

# KEY POINTS

NAME/DATE/SUBJECT

기계번역 기술 개요 및 동향

2021. 7. 19 (월) / 서예애

## NOTES

### 1. 기계번역 방법

- 1.1 규칙기반 기계번역 (RBMT, Rule-Based)
- 1.2 통계기반 기계번역 (SMT, Statistical)
- 1.3 신경망 기반 기계번역 (NMT)

### 2. 주의집중 기반 인코더-디코더 구조

- 2.1 인코더-디코더 구조
- 2.2 주의집중 기법

### 3. 신경망 기반 기계번역 모델 구조

- 3.1 순환신경망 기반 NMT
- 3.2 합성곱신경망 기반 NMT
- 3.3 자기집중 기반 NMT

### 4. 최신 연구주제

- 4.1 양국어 코퍼스 부족
- 4.2 문맥 정보의 확장

### 5. 결론

## SUMMARY

# KEY POINTS

NAME/DATE/SUBJECT

## NOTES

### 1. 기계번역 방법

#### 1.1 규칙기반 기계번역

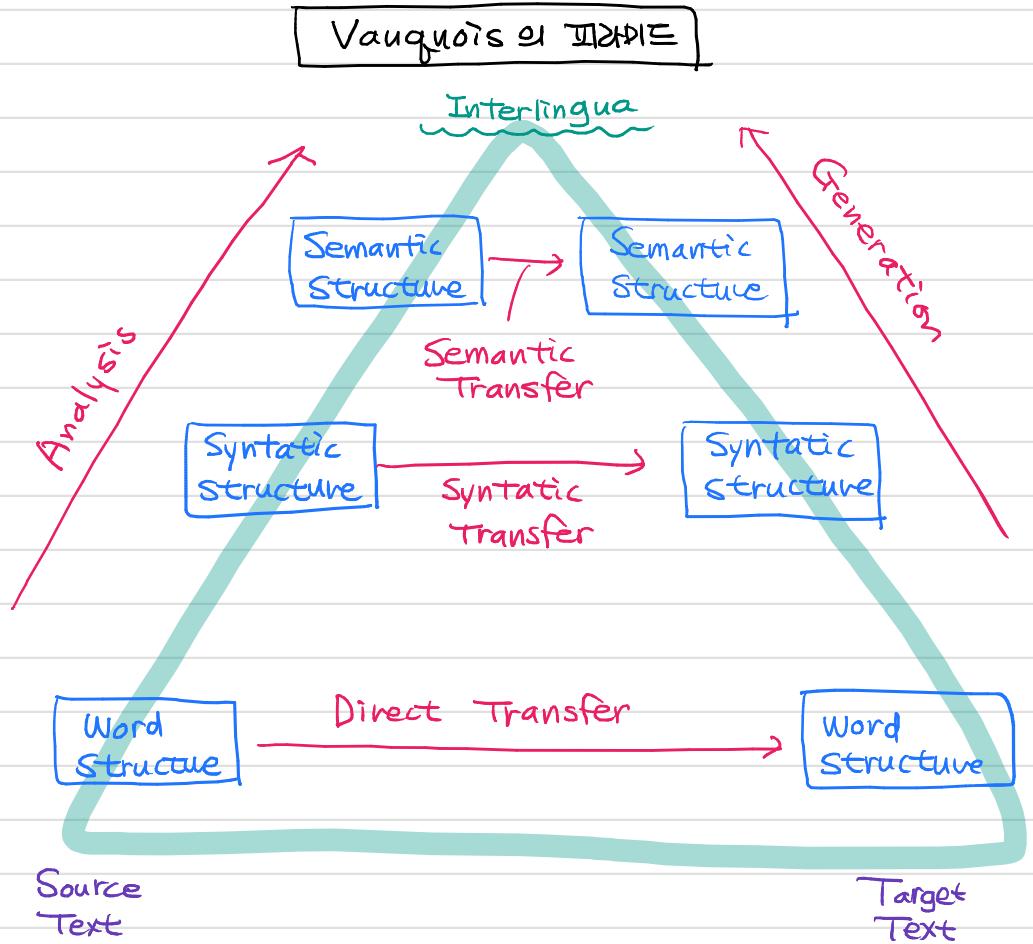
언어 : 문법 규칙의 집합으로 기준  
언어 규칙에 따라 번역하고자 할 연구

언어학적 고찰



## SUMMARY

## NOTES



- 언어적 유사성이 큰 언어쌍 : 형태소 분석 후 변환하는 직접 번역 방식
- 중간언어 (interlingua) 방식 : 다국어 번역에 적합
  - 문장의 의미를 언어독립적으로 나타낼 수 있는 표준언어를 정의
  - 분석과 새롭고 번역을 수행하고자 한 방식

## KEY POINTS



- GPSG
- HPSG
- LFG
- DG

## NAME/DATE/SUBJECT

## NOTES

### · 대부분의 언어학

: 구조분석에 기반한 간접 변환 방식으로 개발



언어의 체태소 및 구문구조를 문법적으로  
나타낼 수 있어야 함



- 문법에  
기반한  
연구방법들!
- GPSG (Generalized Phrase Structure Grammar)  
· HPSG (Head-driven Phrase Structure Grammar)  
· LFG (Lexical Functional Grammar)  
· DG (Dependency Grammar)

- 모든 언어구조를 규칙화하는 데 성공
- 말투에서 추출한 확률정보에 기반, 번역하고자 하는 시도.

