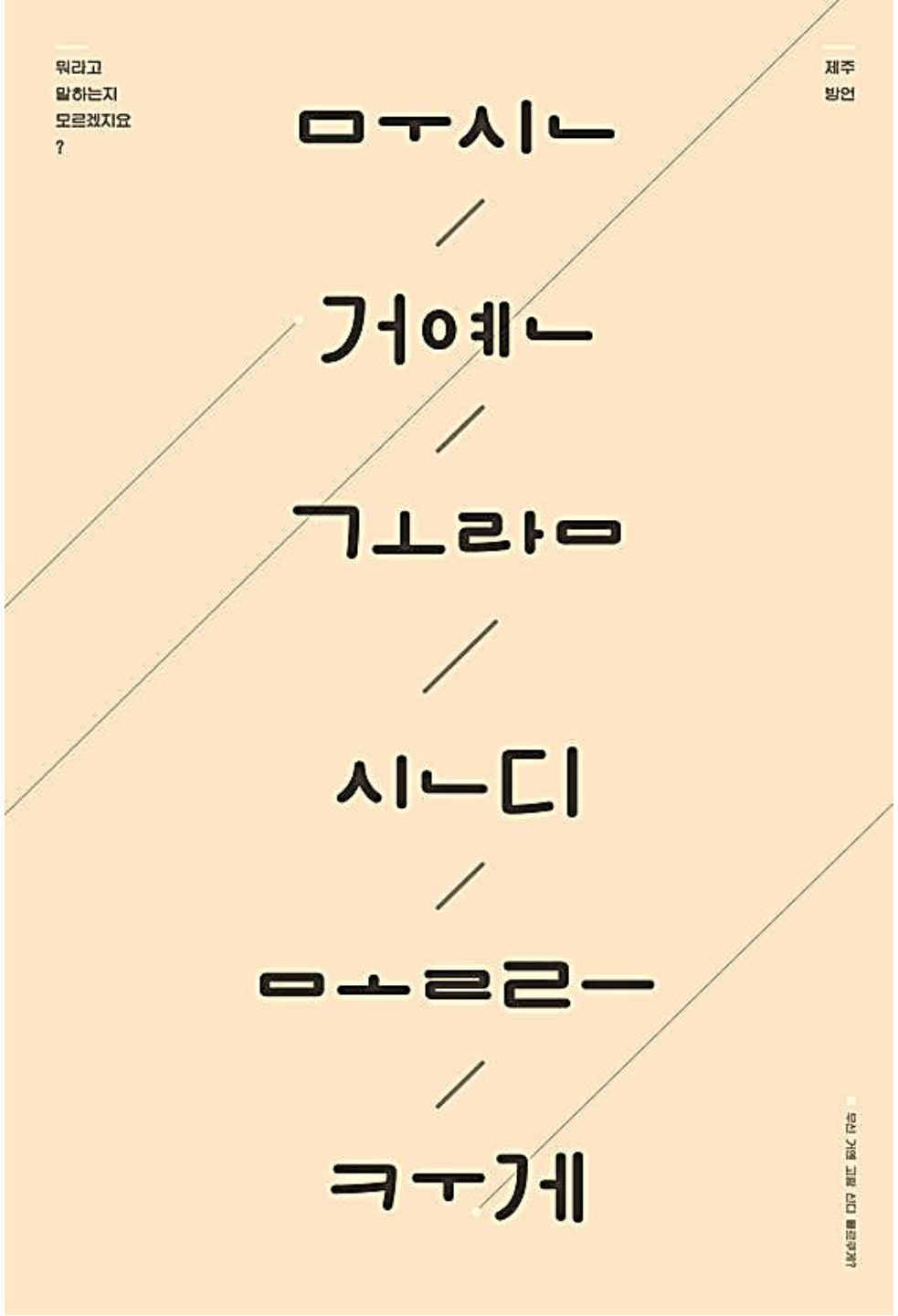
1 조
제주어
번역
프로젝트



1조 Final Project

introduce

임 영택

프로젝트 매니저
데이터 정제
모델링

송 지영

자료 수집
데이터 정제
모델링

김 주현

자료 수집
데이터 정제

장 하림

자료 수집
데이터 정제
프로젝트 정리

정 규진

자료 수집
데이터 정제
사전 만들기

contents

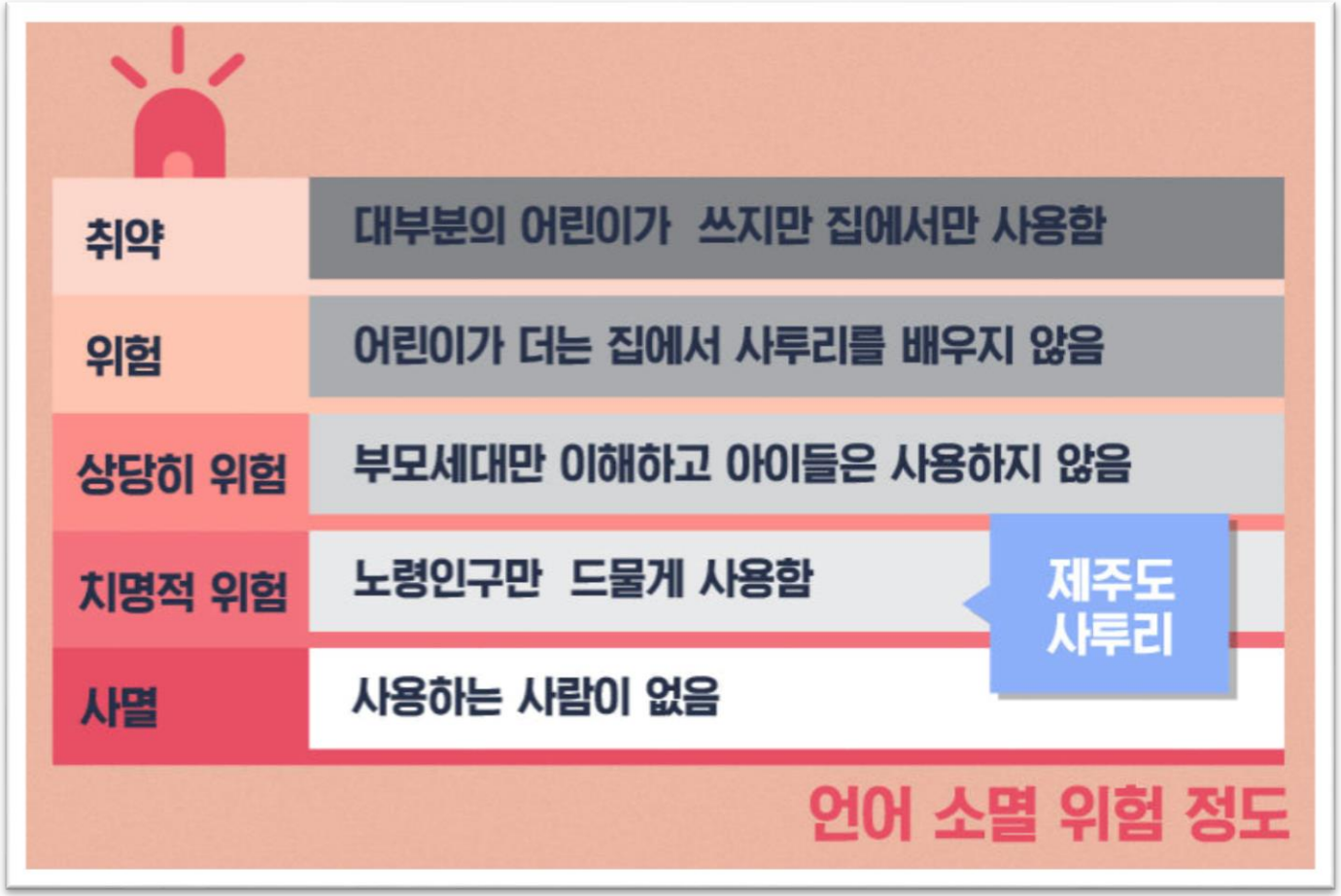
제주어 란?

“ 훈민정음에 가장 가까운 언어 ”

제 주 어

제주어는 한글의 제작원리를 보여주는 언어로,
그 특수성과 언어학적 가치를
국어학자나 언어학자들에게 인정 받았다

분석 배경



아틀라스 : 소멸 위기에 처한 언어 지도

제주어는 2010년

유네스코가 지정한

언어 소멸 위험 정도에서

4단계인

사멸 직전에 위치한다.

아틀라스에 따르면 지난 1세기 동안 지구상에서 200여개의 언어가 사라졌다.

최종 목표

제주어 -> 표준어

표준어 -> 제주어

로 번역해서 사람들에게

접근성이 용이하게끔 인식을

변화시켜 언어 소멸단계를 낮추는 것

데이터 설명

제주 민요

봉지가, 질군악,
산천초목, 오돌또기 등

제주 일상어

솔문 독새기 호나 줍서.
먹돌도 똘람 시민 고망이 난다. 햄시민 된다

제주 속담

친정이 가도 못 얻은 저냑ᄉᆞᆷ심
바당에 가민 얻나.

