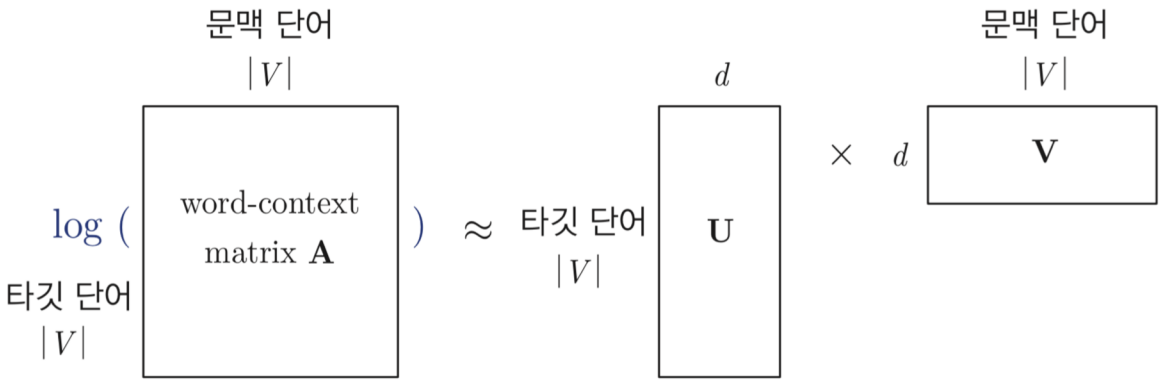


Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems* (pp. 2177-2185).

단어 수준 임베딩

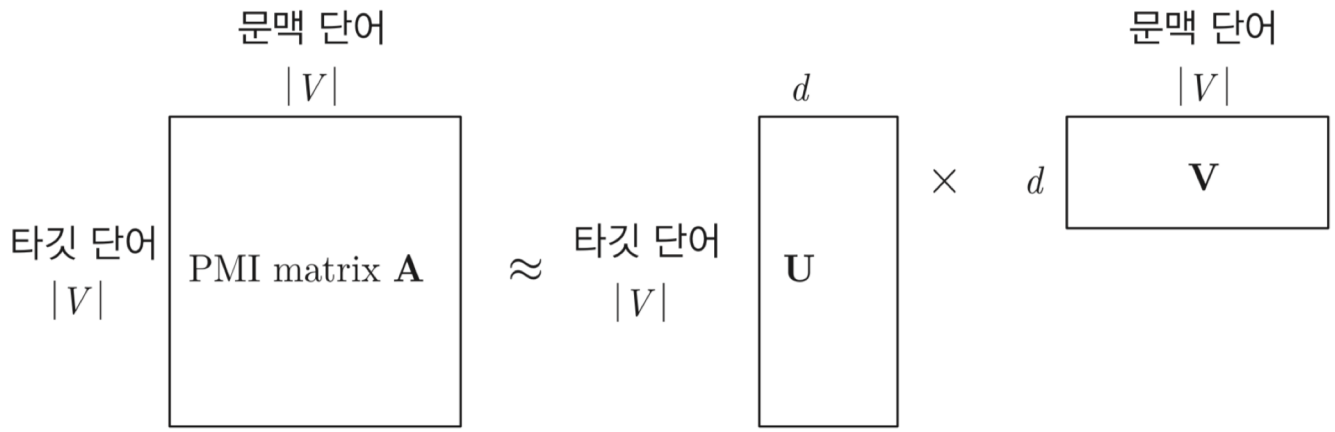
- GloVe

$$\mathcal{J} = \sum_{i,j=1}^{|V|} f(\mathbf{A}_{ij})(\mathbf{U}_i \cdot \mathbf{V}_j + \mathbf{b}_i + \mathbf{b}_j - \log \mathbf{A}_{ij})^2$$



단어 수준 임베딩

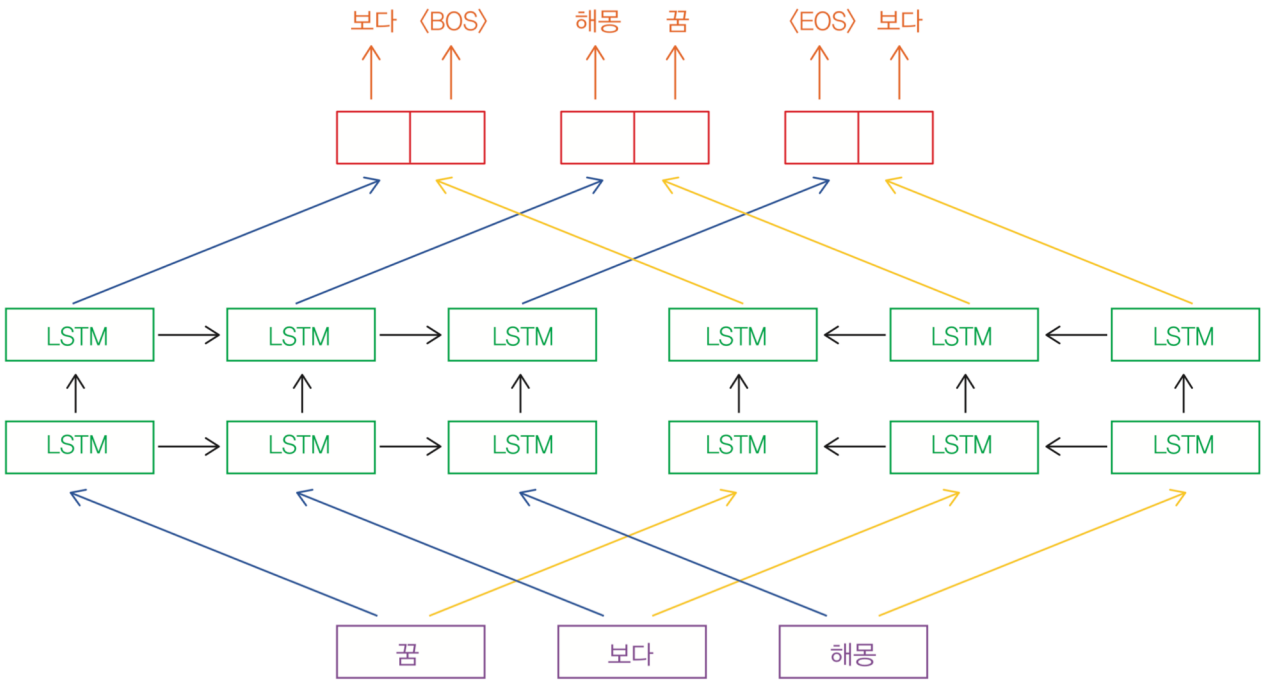
- Swivel



문장 수준 임베딩

- ELMo

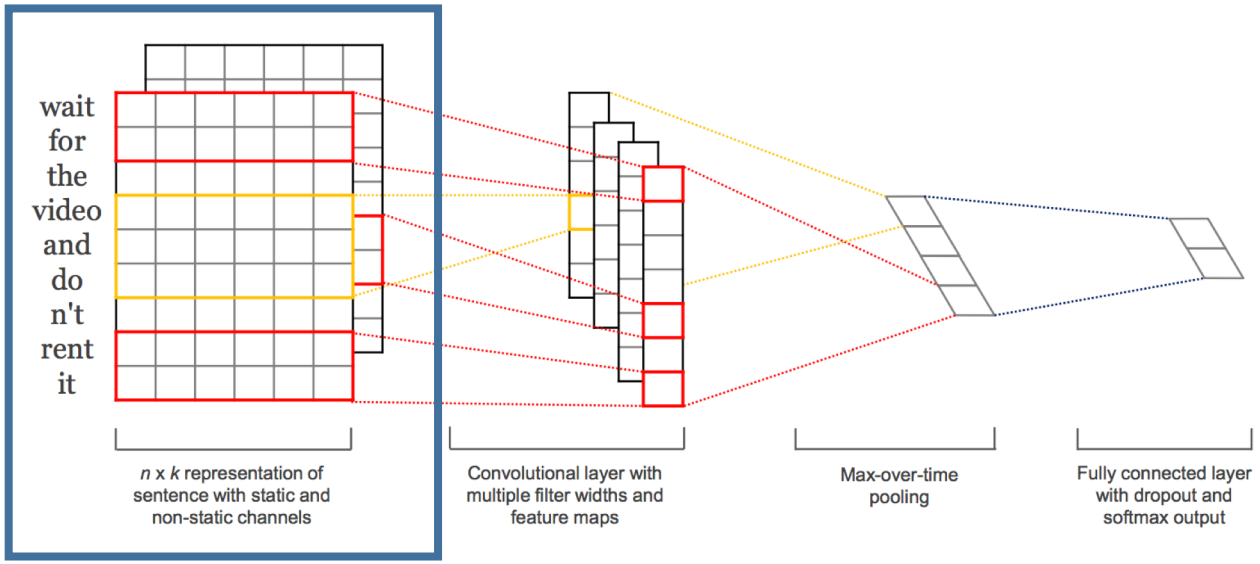
CNN + Bi-LSTM



문장 수준 임베딩

- CNN이 포착하는 정보

Conv filter : ngram detector (특정 구의 존재/부재 확인)