## SUMMARY

- Noisy Channel Model
- 베이지안 규칙

## NOTES

## 1.2 통계기반 기계번역

- 원시 언어문장  $f_i^j = f_1, \dots, f_j$

↑  
대응

- 목적 언어문장  $e_i^x = e_1, \dots, e_x$



노이지 채널 모델에 따라

확률  $P(e_i^x | f_i^j)$  을 최대화시키는  
문장  $e_i^x$  를 최종 번역 문장으로 결정.

$$\tilde{e} = \arg \max_e P(e_i^x | f_i^j)$$

번역모델

↓  
베이지안 규칙에 의해  
나눌 수 있음

$$= \arg \max_e P(f_i^j | e_i^x) P(e_i^x)$$

언어모델

구 단위 SMT (Phrase-based SMT, PBSMT)

단어 번역 시, 주변 단어들을 문맥정보를 반영하기 위해  
단어가 아닌 구(phrase)를 번역의 기본단위로 하여  
 $P(f_i^j | e_i^x)$  를 모델화한 SMT

Continued

## NOTES

## [ PBSMT ]

"구": RBMT에서의 언어학적 의미가 아니라,  
표준적으로 말하는 연속된 단어열



주어진 워시언어 문장을 연속된 구들의 조각으로 나눈 후,  
각각을 구단위 번역패턴을 이용하여 번역하고  
이를 최종 결합



1 : 1

구단위 번역 패턴을

워시언어 문장 - 목적언어 문장

1:1로 매핑된 양국어 코퍼스로부터 추출

워시언어 구에 대응하는 목적언어 구가  
확률을 기반으로 계산된 개수와 함께 저장

- PBSMT : 연속 단어열을 번역의 단위로 사용하므로  
한국어-영어처럼 어순이 크게 다를 언어 같은  
번역에 어려움이 있음.



## SUMMARY

## KEY POINTS

- 하위문장을 기반한 튜닝?  
(SMT의 균형적 문제점?)

## NAME/DATE/SUBJECT

### NOTES



**HSMT** (Hierarchical Phrase - Based Statistical Machine Translation)

: 하위 구 (sub-phrase)를 가질 수 있는 구 (hierarchical phrase)를 번역의 기본단위로 함.



↳ SOTA PBSMT의 성능을 뛰어넘음.

### 1.3 신경망 기반 기계번역

(Neural Machine Translation, NMT)

- 초기 : SMT의 기본구조를 그대로 유지하면서 언어별 등에 하위문장을 신경망으로 대체하는 형태  
⇒ SMT의 균형적 문제인 하위 문장을 기반한 튜닝의 한계를 극복할 수 없었음.

2014

. Sequence-to Sequence learning with neural networks

2014 . On the properties of neural machine translation:

2015 Encoder - Decoder approaches

↓ . Neural Machine Translation by Jointly learning to align and translate

NMT를 하나의 신경망으로 모델링하여 학습

SOTA SMT의 성능을 뛰어넘는 NMT 시대로 ... ☺

## SUMMARY

## NOTES

## 2. 어텐션 기반의 인코더 - 디코더 구조

- 현재 대부분의 NMT 모델 : 어텐션을 도입해

종말단 (end-to-end) 방식의 인코더 - 디코더 구조.

- Sutskever, "Sequence to sequence learning with neural networks" (2014)
- Cho K, "On the properties of neural machine translation: Encoder - Decoder approaches" (2015)
- Bahdanau, "Neural machine translation by jointly learning to align and translate" (2015)

end-to-end encoder-decoder 모델 구조 확장에  
기여한 대표적 초기 연구들

## 2.1 인코더 - 디코더 구조

- 인코더가 입력문장 단어들을 읽어 고정된 길이의 벡터를 변환

↓ 전달

- 디코더 : 전달받은 벡터를 기반으로 출력문장 단어 생성

## SUMMARY

## KEY POINTS

· 공지학습: 10P ↗

## NAME/DATE/SUBJECT

### NOTES

#### 학습자 관점

- 원시 언어문장 :  $X = X_1, \dots, X_I$

- 목적 언어문장 :  $Y = Y_1, \dots, Y_J$

$$P(Y|X, \theta) = \prod_{j=1}^J P(Y_j | X, Y_{<j}, \theta)$$

신경망 파라미터

$Y_{<j} : Y_1, \dots, Y_{j-1}$  위치  $j$  전까지의 부분 번역문

$$\theta = \operatorname{argmax}_{\theta^*} \sum_{m=1}^M \log P(Y^m | X^m, \theta^*)$$

학습데이터를 주어진  $M$  개의 양극어 문장을 대상으로

최대우도 추정기법 (maximum log-likelihood estimation)  
으로 최적화.