단어 사전

셍 거, 족은말젯, 뎅강, 돌코롬 등

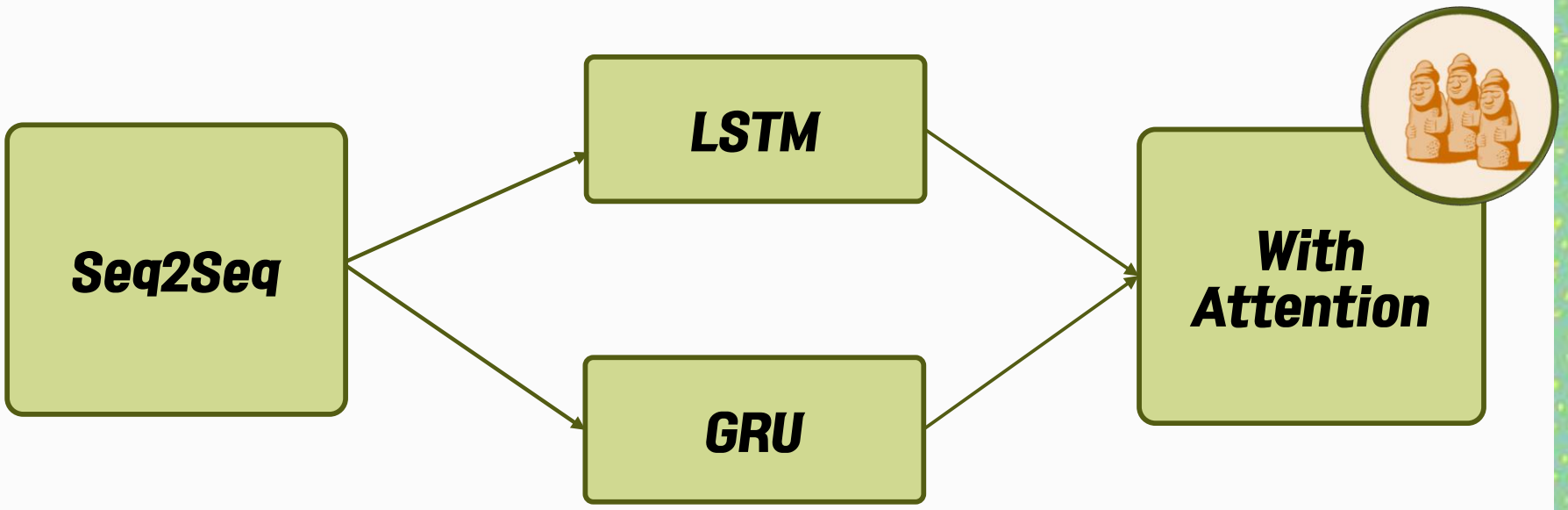
출처 : 제주특별자치도

품사 태깅 (사전 만들기)

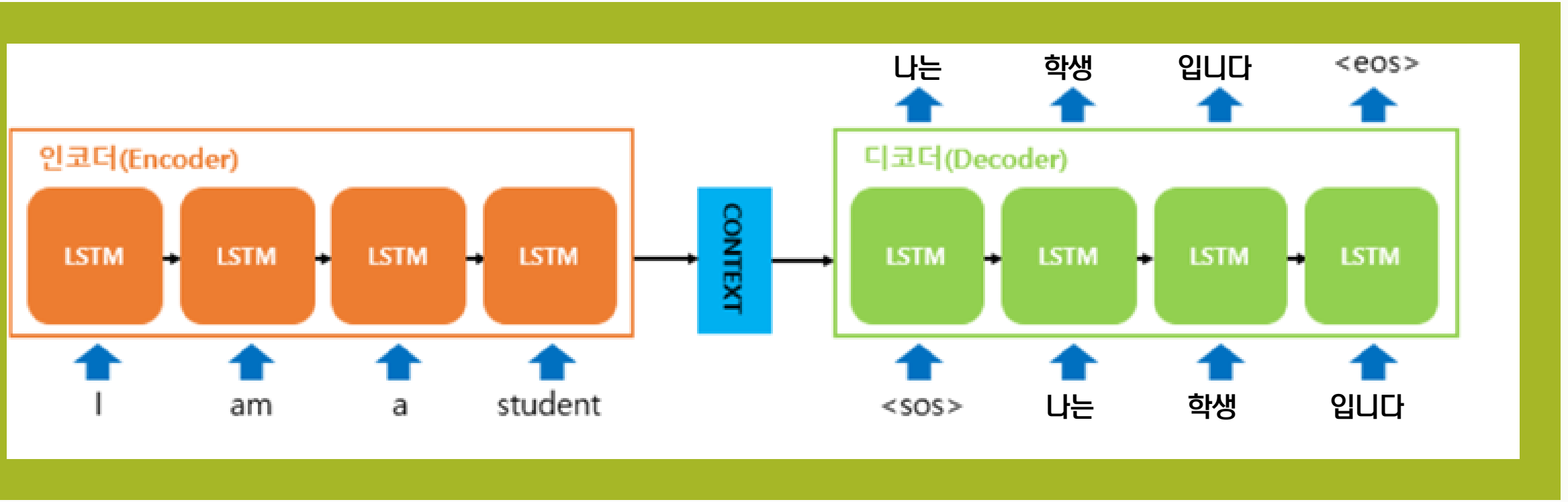
아야	Noun
아이고	Exclamation
아이고게	Exclamation
아이구	Exclamation
아진	Noun
아척	Noun
아프게	Adjective
아프곡	Verb
악살호다	Verb
안	Adverb
안녕	Noun
안녕하우꽈	Verb
안되카	Adverb
안되카마씀	Verb
안되쿠다	Adjective
안된댄	Adjective