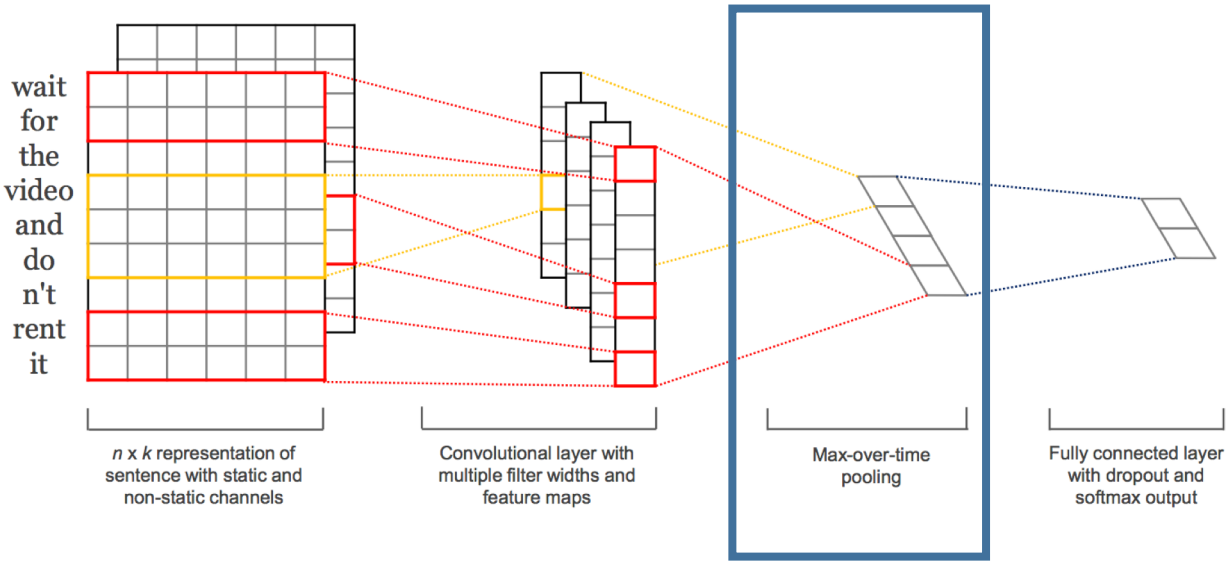


Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

문장 수준 임베딩

- CNN이 포착하는 정보

Max-pooling : threshold filter (불필요한 구 버림)

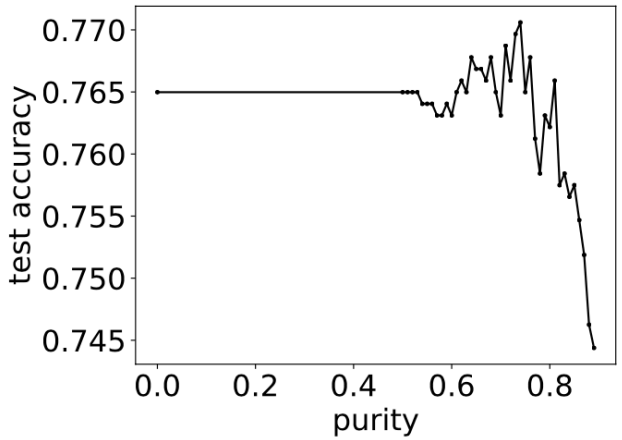


Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

문장 수준 임베딩

- CNN이 포착하는 정보

Max-pooling : threshold filter (불필요한 구 버림)

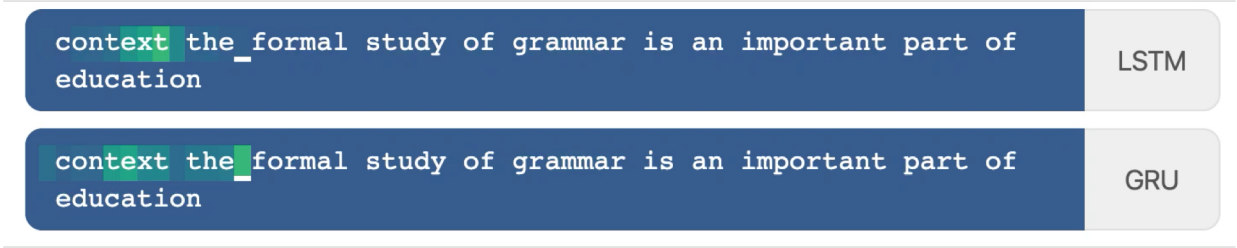


Jacovi, A., Shalom, O. S., & Goldberg, Y. (2018). Understanding convolutional neural networks for text classification. *arXiv preprint arXiv:1809.08037*.

문장 수준 임베딩

- LSTM이 포착하는 정보

Long term dependency (left to right)



https://distill.pub/2019/memorization-in-rnns

문장 수준 임베딩

- ELMo

Fine-tuning 때 학습, 토큰별로 생성

$$\mathrm{ELMo}_k^{\mathrm{task}} = \gamma^{\mathrm{task}} \sum_{j=0}^{L} s_j^{\mathrm{task}} \mathbf{h}_{k,j}^{\mathrm{LM}}$$

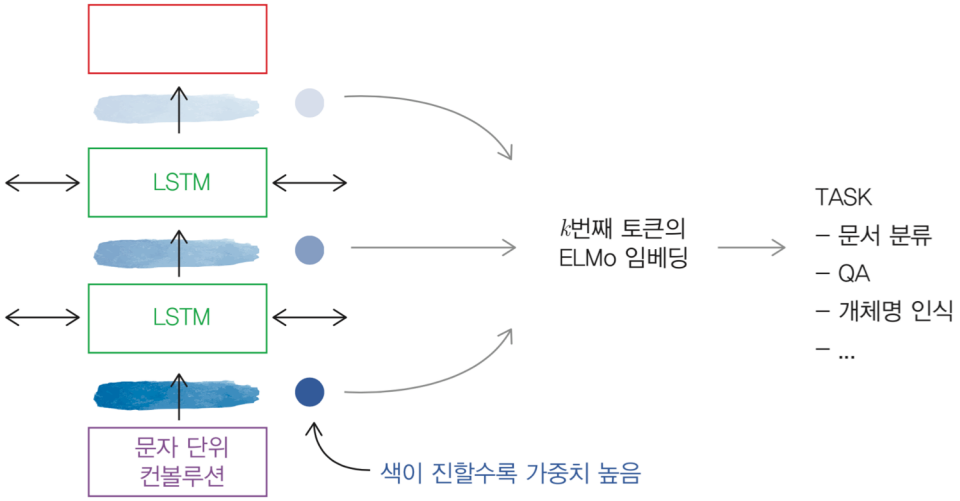


그림 5-19 ELMo 임베딩

문장 수준 임베딩

- Transformer block

Self-Attention
Scale
Dot-Product
Multi-head

문장 수준 임베딩

- Transformer block

Self-Attention (Q=K, 문장 길이 관계없이 모든 쌍의 관계 고려)
Scale
Dot-Product
Multi-head

$$\text{softmax}(\frac{\mathbf{QK}^\top}{\sqrt{d_k}})\mathbf{V} = \begin{matrix} \\ \text{드디어} \\ \text{금요일} \\ \text{이다} \end{matrix} \begin{matrix} \begin{matrix} \text{드디어} & \text{금요일} & \text{이다} \end{matrix} \\ \begin{pmatrix} 0.2 & 0.7 & 0.1 \\ \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots \end{pmatrix} \end{matrix} \begin{pmatrix} \mathbf{V}_{\text{드디어}} \\ \mathbf{V}_{\text{금요일}} \\ \mathbf{V}_{\text{이다}} \end{pmatrix}$$

$$= \begin{matrix} \text{드디어} \\ \text{금요일} \\ \text{이다} \end{matrix} \begin{pmatrix} 0.2\mathbf{V}_{\text{드디어}} + 0.7\mathbf{V}_{\text{금요일}} + 0.1\mathbf{V}_{\text{이다}} \\ \cdots \\ \cdots \end{pmatrix}$$