## SUMMARY

## NOTES

Π : 곱집합 (=데카르트 곱)

- 각 집합의 원소를 각 서분으로 하는 튜플들의 집합

$$A \times B = \{(a, b) | a \in A, b \in B\}$$

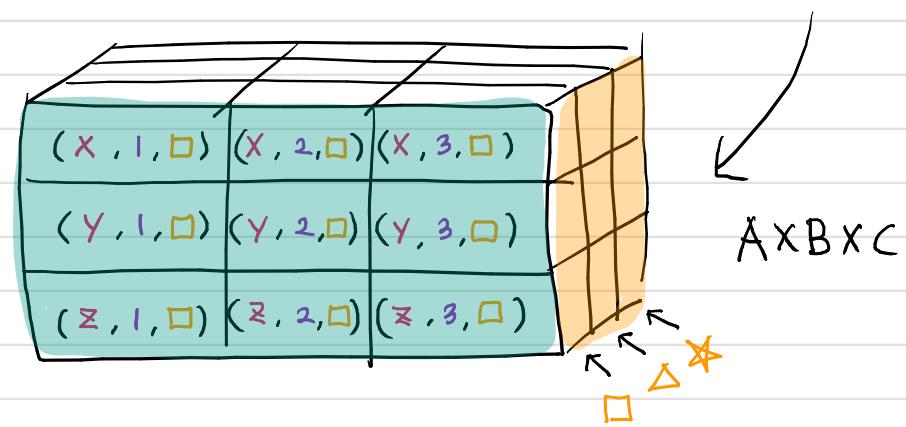
$$A = \{x, y, z\}$$

$$B = \{1, 2, 3\}$$

	1	2	3
x	(x, 1)	(x, 2)	(x, 3)
y	(y, 1)	(y, 2)	(y, 3)
z	(z, 1)	(z, 2)	(z, 3)

$C = \{\square, \Delta, \star\}?$

우선 A, B 곱집합을 만들 후,  
마지막에 C의 원소를 추가



## SUMMARY

## NOTES

$\text{BiRNN}$  :

저방향 RNN  $\vec{f}$  : 이례문장을 저방향 ( $x_1, \dots, x_T$ )으로  
 읽어들여 은닉상태 ( $\vec{h}_1, \dots, \vec{h}_T$ )로 변화  
 +
 역방향 RNN  $\overset{\leftarrow}{f}$  : 역방향 ( $x_T, \dots, x_1$ )의  
 은닉상태 ( $\overset{\leftarrow}{h}_1, \dots, \overset{\leftarrow}{h}_T$ )로 변화

-  $x_i$  의 최종 은닉상태  $h_i$  :  $\overset{\rightarrow}{h}_i + \overset{\leftarrow}{h}_i$   
 $[\overset{\rightarrow}{h}_i^\top; \overset{\leftarrow}{h}_i^\top]^\top$

- 역방향을 고려할 수 없는  
 실시간 번역 (Simultaneous Machine Translation)  
 같은 경우를 제외하고 대부분의 RNN-인코더는  
BiRNN으로 구현

## NOTES

## 2.2 어텐션 기법



RNN

RNN기반 인코더-디코더 구조에서,

입력문장  $\underline{x} = \underline{x_1, \dots, x_I}$

인코더 은닉상태 정보를 기반으로 문맥벡터 생성.

시간  $i$  에서의 인코더의 은닉상태

$$\underline{h_i} = f(x_i, h_{i-1})$$

비선형함수

$$c = g(f(h_1, \dots, h_I)) = \underline{h_I}$$

마지막  $I$  시간의 인코더의 은닉상태

## NOTES

- 디코더 :  $y_j$ 의 확률값을 예측하기 위해

$y_{j-1}$  (이전 출력단어)  
 $s_j$  (시간  $j$ 에서의 디코더의 은닉상태)  
 $c$  (문맥 베터)  
 알려문장  $X$  전체 정보를 함축하여 저장했다고  
 간주되는 베터  
 $y$ 의 모든 출력단어 생성 시 동일한 문맥ベ터 사용  
 를 이용하여 아래와 같은 확률로 출력문장  
 $y = y_1, \dots, y_j$  을 생성

$$P(y_j | y_1^{j-1}, x) = g(y_{j-1}, s_j, c)$$

- 문맥ベ터  $c$  가 긴 알려문장의 베터에 필요한 정보를 충분히 담을 수 없음  
 $\Rightarrow$  문장의 길이가 길어지면 베터의 성능이 가파르게 저하

$\downarrow$   
 이 문제를 해결하기 위해 어려운 방법 제안

- 저결점수 계산 향수

### ① Additive

Score ( $s_j, h_i$ )

$$= v^T \tanh(W[s_j; h_i])$$

### ② Dot - Product

Score ( $s_j, h_i$ )

$$= s_j^T h_i$$

### ③ Scaled Dot - Product

$$\text{Score} (s_j, h_i) = \frac{s_j^T h_i}{\sqrt{d}}$$

이것은 차원의 차원

어텐션

출격단어 생성 시에 입력단어의 어느 부분을 강조적으로 고려할지 단어 간 저결관계에 기반하여 알려줌!

$$C_j = \sum_{i=1}^n \alpha_{ji} h_i$$

문맥ベクトル

$\{\alpha_{j1}, \dots, \alpha_{jn}\}$   
( $y_j$ 의 결정에 각 입력단어들이 기여한 확률)