OKT	Komoran	품사
Noun	NN	명사
Adverb	MA	부사
Determiner	NP	대명사
Exclamation	IC	감탄사
Adjective	VA	형용사
Suffix	XS	접미사
Verb	VV	동사
Josa	JK	조사
number	SN	숫자

Model Selection



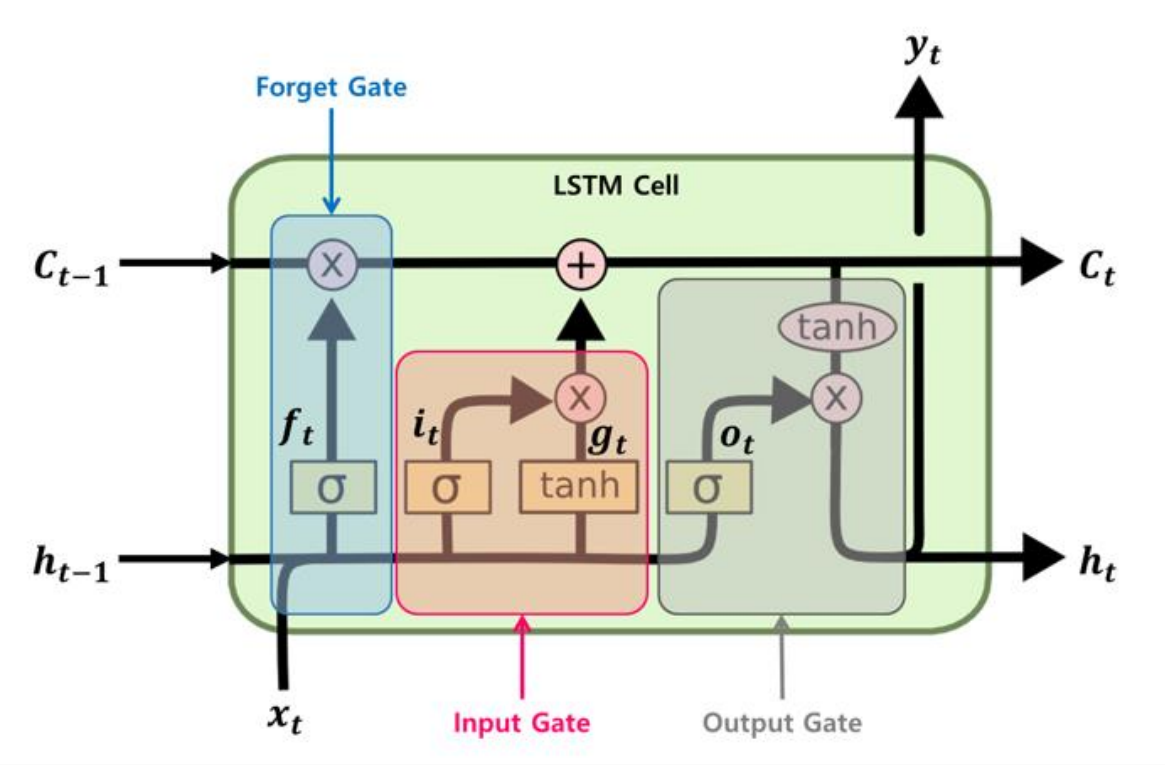
Seq2Seq (LSTM)