그림 5-22 Scaled Dot-Product Attention 예시

문장 수준 임베딩

- Transformer block

Self-Attention
Scale
Dot-Product
Multi-head

$$\text{softmax}(\frac{\mathbf{QK}^\top}{\sqrt{d_k}})\mathbf{V} = \begin{array}{c} \\ \text{드디어} \\ \text{금요일} \\ \text{이다} \end{array} \begin{array}{c} \begin{array}{ccc} \text{드디어} & \text{금요일} & \text{이다} \end{array} \\ \begin{pmatrix} 0.2 & 0.7 & 0.1 \\ \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots \end{pmatrix} \end{array} \begin{pmatrix} \mathbf{V}_{\text{드디어}} \\ \mathbf{V}_{\text{금요일}} \\ \mathbf{V}_{\text{이다}} \end{pmatrix}$$

$$= \begin{array}{c} \\ \text{드디어} \\ \text{금요일} \\ \text{이다} \end{array} \begin{pmatrix} 0.2\mathbf{V}_{\text{드디어}} + 0.7\mathbf{V}_{\text{금요일}} + 0.1\mathbf{V}_{\text{이다}} \\ \cdots \\ \cdots \end{pmatrix}$$

그림 5-22 Scaled Dot-Product Attention 예시

소프트맥스 그래디언트

$$\frac{\partial \mathbf{y}_i}{\partial \mathbf{x}_i} = \mathbf{y}_i(1 - \mathbf{y}_i)$$

$$\frac{\partial \mathbf{y}_i}{\partial \mathbf{x}_j} = -\mathbf{y}_i\mathbf{y}_j$$

문장 수준 임베딩

- Transformer block

Self-Attention
Scale
Dot-Product (Q, K 간 유사도 포착 + 계산 효율 높음)
Multi-head

$$\text{softmax}\left(\frac{\mathbf{QK}^\top}{\sqrt{d_k}}\right)\mathbf{V} = \begin{array}{c} \\ \text{드디어} \\ \text{금요일} \\ \text{이다} \end{array} \begin{array}{c} \begin{array}{ccc} \text{드디어} & \text{금요일} & \text{이다} \end{array} \\ \begin{pmatrix} 0.2 & 0.7 & 0.1 \\ \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots \end{pmatrix} \end{array} \begin{pmatrix} \mathbf{V}_{\text{드디어}} \\ \mathbf{V}_{\text{금요일}} \\ \mathbf{V}_{\text{이다}} \end{pmatrix}$$

$$= \begin{array}{c} \\ \text{드디어} \\ \text{금요일} \\ \text{이다} \end{array} \begin{pmatrix} 0.2\mathbf{V}_{\text{드디어}} + 0.7\mathbf{V}_{\text{금요일}} + 0.1\mathbf{V}_{\text{이다}} \\ \cdots \\ \cdots \end{pmatrix}$$

그림 5-22 Scaled Dot-Product Attention 예시

문장 수준 임베딩

- Transformer block

Self-Attention
Scale
Dot-Product
Multi-head (여러 명의 독자)

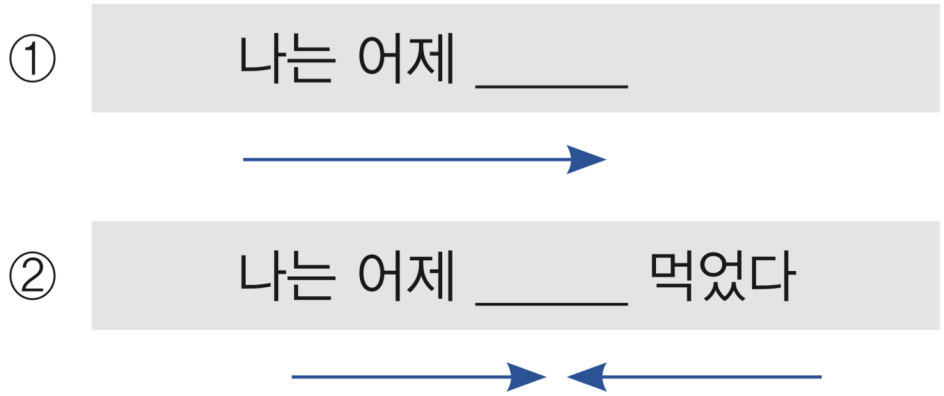
$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, ..., \text{head}_h)\mathbf{W}^{\text{O}}$$

$$\text{head}_i = \text{Attention}(\mathbf{Q}\mathbf{W}_i^{\text{Q}}, \mathbf{K}\mathbf{W}_i^{\text{K}}, \mathbf{V}\mathbf{W}_i^{\text{V}})$$

문장 수준 임베딩

- BERT

Bi-directional Language Model



문장 수준 임베딩

- BERT

Bi-directional Language Model

$$\text{softmax}(\frac{\mathbf{QK}^\top}{\sqrt{d_k}})\mathbf{V} = \begin{matrix} \\ \text{꿈} \\ \text{보다} \\ \text{해몽} \end{matrix} \begin{matrix} \begin{matrix} \text{꿈} & \text{보다} & \text{해몽} \end{matrix} \\ \begin{pmatrix} 0 & 0 & 0 \\ 1.0 & 0 & 0 \\ 0.9 & 0.1 & 0 \end{pmatrix} \end{matrix} \begin{pmatrix} \mathrm{V}_{\text{꿈}} \\ \mathrm{V}_{\text{보다}} \\ \mathrm{V}_{\text{해몽}} \end{pmatrix}$$

$$= \begin{matrix} \\ \text{꿈} \\ \text{보다} \\ \text{해몽} \end{matrix} \begin{pmatrix} 0.\mathrm{V}_{\text{꿈}} + 0.\mathrm{V}_{\text{보다}} + 0.\mathrm{V}_{\text{해몽}} \\ 1.0\mathrm{V}_{\text{꿈}} + 0.\mathrm{V}_{\text{보다}} + 0.\mathrm{V}_{\text{해몽}} \\ 0.9\mathrm{V}_{\text{꿈}} + 0.1\mathrm{V}_{\text{보다}} + 0.\mathrm{V}_{\text{해몽}} \end{pmatrix} \begin{matrix} \text{정답} \\ \text{꿈} \\ \text{보다} \\ \text{해몽} \end{matrix}$$

그림 5-27 GPT의 학습

문장 수준 임베딩

- BERT

Bi-directional Language Model

$$\text{softmax}(\frac{\mathbf{QK}^\top}{\sqrt{d_k}})\mathbf{V} = \begin{matrix} \\ \text{[MASK]} \\ \text{보다} \\ \text{해몽} \end{matrix} \begin{matrix} \begin{matrix} \text{[MASK]} & \text{보다} & \text{해몽} \end{matrix} \\ \begin{pmatrix} 0.3 & 0.1 & 0.6 \\ \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots \end{pmatrix} \end{matrix} \begin{pmatrix} \mathrm{V}_{\text{[MASK]}} \\ \mathrm{V}_{\text{보다}} \\ \mathrm{V}_{\text{해몽}} \end{pmatrix}$$

$$= \begin{matrix} \text{[MASK]} \\ \text{보다} \\ \text{해몽} \end{matrix} \begin{pmatrix} 0.3\mathrm{V}_{\text{[MASK]}} + 0.1\mathrm{V}_{\text{보다}} + 0.6\mathrm{V}_{\text{해몽}} \\ \cdots \\ \cdots \end{pmatrix}$$

정답

꿈

•

•

그림 5-28 BERT의 학습

문장 수준 임베딩

- BERT

Bi-directional Language Model

Tasks	Dev Set MNLI-m (Acc)	QNLI (Acc)	MRPC (Acc)	SST-2 (Acc)	SQuAD (F1)
$BERT_{BASE}$	84.4	88.4	86.7	92.7	88.5
No NSP	83.9	84.9	86.5	92.6	87.9
LTR & No NSP	82.1	84.3	77.5	92.1	77.8
+ BiLSTM	82.1	84.1	75.7	91.6	84.9