$$\alpha_{ji} = \frac{\exp(e_{ji})}{\sum_{k=1}^n \exp(e_{jk})}$$

↑  
softmax 과정을 거쳐 0~1 사이의 값을 가지게 된 저결 가중치

$$e_{ji} = A(s_{j-1}, h_i)$$

j 번째 입력단어  $x_i$ 와

j 번째 출격단어  $y_j$  간의 저결 연관도를 저수화

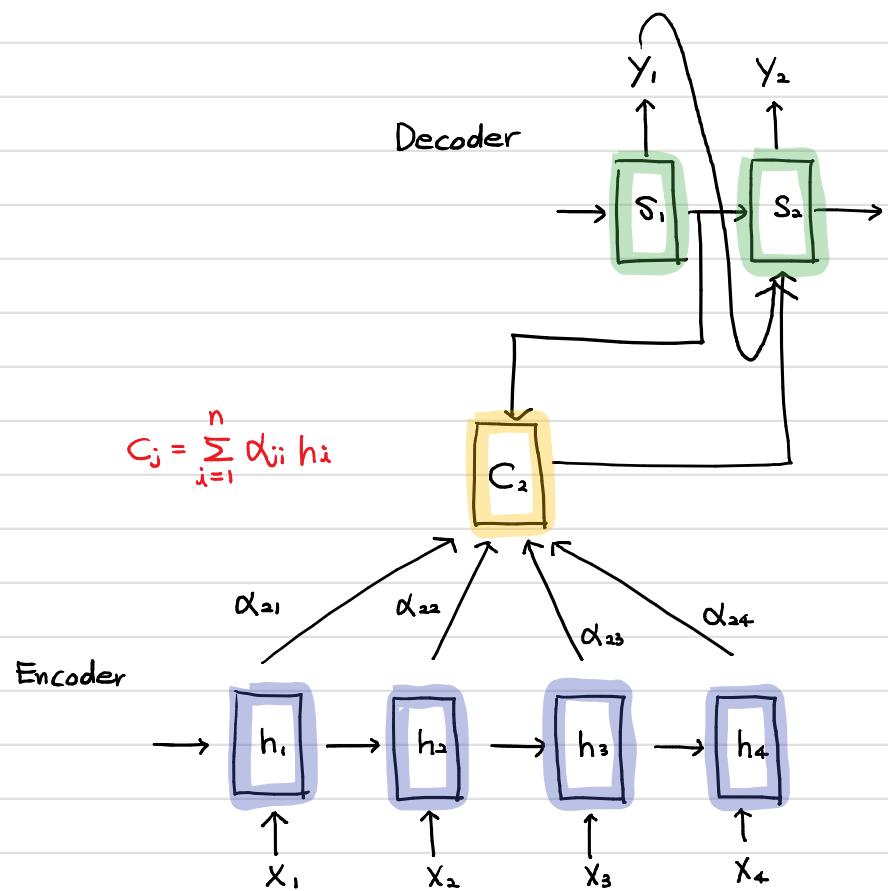
## NOTES

- 어텐션을 적용한 디코더

: 각 출력 단어별로 별도의 문맥벡터  $C_j$ 를 가지.

$$P(y_j | y_1^{j-1}, x) = g(y_{j-1}, s_j, \underline{C_j})$$

RNN 기반 어텐션을 적용된 인코더-디코더 구조



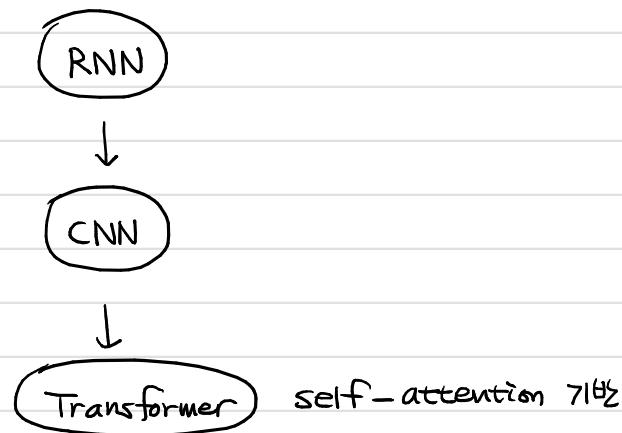
## SUMMARY

## NOTES

"Attention is all you need"

; 어텐션을 보다 적극적으로 적용한 self-attention  
기반의 NMT 제안.

### 3. 신경망 기반 기계번역 모델 구조



#### 3.1 순환신경망 기반 NMT

- 대표적인 RNN기반 NMT: Google GNMT, RNMT+



## NOTES

## GNMT

각각 8개의 LSTM을 쌓은 인코더, 디코더  
어려운 처리를 위한 1개의 순방향 신경망

- 기존 NMT 모델들을 대체의 RNN을 쌓을 때  
기울기 소실 / 폭주로 인해 학습이 잘 되지 않음



"잔차 연결" (Residual Connection) 을 도입하여 해결  
이전 층의 입력값을 다음 층의 입력값에 반영

- 단어 단위 베이스에서 발생하는 OOV (Out-of-Vocabulary)



하부단어 (sub-word) 기반 아버디드으로 해결

여러 기법의 적용으로 GNMT는 속도와 성능 면에서 큰 개선  
SMT를 NMT를 교체할 수 있게 됨?

## RNMT +

GNMT와 유사한 RNN NMT 구조에

Transformer에서 제안한 멀티헤드 어려운 기법 도입

WMT'14의 영어-독일어, 영어-프랑스어 번역 테스크에서  
Transformer를 근소한 차이로 앞섰다

## SUMMARY

## NOTES

## 3. 2 헉사급신경망 기반 NMT

## - CNN

① RNN에 비해 복잡화가 용이, 속도가 빠름.

② 커널 크기만큼의 단어를 한번에 문맥정보로 볼 수 있으므로

단어에 있는 단어들의 접근이 RNN의 순차적 접근보다 용이.