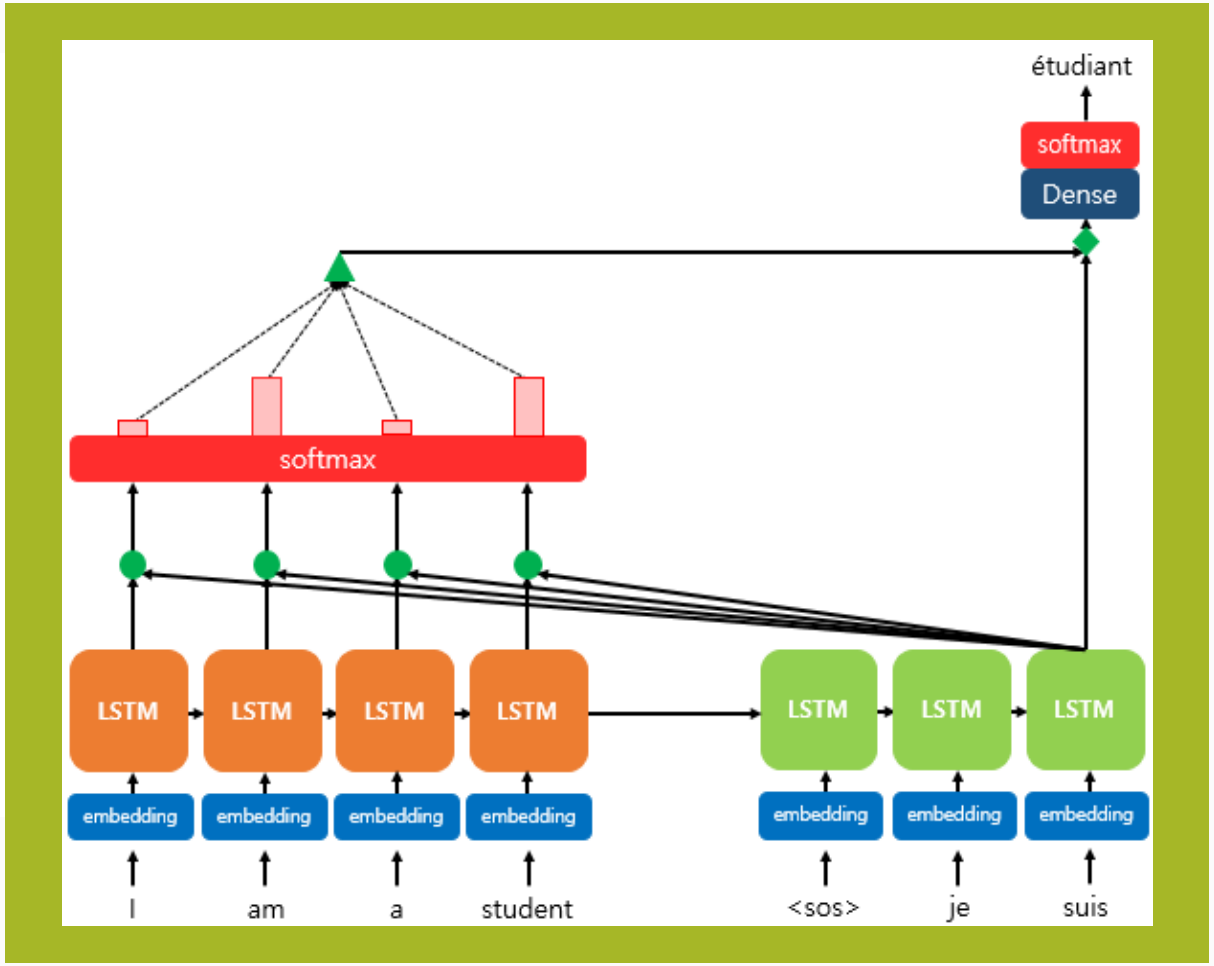


LSTM 결과

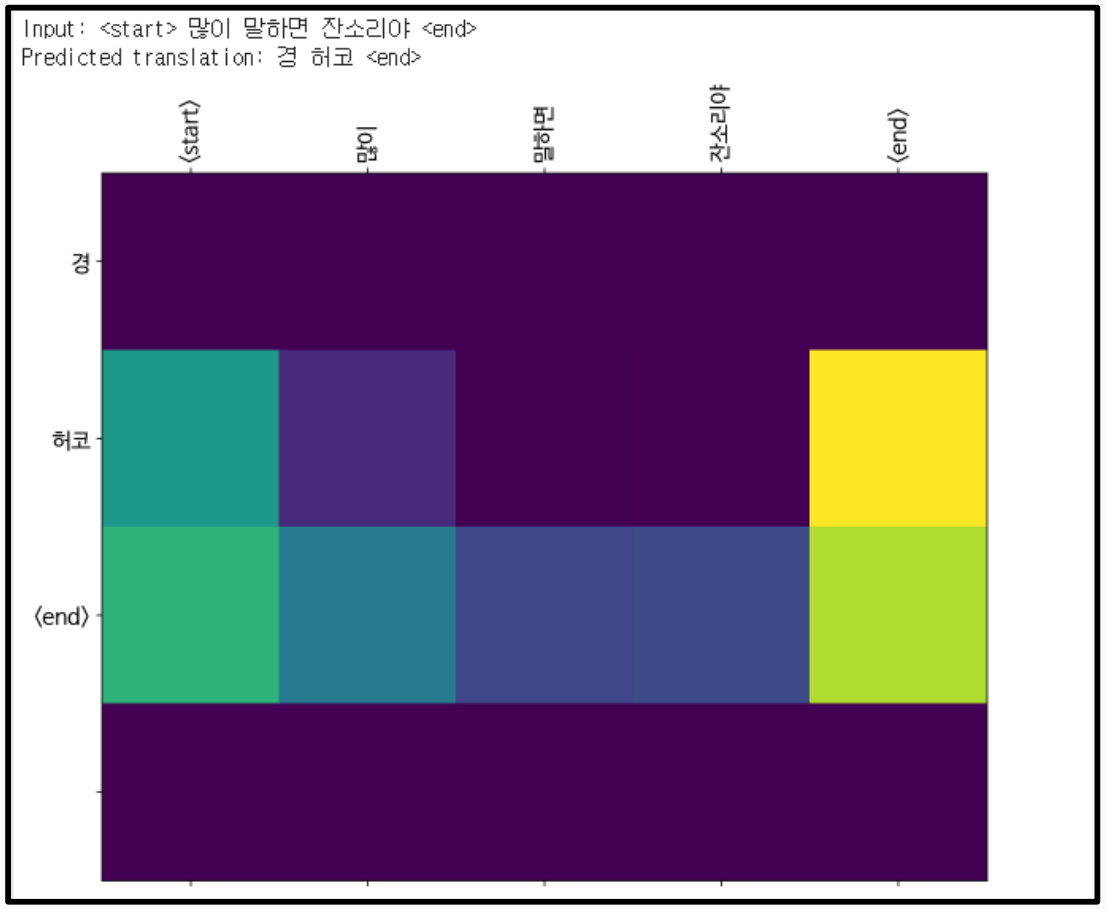
```
-----------------------------------
입력 문장: 어멍
정답 문장:  어머니
번역기가 번역한 문장:  아지
-----------------------------------
입력 문장: 마농
정답 문장:  마늘
번역기가 번역한 문장:  아지
-----------------------------------
입력 문장: 과지
정답 문장:  무당입는옷
번역기가 번역한 문장:  아지
-----------------------------------
입력 문장: 돌생기
정답 문장:  돌맹이
번역기가 번역한 문장:  아지
```

LSTM with Attention





LSTM with Attention 결과




```
translate(lines[1])