GPT

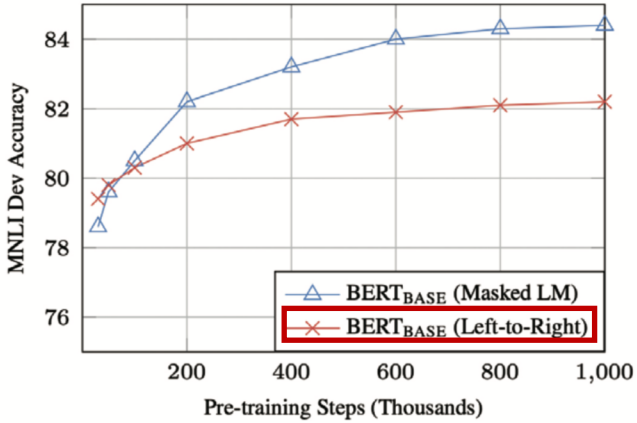


그림 5-29 BERT 대 GPT 성능 차이(Devlin et al., 2018)

문장 수준 임베딩

- BERT

Masked Language Model

- 전체 학습 데이터 토큰의 15%를 마스킹한다.
- 마스킹 대상 토큰 가운데 80%는 실제 빈칸으로 만들고, 모델은 그 빈칸을 채운다. 예: 발 없는 말이 [MASK] 간다 → 천리
- 마스킹 대상 토큰 가운데 10%는 랜덤으로 다른 토큰으로 대체하고, 모델은 해당 위치의 정답 단어가 무엇일지 맞추도록 한다. 예: 발 없는 말이 [컴퓨터] 간다 → 천리
- 마스킹 대상 토큰 가운데 10%는 토큰 그대로 두고, 모델은 해당 위치의 정답 단어가 무엇일지 맞추도록 한다. 예: 발 없는 말이 [천리] 간다 → 천리

문장 수준 임베딩

- BERT

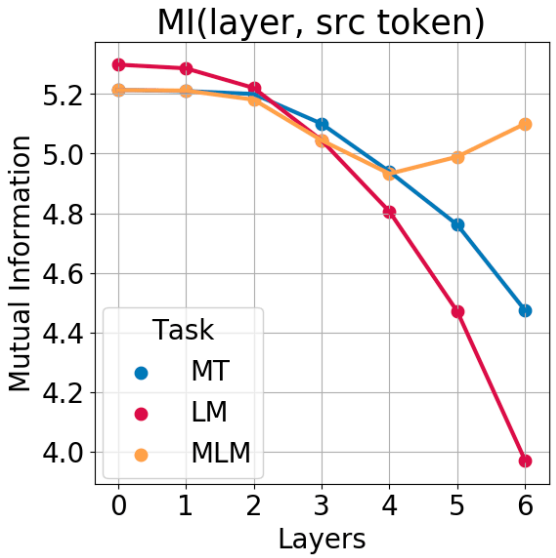
Masked Language Model

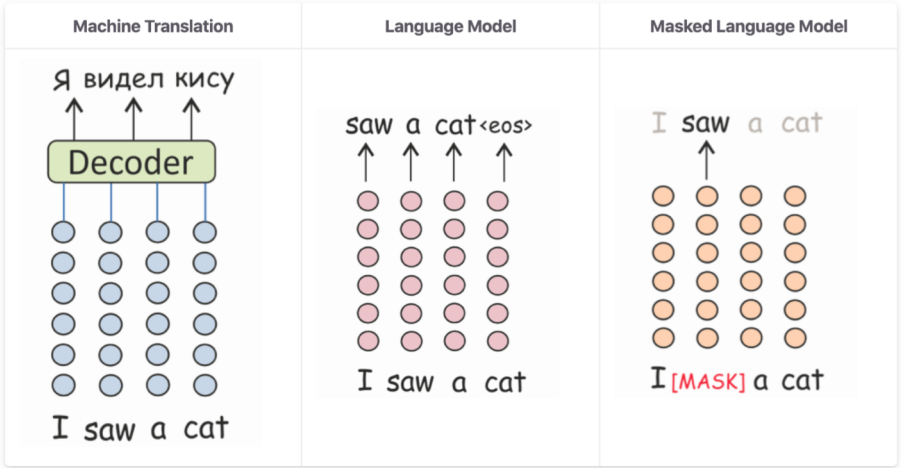
- 발 없는 말이 [MASK] 간다의 빈칸을 채워야 하기 때문에 문장 내 어느 자리에 어떤 단어를 쓰는 게 자연스러운지 앞뒤 문맥을 읽어낼 수 있게 된다.
- 발 없는 말이 천리 간다 발 없는 말이 컴퓨터 간다를 비교해 보면서 주어진 문장이 의미/문법상 비문인지 아닌지 가려낼 수 있다.
- 모델은 어떤 단어가 마스킹될지 전혀 모르기 때문에 문장 내 모든 단어 사이의 의미적, 문법적 관계를 세밀히 살피게 된다.

문장 수준 임베딩

- Transformer의 dynamics

언어모델과 기계번역 모델은 입력 정보를 갈수록 잊는다
즉, 새로운 정보(다음 또는 번역 토큰)를 생성하는 데 주목한다



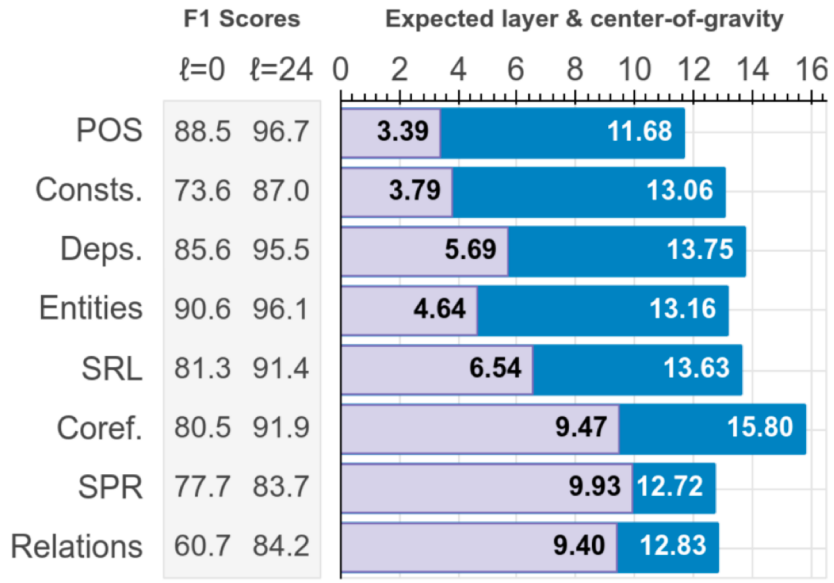


Voita, E., Sennrich, R., & Titov, I. (2019). The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives. *arXiv preprint arXiv:1909.01380*.

문장 수준 임베딩

- Transformer의 dynamics

트랜스포머의 하위 레이어에서는 syntactic 정보(POS 등)를,
상위 레이어에서는 semantic 정보(Coref. 등)를 함축한다

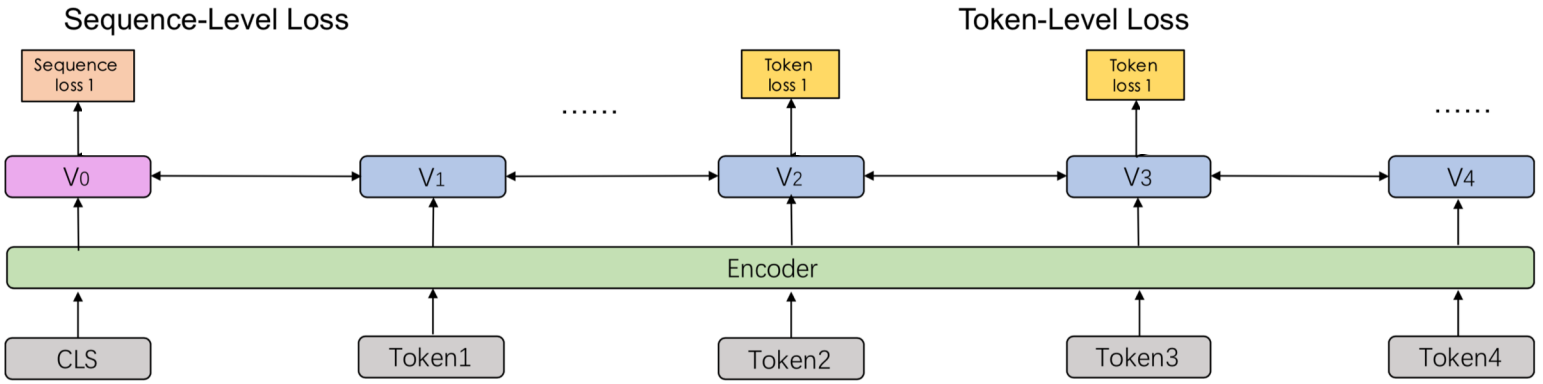


Tenney, I., Das, D., & Pavlick, E. (2019). BERT Rediscovers the Classical NLP Pipeline. arXiv preprint arXiv:1905.05950.