( · Conv S2S )      인코더 - 디코더 모두 CNN으로 구현한  
 ( · Slice Net )      대표적 CNN 기반 NMT

## Conv S2S

- 1차원 CNN에 GLU (Gated Linear Unit) 연결
- 잔차연결
- Multi-Hop Attention  
: 디코더의 각 층마다 어텐션 기법 적용

- 단어의 위치정보는 번역 시 중요한 정보

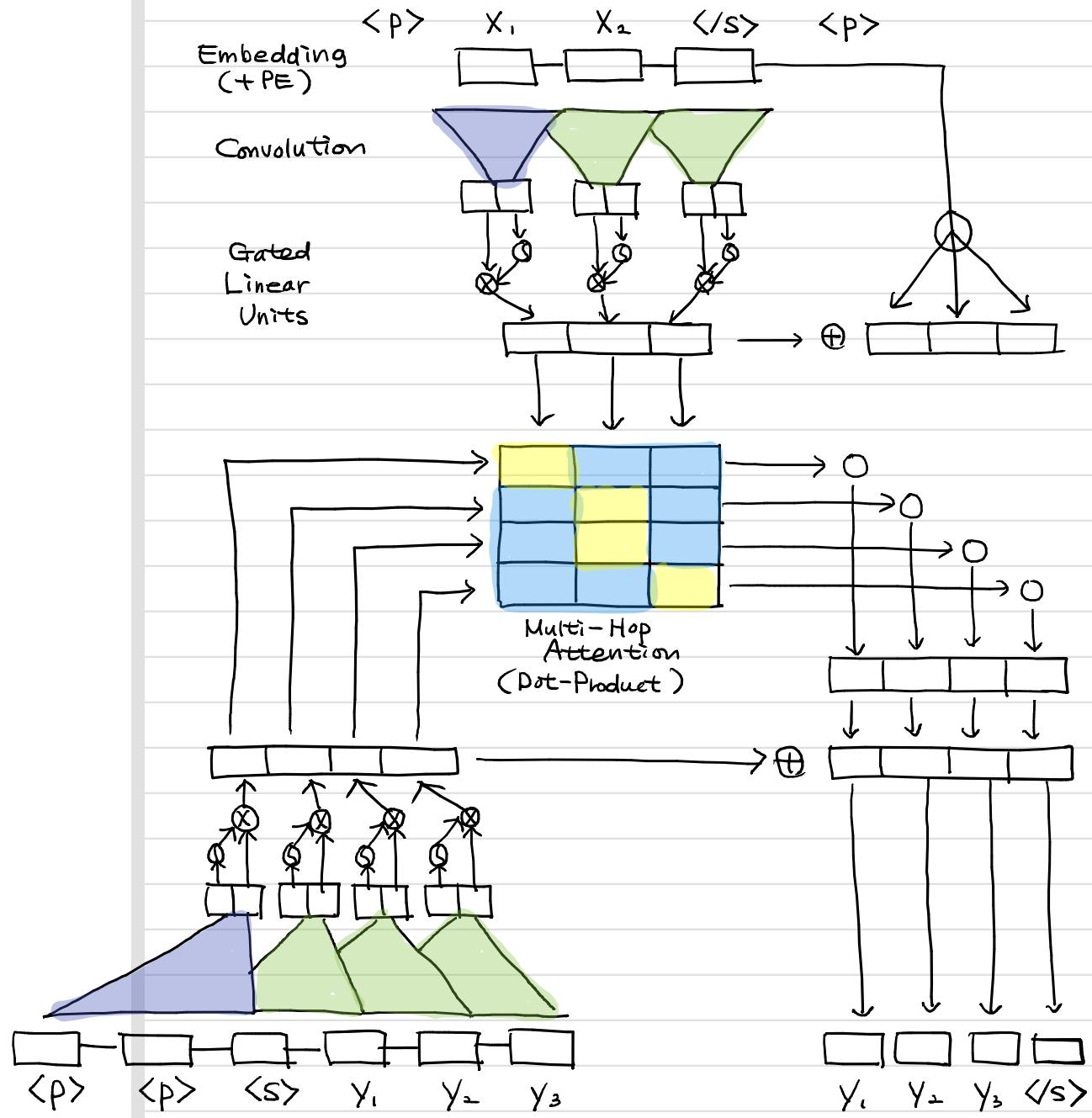
- CNN은 RNN과 달리 positional encoding을 통해  
위치정보를 알려줌

- ConvS2S : 단어 아베딩 정보에 절대위치 정보를 더하는 계산을  
위치 정보 인코딩



## NOTES

## (Encoder)



## SUMMARY

## KEY POINTS

NAME/DATE/SUBJECT

## NOTES

### 3.3 Self-Attention 기반 NMT

## Transformer

기존 RNN, CNN을 적용하지 않고  
어텐션에 전적으로 의존

기준에 적용되며 인코더 - 디코더 어텐션과 함께  
인코더 - 디코더에 self-Attention 적용

- 인코더 - 디코더 어텐션 : 앤솔러레이터와 총액터 간의 연관관계 고려
  - 인코더 어텐션 : 앤솔러레이터들끼리 연관관계
  - 디코더 어텐션 : 총액터들끼리 연관관계

• 멀티헤드 어텐션  
:  $Q, K, V$  를 대상으로  $h$  개의 서로 다른 선형 투영 (linear projection)  
을 수행함으로써  $h$  개의 다른 부분 망을 상에서의  
어텐션 차수를 계산할 수 있게 함.