```

```
Input: 너와 함께 하면 즐거움이 열배야 <end>
Predicted translation: 느영 고찌 글민 지꺼짐이 열배여 <end>
```

```
translate(lines[2])
```

```
Input: 같이 가요 함께 해요 <end>
Predicted translation: 촘말로 곱고 몬트락허우다 <end>
```

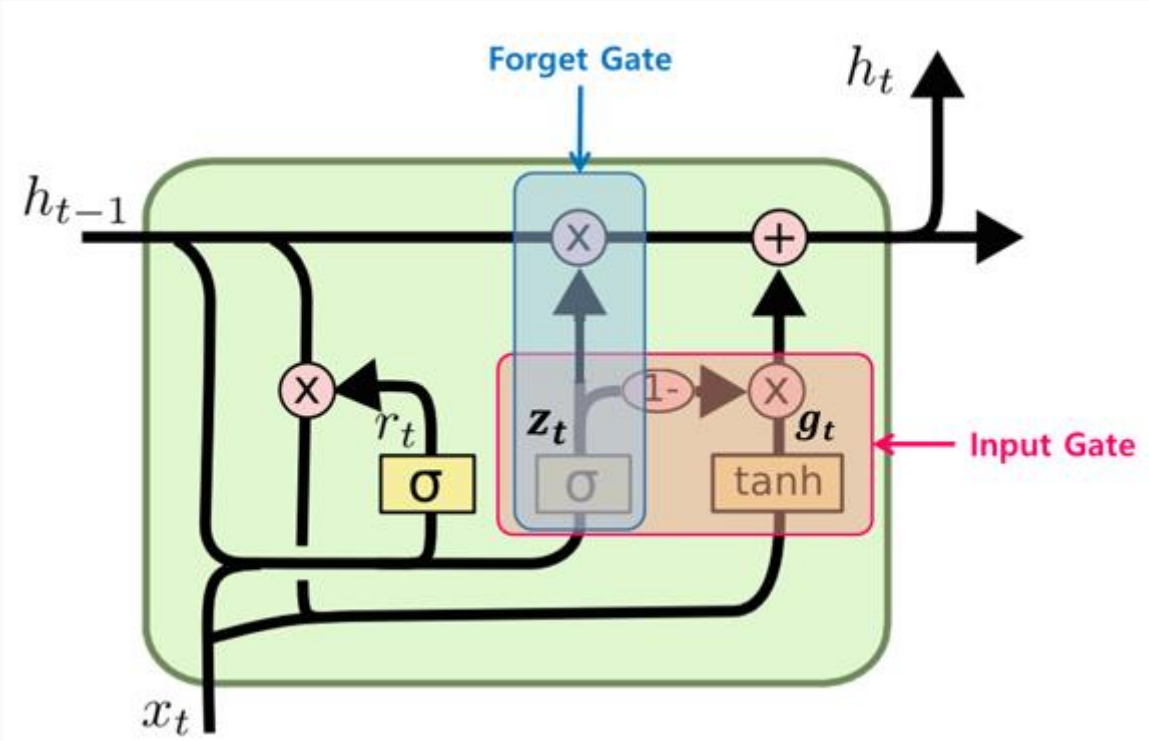
```
translate(lines[4])
```

```
Input: 이렇게 예쁜 날 공기 좋고 사람 좋고 <end>
Predicted translation: 영도 곱닥헌 날 공기 좋고 <end>
```

```
translate(lines[5])
```

```
Input: 너와 함께 하니 무슨 걱정 있으랴 <end>
Predicted translation: 경 돌암서? <end>
```