임베딩에 문법 정보 녹이기

- Continual Learning

  catastrophic forgetting 막기 위해 순차적으로 학습

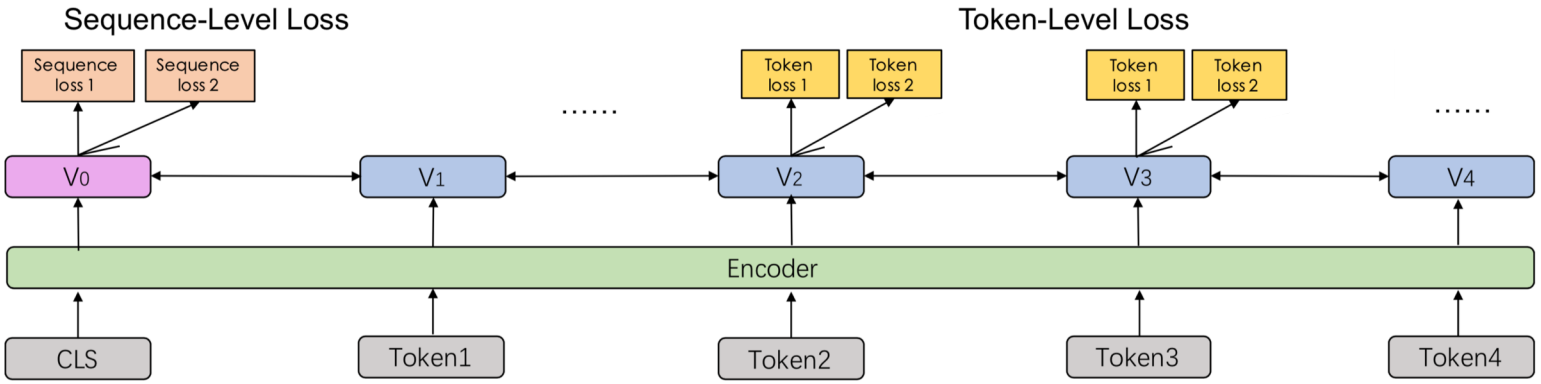


Sun, Y., Wang, S., Li, Y., Feng, S., Tian, H., Wu, H., & Wang, H. (2019). Ernie 2.0: A continual pre-training framework for language understanding. *arXiv preprint arXiv:1907.12412*.

임베딩에 문법 정보 녹이기

- Continual Learning

catastrophic forgetting 막기 위해 순차적으로 학습

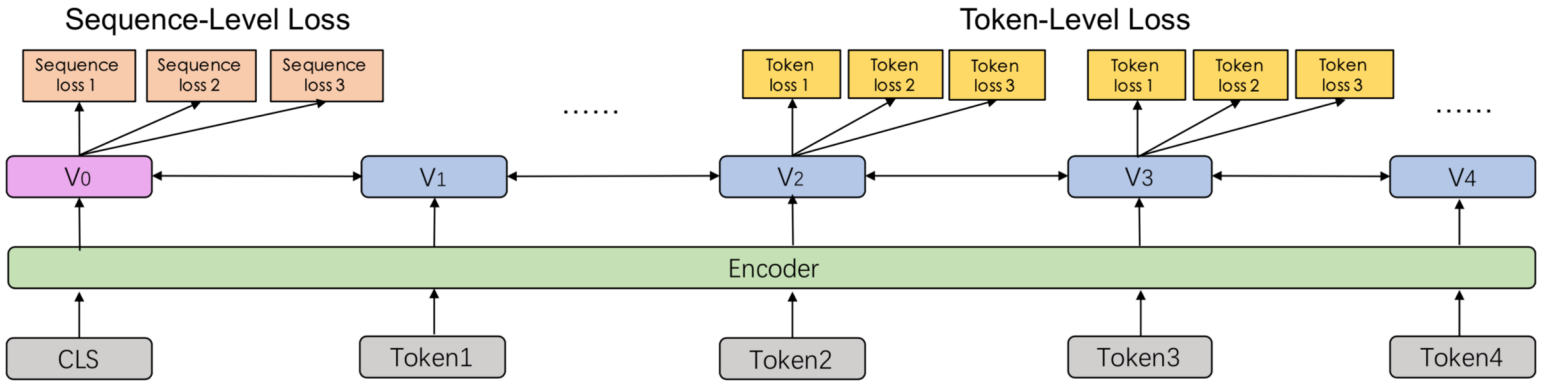


Sun, Y., Wang, S., Li, Y., Feng, S., Tian, H., Wu, H., & Wang, H. (2019). Ernie 2.0: A continual pre-training framework for language understanding. *arXiv preprint arXiv:1907.12412*.

임베딩에 문법 정보 녹이기

- Continual Learning

  catastrophic forgetting 막기 위해 순차적으로 학습



Sun, Y., Wang, S., Li, Y., Feng, S., Tian, H., Wu, H., & Wang, H. (2019). Ernie 2.0: A continual pre-training framework for language understanding. *arXiv preprint arXiv:1907.12412*.

임베딩에 문법 정보 녹이기

- Word-aware Pre-training task

Knowledge Masking Task
개체(named entity), 구(phrase) 단위 마스킹 (ERNIE1.0)
continual learning의 최초 모델에 적용

Capitalization Prediction Task
해당 단어가 대문자인지 소문자인지 예측

Token-Document Relation Prediction task
해당 토큰이 원본 문서의 다른 segment에 나타나는지 예측
키워드(동일 문서에 자주 나타나는 단어) 파악 가능

Sun, Y., Wang, S., Li, Y., Feng, S., Tian, H., Wu, H., & Wang, H. (2019). Ernie 2.0: A continual pre-training framework for language understanding. *arXiv preprint arXiv:1907.12412*.

임베딩에 문법 정보 녹이기

- Structure-aware Pre-training task

Sentence Reordering Task
원래 세그먼트를 1~m개로 나누고 랜덤 셔플한 뒤 순서 맞추기
문장 간 관계를 파악하게 됨

Sentence Distance task
3범주 분류 문제
0=동일 문서 앞뒤 문장
1=앞뒤는 아니지만 동일 문서
2=다른 문서에서 뽑힌 두 문장

Sun, Y., Wang, S., Li, Y., Feng, S., Tian, H., Wu, H., & Wang, H. (2019). Ernie 2.0: A continual pre-training framework for language understanding. *arXiv preprint arXiv:1907.12412*.

임베딩에 문법 정보 녹이기

- Semantic-aware Pre-training task

Discource Relation Task
두 문장이 의미적으로, 비유적으로 유사한지 여부 예측

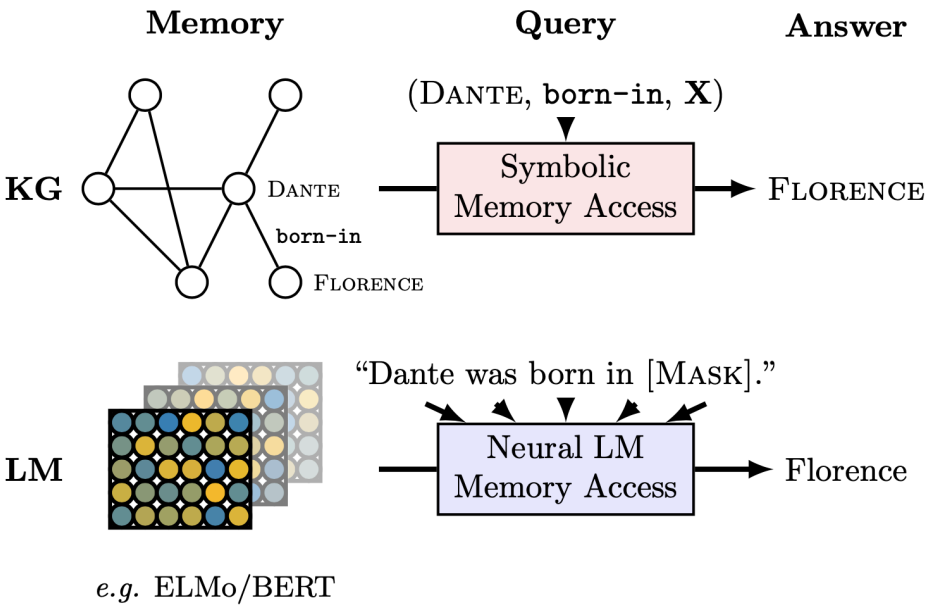
Sentence Distance task
앞 문장을 쿼리/뒤 문장을 검색 문서 제목으로 간주, 3범주 분류 문제
0=강한 관련성, 쿼리를 실제 검색했을 때 해당 문서 제목을 클릭
1=약한 관련성, 해당 쿼리를 던지면 검색은 되지만 미클릭
2=관련성 제로

Sun, Y., Wang, S., Li, Y., Feng, S., Tian, H., Wu, H., & Wang, H. (2019). Ernie 2.0: A continual pre-training framework for language understanding. *arXiv preprint arXiv:1907.12412*.