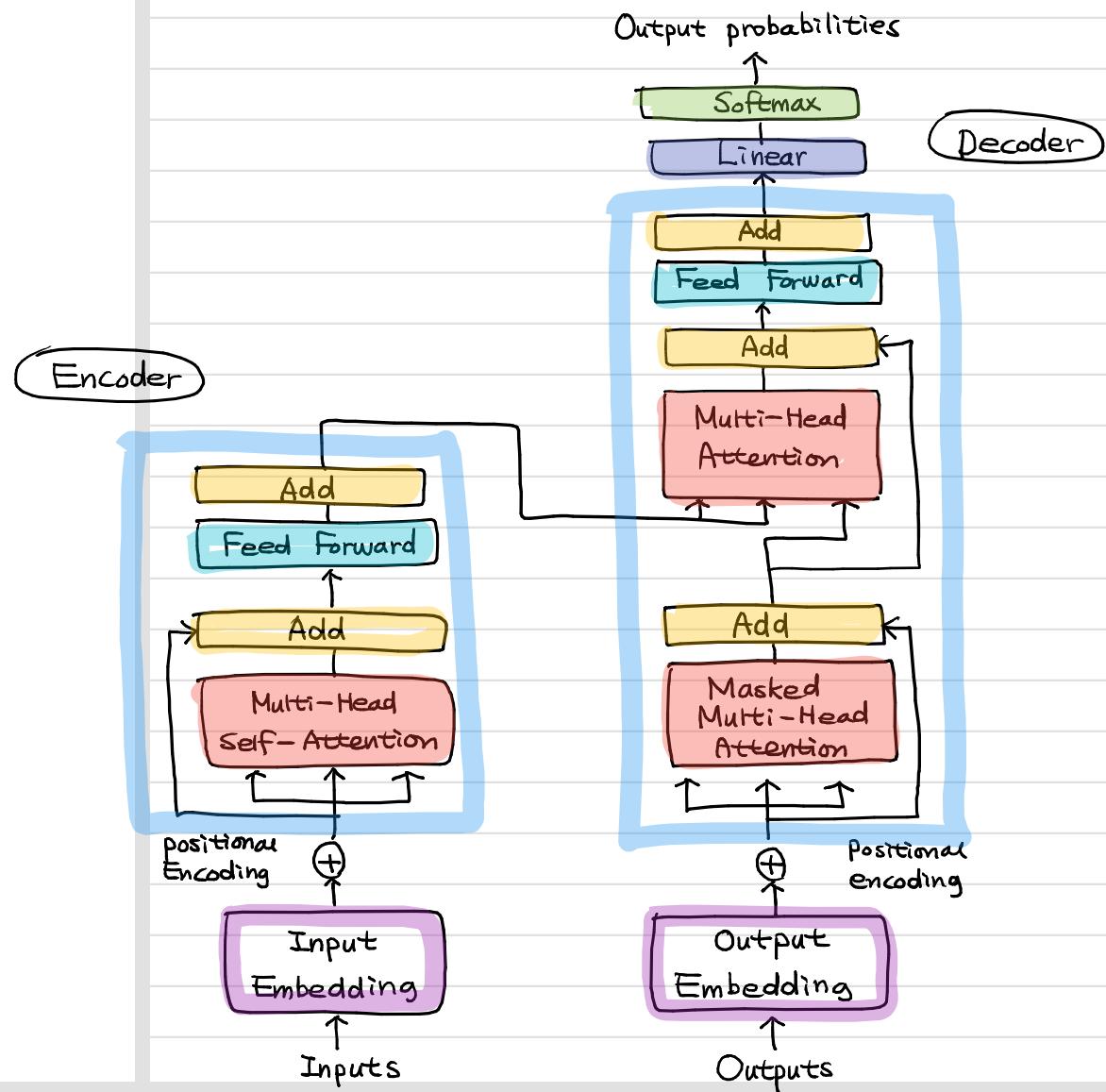
$$\text{Multi Head } (\mathbf{Q}, \mathbf{k}, \mathbf{v}) = \text{Concat}(\text{head}_1, \dots, \text{head}_n) \mathbf{W}^o$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

## SUMMARY

## NOTES

- 기존의 절대위치 기반 위치 인코딩은 강건성 등에서 문제가 있음.  
⇒ 사인함수와 코사인함수를 이용한 저주파 위치 인코딩  
( Sinusoidal Positional Encoding )을 수용



## SUMMARY

# KEY POINTS

· pivot translation



pivot language

: 3rd or intermediate  
language



bridge the gap  
between  
language pairs.

# NAME/DATE/SUBJECT

## NOTES

### 4. 출시 연구주제

아직 해결해야 할 문제점들이 많음!

#### 4.1 양자어コーパス 부족

- NMT는 대량의 양자어コーパ스를 필요로 함
- 실제 학습에 필요한 충분한 코퍼스가 주어지지 않을 경우  
RBMT나 SMT가 더 좋은 성능을 보임!

저자원 (Low-resource) 환경에서 NMT를 개발하는 방법

#### ① 상대적으로 자원이 풍부한 언어를 기반으로 전이학습

(Transfer learning)

: 자원이 풍부한 영-프 등 언어상을 먼저 학습 후  
영-베 같은 저자원 언어상을 이어서 학습

#### — ② 다국어번역, 피봇번역 등을 통해 번역지식 공유

(pivot translation)

: 멀티태스크 학습 (Multi-task learning)을 이용하여  
자원이 풍부한 언어에서 학습한 정보를 저자원 언어상이 공유

“Google’s multilingual neural machine translation  
system: Enabling zero-shot translation”

- 특정 언어상에 대한 양자어コーパ스가 없을 경우에도  
여러와 같이 풍부한 학습코퍼스를 가지고 있는 언어에 기대어  
제로샷 번역을 할 수 있음을 보임.