GRU with Attention

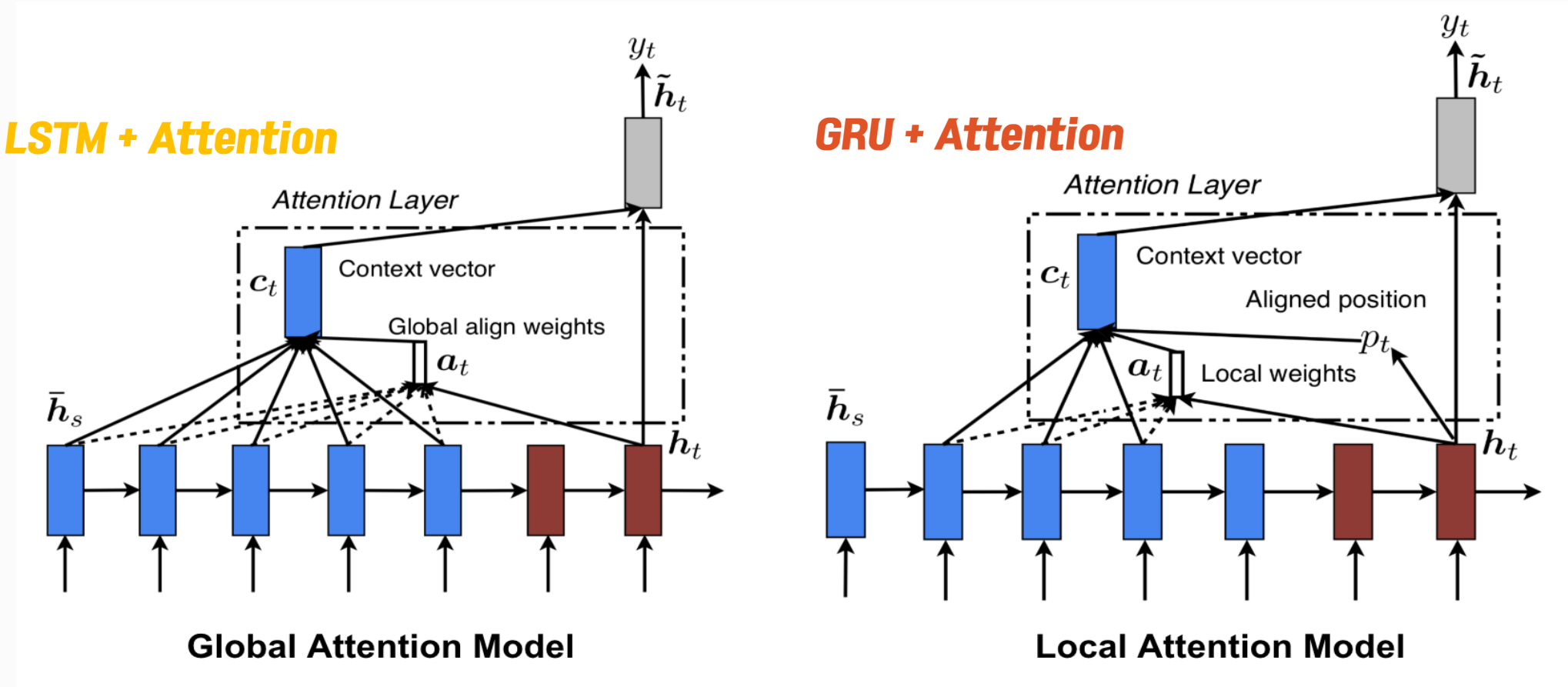


1. Bahdanau attention
2. Luong attention (general score)

Luong $$\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) = \begin{cases} \boldsymbol{h}_t^\top \bar{\boldsymbol{h}}_s & \textit{dot} \\ \boldsymbol{h}_t^\top \boldsymbol{W}_a \bar{\boldsymbol{h}}_s & \textit{general} \\ \boldsymbol{v}_a^\top \tanh\left(\boldsymbol{W}_a[\boldsymbol{h}_t; \bar{\boldsymbol{h}}_s]\right) & \textit{concat} \end{cases}$$

$$\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) = \begin{cases} \boldsymbol{h}_t^\top \boldsymbol{W} \bar{\boldsymbol{h}}_s & \text{[Luong's multiplicative style]} \\ \boldsymbol{v}_a^\top \tanh\left(\boldsymbol{W_1} \boldsymbol{h}_t + \boldsymbol{W_2} \bar{\boldsymbol{h}}_s\right) & \text{[Bahdanau's additive style]} \end{cases}$$

GRU with Attention



GRU with Attention 결과

"촐람생이 " -> "경솔한사람"
"좀짐팽이 " -> "종아리"
"촘지금 " -> "참기름"
"베불레기 " -> "아기옷"
"빈네 " -> "비녀"
"주멩기 " -> "주머니"
"갈체 " -> "상태기"
"도새기 " -> "돼지"
"빙아리 " -> "병아리"
"간드랑호다 " -> "시원하다"
"썸찌근호다 " -> "지긋지긋하다"
"그저끼 " -> "그저께"
"구숭호다 " -> "불평을말하다"
"속솜호다 " -> "잠잠하다"