임베딩에 내재한 정보

- 사고력/상식의 원천은 곧 암기이다?

Knowledge Graph vs Language Model



Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., & Riedel, S. (2019). Language Models as Knowledge Bases?. *arXiv preprint arXiv:1909.01066*.

임베딩에 내재한 정보

- 사고력/상식의 원천은 곧 암기이다?

어휘 문법 정보뿐 아니라 지식까지 통으로 외운다

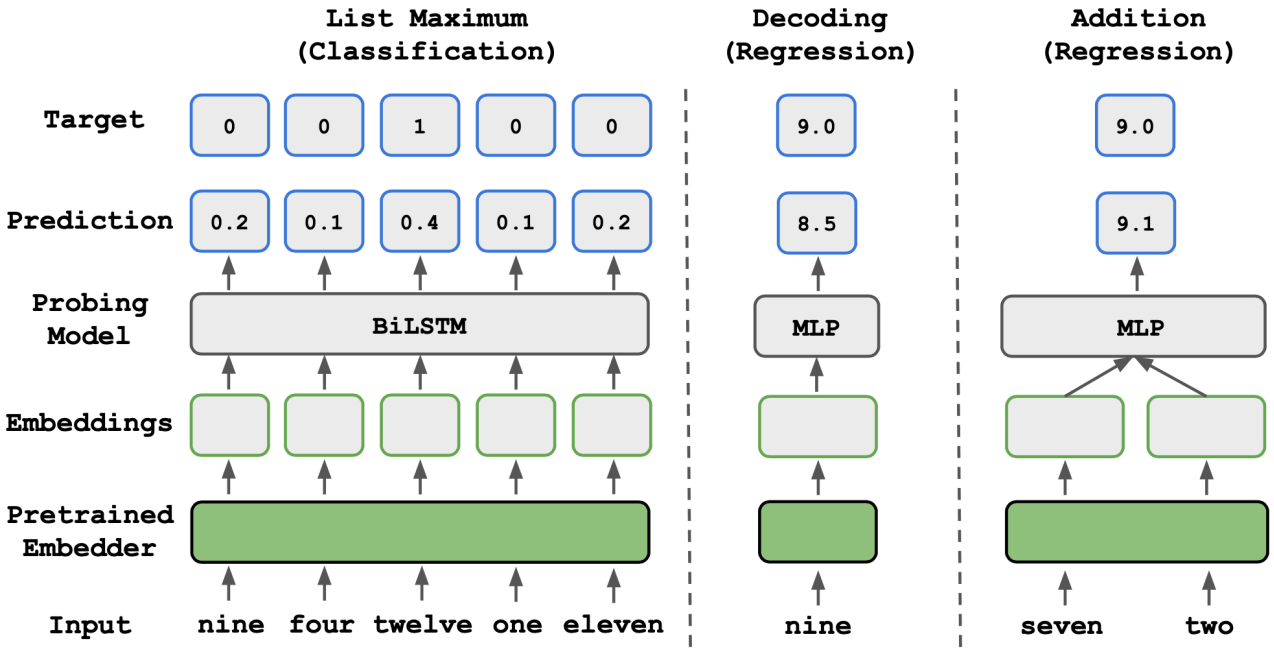
	Relation	Query	Answer	Generation
T-Rex	P19	Francesco Bartolomeo Conti was born in ____.	Florence	Rome [-1.8], **Florence [-1.8]**, Naples [-1.9], Milan [-2.4], Bologna [-2.5]
	P20	Adolphe Adam died in ____.	Paris	**Paris [-0.5]**, London [-3.5], Vienna [-3.6], Berlin [-3.8], Brussels [-4.0]
	P279	English bulldog is a subclass of ____.	dog	dogs [-0.3], breeds [-2.2], **dog [-2.4]**, cattle [-4.3], sheep [-4.5]
	P37	The official language of Mauritius is ____.	English	**English [-0.6]**, French [-0.9], Arabic [-6.2], Tamil [-6.7], Malayalam [-7.0]
	P413	Patrick Oboya plays in ____ position.	midfielder	centre [-2.0], center [-2.2], **midfielder [-2.4]**, forward [-2.4], midfield [-2.7]
	P138	Hamburg Airport is named after ____.	Hamburg	Hess [-7.0], Hermann [-7.1], Schmidt [-7.1], **Hamburg [-7.5]**, Ludwig [-7.5]
	P364	The original language of Mon oncle Benjamin is ____.	French	**French [-0.2]**, Breton [-3.3], English [-3.8], Dutch [-4.2], German [-4.9]
	P54	Dani Alves plays with ____ .	Barcelona	Santos [-2.4], Porto [-2.5], Sporting [-3.1], Brazil [-3.3], Portugal [-3.7]
	P106	Paul Toungui is a ____ by profession .	politician	lawyer [-1.1], journalist [-2.4], teacher [-2.7], doctor [-3.0], physician [-3.7]
	P527	Sodium sulfide consists of ____.	sodium	water [-1.2], sulfur [-1.7], **sodium [-2.5]**, zinc [-2.8], salt [-2.9]
	P102	Gordon Scholes is a member of the ____ political party.	Labor	Labour [-1.3], Conservative [-1.6], Green [-2.4], Liberal [-2.9], **Labor [-2.9]**
	P530	Kenya maintains diplomatic relations with ____.	Uganda	India [-3.0], **Uganda [-3.2]**, Tanzania [-3.5], China [-3.6], Pakistan [-3.6]
	P176	iPod Touch is produced by ____.	Apple	**Apple [-1.6]**, Nokia [-1.7], Sony [-2.0], Samsung [-2.6], Intel [-3.1]
	P30	Bailey Peninsula is located in ____.	Antarctica	**Antarctica [-1.4]**, Bermuda [-2.2], Newfoundland [-2.5], Alaska [-2.7], Canada [-3.1]
	P178	JDK is developed by ____.	Oracle	IBM [-2.0], Intel [-2.3], Microsoft [-2.5], HP [-3.4], Nokia [-3.5]
	P1412	Carl III used to communicate in ____.	Swedish	German [-1.6], Latin [-1.9], French [-2.4], English [-3.0], Spanish [-3.0]
	P17	Sunshine Coast, British Columbia is located in ____.	Canada	**Canada [-1.2]**, Alberta [-2.8], Yukon [-2.9], Labrador [-3.4], Victoria [-3.4]
	P39	Pope Clement VII has the position of ____ .	pope	cardinal [-2.4], Pope [-2.5], **pope [-2.6]**, President [-3.1], Chancellor [-3.2]
	P264	Joe Cocker is represented by music label ____.	Capitol	EMI [-2.6], BMG [-2.6], Universal [-2.8], **Capitol [-3.2]**, Columbia [-3.3]
	P276	London Jazz Festival is located in ____.	London	**London [-0.3]**, Greenwich [-3.2], Chelsea [-4.0], Camden [-4.6], Stratford [-4.8]
	P127	Border TV is owned by ____.	ITV	Sky [-3.1], **ITV [-3.3]**, Global [-3.4], Frontier [-4.1], Disney [-4.3]
	P103	The native language of Mammootty is ____.	Malayalam	**Malayalam [-0.2]**, Tamil [-2.1], Telugu [-4.8], English [-5.2], Hindi [-5.6]
	P495	The Sharon Cuneta Show was created in ____.	Philippines	Manila [-3.2], **Philippines [-3.6]**, February [-3.7], December [-3.8], Argentina [-4.0]
ConceptNet	AtLocation	You are likely to find a overflow in a ____.	drain	sewer [-3.1], canal [-3.2], toilet [-3.3], stream [-3.6], **drain [-3.6]**
	CapableOf	Ravens can ____.	fly	**fly [-1.5]**, fight [-1.8], kill [-2.2], die [-3.2], hunt [-3.4]
	CausesDesire	Joke would make you want to ____.	laugh	cry [-1.7], die [-1.7], **laugh [-2.0]**, vomit [-2.6], scream [-2.6]
	Causes	Sometimes virus causes ____.	infection	disease [-1.2], cancer [-2.0], **infection [-2.6]**, plague [-3.3], fever [-3.4]
	HasA	Birds have ____.	feathers	wings [-1.8], nests [-3.1], **feathers [-3.2]**, died [-3.7], eggs [-3.9]
	HasPrerequisite	Typing requires ____.	speed	patience [-3.5], precision [-3.6], registration [-3.8], accuracy [-4.0], **speed [-4.1]**
	HasProperty	Time is ____.	finite	short [-1.7], passing [-1.8], precious [-2.9], irrelevant [-3.2], gone [-4.0]
	MotivatedByGoal	You would celebrate because you are ____.	alive	happy [-2.4], human [-3.3], **alive [-3.3]**, young [-3.6], free [-3.9]
	ReceivesAction	Skills can be ____.	taught	acquired [-2.5], useful [-2.5], learned [-2.8], combined [-3.9], varied [-3.9]
	UsedFor	A pond is for ____.	fish	swimming [-1.3], fishing [-1.4], bathing [-2.0], **fish [-2.8]**, recreation [-3.1]

Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., & Riedel, S. (2019). Language Models as Knowledge Bases?. *arXiv preprint arXiv:1909.01066*.