## SUMMARY

## NOTES

- Back translation  
 (= reverse translation)  
 : the process of  
 re-translating content  
 from TL back to SL

- 기존 단일 언어쌍 NMT 모델의 구조와 학습방법을 따르되,  
 다른 언어 처리를 위해 입력문장 뒷에 목적언어를 "붙여주는  
 특수토큰 추가,"
- 여러 언어를 통합한 사전 사용.
- 여러가지 언어들이 중간언어 형태의 표준을 학습  
 ⇒ 중간언어 방식 번역의 가능성을 보여줌.

③ 대량의 단일어 코퍼스를 활용  
(monolingual corpus)

- 예번역 (Back translation)
    - : 학습 시 목적언어 단일어 코퍼스를 이용
  - 결합 예번역 (Joint Back-Translation)
    - : 목적언어, 원시언어 단일어 코퍼스를 모두 이용
- "Iterative back-translation for  
neural machine translation"
- Dual Learning
    - : 강화학습을 통해 양방향 번역시스템을 동시학습
- "Dual learning for machine translation"

## NOTES

- Dual transfer learning  
: 주변분포 정규화와 전이학습에 기반  
" Dual transfer learning for neural machine translation with marginal distribution regularization "
- 디코딩 시 목적언어 코퍼스에 기반하여 사전학습된  
별도의 언어모델을 활용하는 방법 연구
- Mirror - Generative NMT  
: 기존 방벙들에서 번역모델이나 언어모델이 서로 연관성을  
갖고 학습되지 않는 점을 지적.
  - 원시언어  $\Rightarrow$  목적언어 번역모델
  - 목적언어  $\Rightarrow$  원시언어 번역모델
  - 원시언어 언어모델
  - 목적언어 언어모델

★  
모두 하나로  
경합한 구조를 제안

" Mirror - generative neural machine translation."



## KEY POINTS

NAME/DATE/SUBJECT

## NOTES

### - "Unsupervised Neural Machine Translation"

- 비지도 교차언어 임베딩 기법과 공유인코더 기반  
(Cross-lingual Embedding)

- 여러개 언어로 훈련된 코퍼스에서 각각의 단어 임베딩을 학습

- 자동으로 양방향 사전을 구축

- 이에 기반하여 공동의 임베딩으로 단어를 맵핑

- 디노이즈 인코더와 엣터넷을 이용하여 반복적으로 학습

- (Denoising Auto-Encoder)

## SUMMARY

## NOTES

- "Unsupervised machine translation using monolingual corpora only."

- 단일어 코퍼스에서 각 언어별 단어 이베딩을 학습

- ↓

- 양국어 사전 추출

- ↓

- 이를 기반으로 단어-대-단어 번역을 수행하는 초기 번역모델을 만들

- ↓

- 이를 바탕으로 오토인코더와 역번역에 기반하여 공동의 임재공간에 매핑

- "Explicit cross-lingual pre-training for unsupervised machine translation"

- 교차언어 단어 이베딩의 품질을 높이기 위해

- BERT 기반으로 사전 학습한 Cross-lingual Masked Language Model 을 인코더, 디코더로 활용

- 바탕으로 오토인코더와 역번역을 반복하는 형태를 비지도 학습 수행

## SUMMARY

## NOTES

## — "Revisiting low-resource neural machine translation"

- NMT가 low-resource 언어상에 대해 PBSMT 보다 향상하다는 것은 잘못된 인식이라고 빨빡.
- SMT 보다 낮은 성능을 보이는 이유:  
: 학습 코퍼스가 적은 상황을 고려하여 하이퍼파라미터를 설정하지 않았기 때문이라고 주장

↓

- 하부단어 단위 학습 시 최소빈도 기준값을 양국어 코퍼스의 크기에 맞춰 더 낮게 조정
- 더 작은 NMT 모델
- 더 작은 배치사이즈
- 더 적극적인 드롭아웃 등을 적용.

↓

- 양국어 코퍼스가 부족한 상황에서도 PBSMT 보다 좋은 성능을 보일 수 있음을 증명.

## SUMMARY

## NOTES

## 4. 2 문맥정보의 확장

- 일반적 NMT: 문장 단위를 기본으로 한 모델링.
- 문장을 넘어선 더 많은 문맥정보를 사용하면  
⇒ 더 정확한 번역 가능

생략 / 상호창조 같은 담화 현상 해결 가능



- “문서 단위 번역” 연구 중  
(Document - level Translation)

- ① 문맥정보를 통한 번역메모리 기본으로 가변적으로 관리?  
(Dynamic Translation Memory)
- ② 전체 문서를 문맥정보로 활용하는지?
- ③ 고정된 개수의 주변문장 일부만을 활용하는지?
- 에 따라 분류.

## KEY POINTS

"Learning to remember translation history with a continuous cache."

## NAME/DATE/SUBJECT

### NOTES

#### ① 동적 메모리 기반 방법

##### - 일반적인 메모리

: 고정 개수의  $(C_t, S_t, Y_t)$  트리플 정보를 저장  
어텐션  
버퍼 디코더  
온갖상태  
단어 번역

##### - 이전 번역 단어들의 온갖 표현을 동적 메모리에 저장

##### - Cache 를 NMT에 추가.

문서단위 번역 정보를 저장한

key-value 구조의

동적 번역 메모리 네트워크

##### - 각 디코딩 단계에서

현재 유사언어 문장에서의 문맥을 나타내는

어텐션 문맥벡터를 기준 하여

Cache에서 값을 읽어들여.