임베딩에 내재한 정보

- 임베딩에 숫자의 의미가 얼마나 포함돼 있을까?

Numeracy Probing Model



Wallace, E., Wang, Y., Li, S., Singh, S., & Gardner, M. (2019). Do NLP Models Know Numbers? Probing Numeracy in Embeddings. *arXiv preprint arXiv:1909.07940*.

임베딩에 내재한 정보

- 임베딩에 숫자의 의미가 얼마나 포함돼 있을까?

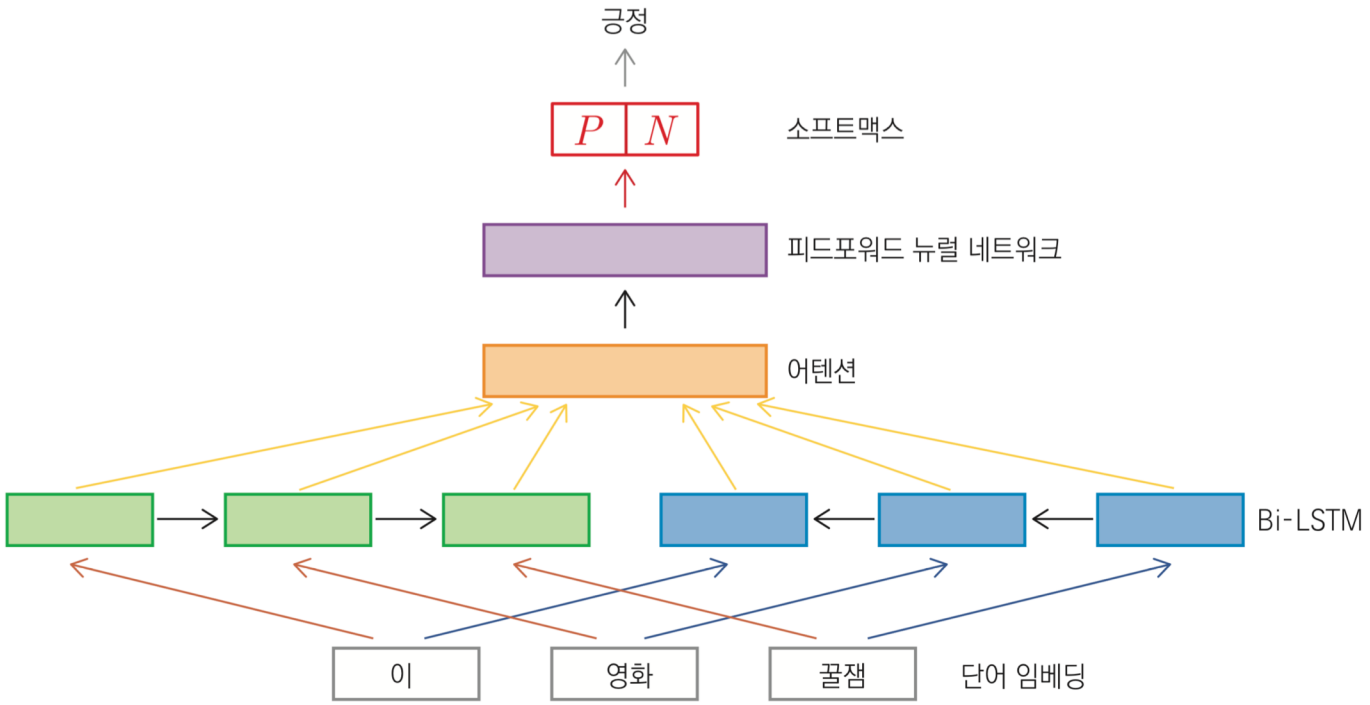
임베딩에 숫자 의미 내포 (random vs pretrained)
Transformer보다 CNN 계열이 잘한다 (지역 정보 포착 유리)
Subword보다 Char 단위가 잘한다 (BERT vs ELMo/Char-*)

Interpolation	**List Maximum (5-classes)**			**Decoding (RMSE)**			**Addition (RMSE)**		
Integer Range	[0,99]	[0,999]	[0,9999]	[0,99]	[0,999]	[0,9999]	[0,99]	[0,999]	[0,9999]
Random Vectors	0.16	0.23	0.21	29.86	292.88	2882.62	42.03	410.33	4389.39
Untrained CNN	0.97	0.87	0.84	2.64	9.67	44.40	1.41	14.43	69.14
Untrained LSTM	0.70	0.66	0.55	7.61	46.5	210.34	5.11	45.69	510.19
Value Embedding	**0.99**	0.88	0.68	**1.20**	11.23	275.50	**0.30**	15.98	654.33
Pre-trained									
Word2Vec	0.90	0.78	0.71	2.34	18.77	333.47	0.75	21.23	210.07
GloVe	0.90	0.78	0.72	2.23	13.77	174.21	0.80	16.51	180.31
ELMo	0.98	0.88	0.76	2.35	13.48	62.20	0.94	15.50	45.71
BERT	0.95	0.62	0.52	3.21	29.00	431.78	4.56	67.81	454.78
Learned									
Char-CNN	0.97	**0.93**	**0.88**	2.50	**4.92**	**11.57**	1.19	**7.75**	**15.09**
Char-LSTM	0.98	0.92	0.76	2.55	8.65	18.33	1.21	15.11	25.37
DROP-trained									
NAQANet	0.91	0.81	0.72	2.99	14.19	62.17	1.11	11.33	90.01
- GloVe	0.88	0.90	0.82	2.87	5.34	35.39	1.45	9.91	60.70

Wallace, E., Wang, Y., Li, S., Singh, S., & Gardner, M. (2019). Do NLP Models Know Numbers? Probing Numeracy in Embeddings. *arXiv preprint arXiv:1909.07940*.

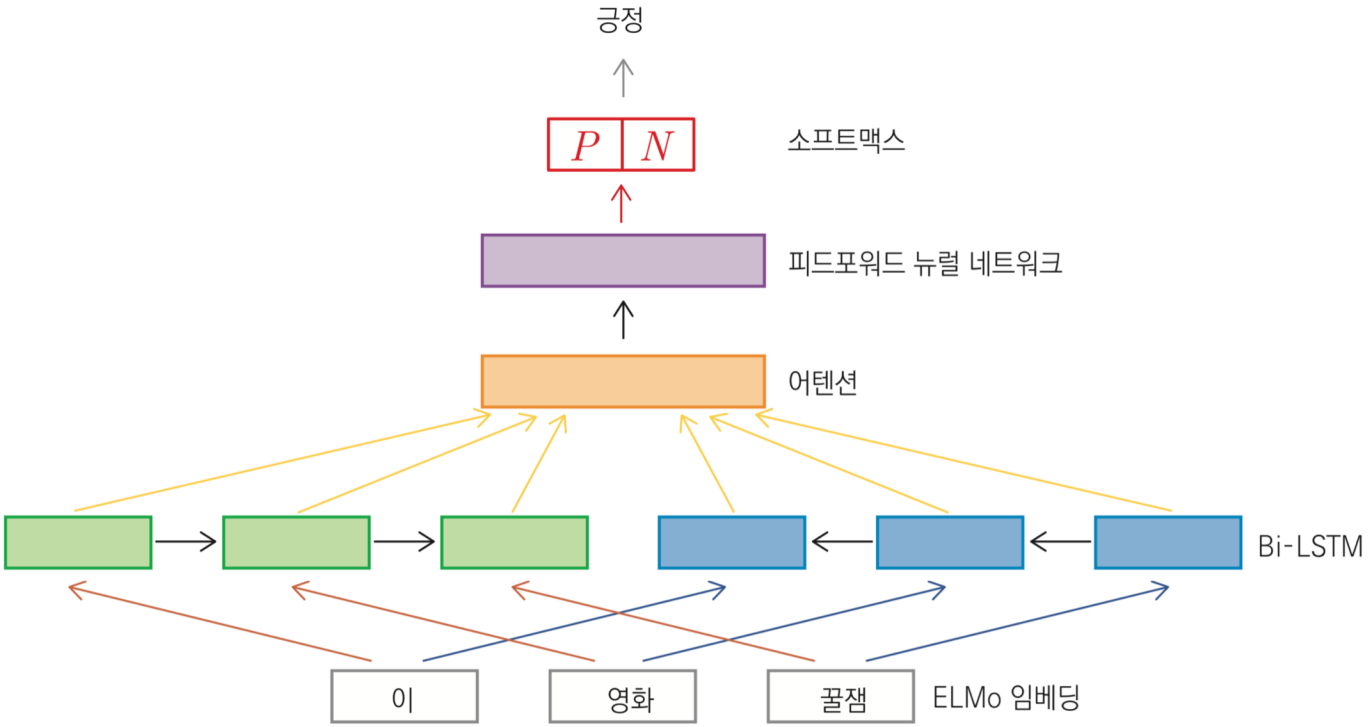
임베딩 파인튜닝

- Word Embeddings



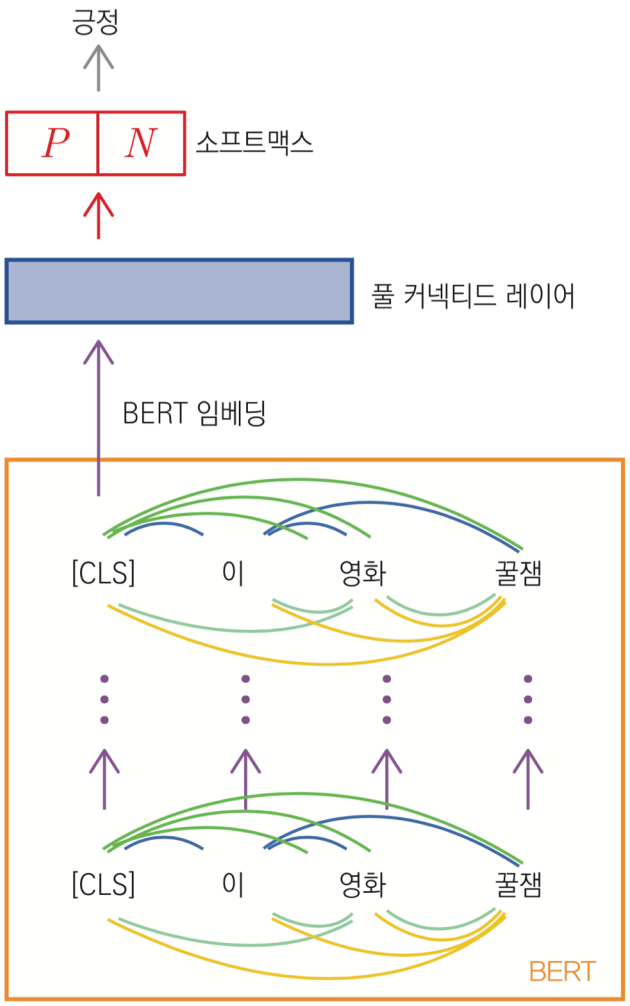
임베딩 파인튜닝

- ELMo Embeddings



임베딩 파인튜닝

- BERT Embeddings



튜토리얼

- Code

  https://github.com/ratsgo/embedding

- Blog

  https://ratsgo.github.io/embedding

- 도커 환경 구성

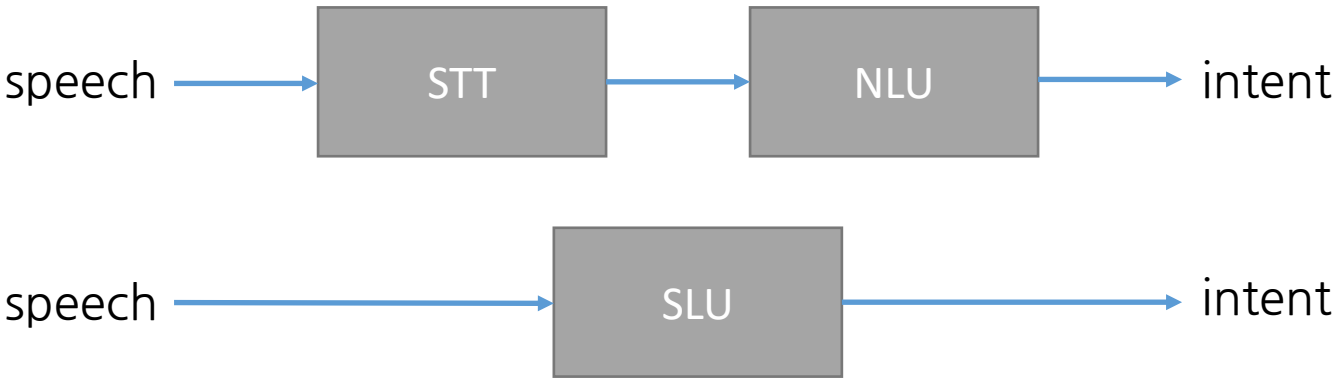
```
git clone https://github.com/ratsgo/embedding.git
cd embedding
docker build -t ratsgo/embedding-gpu -f docker/Dockerfile-GPU .
docker run -it --rm --runtime=nvidia ratsgo/embedding-gpu bash
```

튜토리얼


자연어 처리 모델의 성능을 높이는 핵심 비결
Word2Vec에서 ELMo, BERT까지
한국어 임베딩
Sentence Embeddings Using Korean Corpora
x_2
x_1
이기창 지음
NAVER Chatbot Model 감수

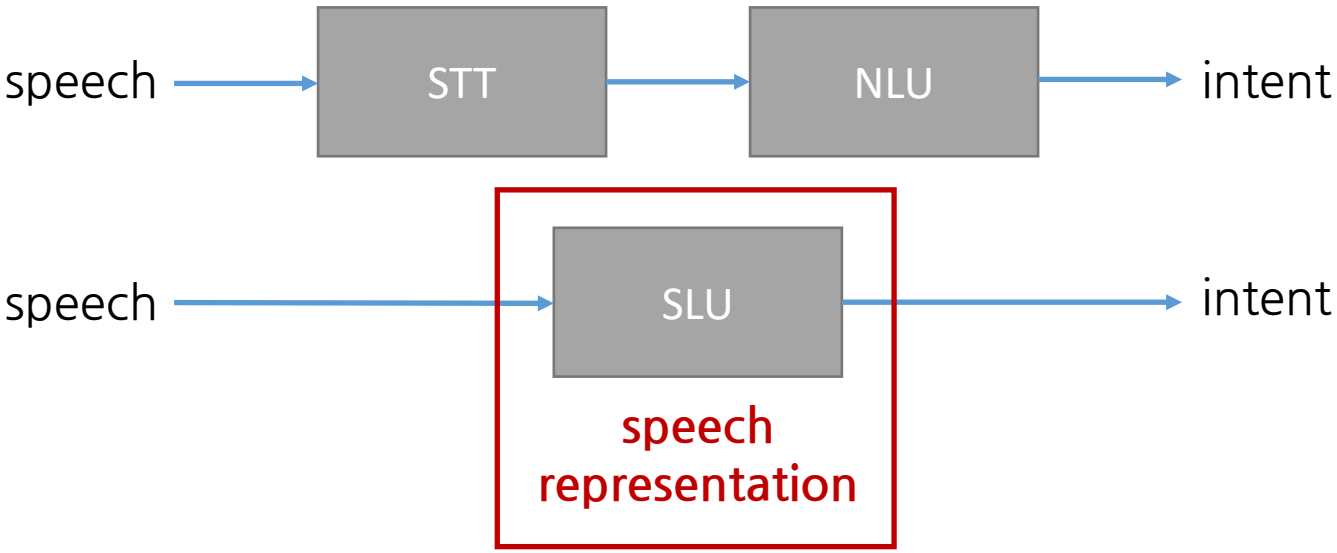
Beyond Text

- Spoken Language Understanding



Beyond Text

- Spoken Language Understanding



Beyond Text

- Universal Representation?