이전 번역 결과들이 저장된 것

목적언어 단어의 벡터값



##### - 정색된 값은 현재 디코더의 온갖상태에 대해서.

##### - 캐쉬는 문장의 번역이 완료된 후 번역결과를 반영하여 갱신

## SUMMARY

## NOTES

② 전체 문서를 문맥 정보로 활용

"Using whole document context in neural machine translation"

단어 아베이지 평균  
(word embedding average)

. 문장의 일부만을 사용하는 대신 더 넓은 문맥을 위해 문서 정보 자체를 사용하는 방법들  
주 가장 간단한 방법

- 문서 내의 OOV를 제외한 모든 단어의 아베이지 평균값을 문서단위 문맥 정보로 사용.

"Selective attention for context aware  
neural machine translation"

- 전체 문서를 문맥으로 고려하되, sparse attention을 이용하여 전체 문서 중에서 관련있는 문장들에 선택적 집중

↓

해당 문장들의 단어만을 고려하는 선택적 어텐션 방법 제안

## KEY POINTS

NAME/DATE/SUBJECT

## NOTES

### ③ 고정된 개수의 주변문장 일부분을 활용

- 문맥정보를 확장하는 상당수의 연구는 이전한 앞뒤의 주변문장만을 문맥정보로 활용

"Exploiting Cross-Sentence Context  
for neural machine translation"



- 문장 RNN과 문서 RNN으로 이루어진 계층적 RNN을 도입
- 이전 3개 워시마다 문장을 문맥정보로 사용.
- 문장 RNN이 각 문장을 입력받아 변화한 마지막 은닉상태를 문장 벡터로 추출
- ↓
- 문서 RNN이 이를 입력받아 문서벡터로 변환
- ↓
- 변환된 문맥정보는 이코더, 디코더의 상태 초기화에 사용  
및 문맥 게이트를 통해 디코더의 보조정보를 NNT에 전달됨.

## SUMMARY

# KEY POINTS

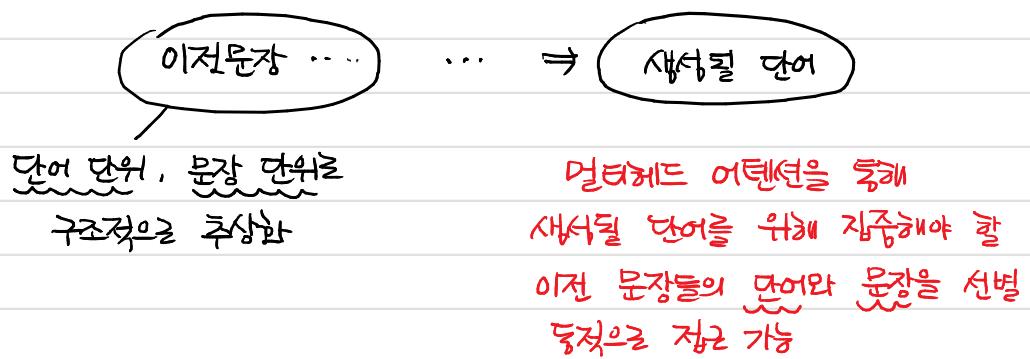
NAME/DATE/SUBJECT

## NOTES

"Document level neural machine translation  
with hierarchical attention networks"

- 구조화된 동적 패턴으로 문서 단위 문맥 정보를 획득하는 계층적 멀티헤드 어텐션 네트워크를 제안 (Hierarchical multi-head attention network, Multi-head HAN)

- 문맥정보 :



- 원시언어 문맥과 목적언어 문맥을 모두 반영하기 위해 2개의 HAN을 NMT에 결합



## SUMMARY

## NOTES

" When a good translation is wrong in context:  
 Context-aware machine translation  
 improves on deixis, ellipsis, and lexical cohesion."

- 문서단위 NMT : 대부분의 학습 코퍼스는 문맥 무관한 문장단위

- 문서단위 문맥정보가 없는 문장단위 번역모델을 먼저 학습하여 고정  
 ↓  
 • 문서단위 번역모델을 연동.

- 문서단위 번역모델 : 문장단위 번역모델의 번역결과를 이겨받아  
 문서단위 문맥정보를 기반으로 저결해주는 역할

- 디코더 : Transformer 디코더 구조  
 Multi-head attention 이 현재암과 +  
 이전 문장들도 이겨받을 수 있도록 수정.

Multi-head attention 블록 하나 더 추가  
 ⇒ 현재 번역 결과와 이전문장 번역 결과 고려

- ★ 서술 저항이 거의 없어 문서단위 문맥정보 반영.

## KEY POINTS

NAME/DATE/SUBJECT

## NOTES

### • Out-of-Distribution?

- : 학습 데이터의 분포와는 다른 분포를 갖는 데이터

분류 문제에서 는  
학습 데이터에 포함되지  
않은 class 를 가지  
데이터를 의미.

### 5. 결론 (추후주제)

- 번역 품질의 저하 없는 비자기회귀 모델 (Non-Autoregressive Model)
- 이미지 등을 추가 정보를 활용하는  
멀티모달 번역 (Multi Modal Translation)
- 실시간 번역 (Simultaneous Machine Translation)
- Speech-to-Text 번역
- OOD (Out-of-Distribution) 문장의 번역에도 강건한 번역
- Transformer 를 놓아하는 새로운 NMT 구조의 설계
- 학습과 추론 시의 불일치성 처리 등

## SUMMARY