"상고지 " -> "무지개"
"고찌 글라 고찌 가게 " -> "같이 가요 함께 해요"
"4. 3 소건때 총 맞앙 턱을 잃엉 경 살았주" -> "4. 3 사건때 총 맞아서 턱을 잃었네 그렇게 살았네"
"돈 한 추룩 뺄라진 추룩 오시록 헌 추룩" -> "돈이 많은 척 잘난 척 점 잖은 척"
"나오는 걸 어떵 헐 수 이시냐" -> "나오는 걸 어떻게 할 수 있겠니"
"바당 바당 제주 바당 보름 보름 보름" -> "바다 바다 바다 바다 바람 바람 바람 바람"
"이녁 가슴 쏘곱엔 " -> "당신의 마음 속엔"
"이어도 이어도 사나" -> "이어도 이어도 사나"
"절이치는 바당더레 강 보난" -> "파도치는 바다로 가 보니"
"까메기 똥 케우리듯" -> "까마귀 똥 헤집듯"
"것 박접허면 죄 짓나" -> "음식을 박대하면 죄 짓는다."

"어멍 " -> "어머니"
"또꼬망 " -> "똥구멍"
"돗괴기 " -> "돼지고기"
"단취 " -> "단추"
"시미옷 " -> "손자용상복"
"요령 " -> "방울"
"작박 " -> "작은바가지"
"잠데 " -> "쟁기"
"물꾸럭 " -> "문어"
"감저 " -> "고구마"
"멘도롱호다 " -> "따뜻하다"
"얼탑지다 " -> "당황하다"
"데싸지다 " -> "자빠지다"
"데껴불다 " -> "던져버리다"
"일고 " -> "일곱"
"통시 " -> "화장실"
"먹엄직이 살암직이 시상" -> "먹어 볼만한 살아 볼만한 세상"
"모멀고루 풀엉 얄롭게 지경" -> "메밀가루 반죽하여 얇게 지져"
"고루삭삭 삐어지국" -> "사방으로 흩어지고"
"경해그네 어떵 되연" -> "그래서 어떻게 되었어"
"정 골아도 빙세기 웃곡" -> "저렇게 말해도 빙그레 웃고"
"말 골암쩌" -> "말을 하네"
"이래 도라앉작 저래 도라앉작" -> "이리 흔들리고 저리 흔들리고"
"아고 삼촌 물꾸럭 나 얼마마씸?" -> "아고 삼촌 문어 한마리 얼마죠?"

* 평가방법 BLEU 란?

입력 값 : The teacher arrived late because of the traffic. (Source Original)

정답 값 : 선생님은 교통체증 때문에 늦게 도착 하셨다. (Reference Translation)

예측 값 :

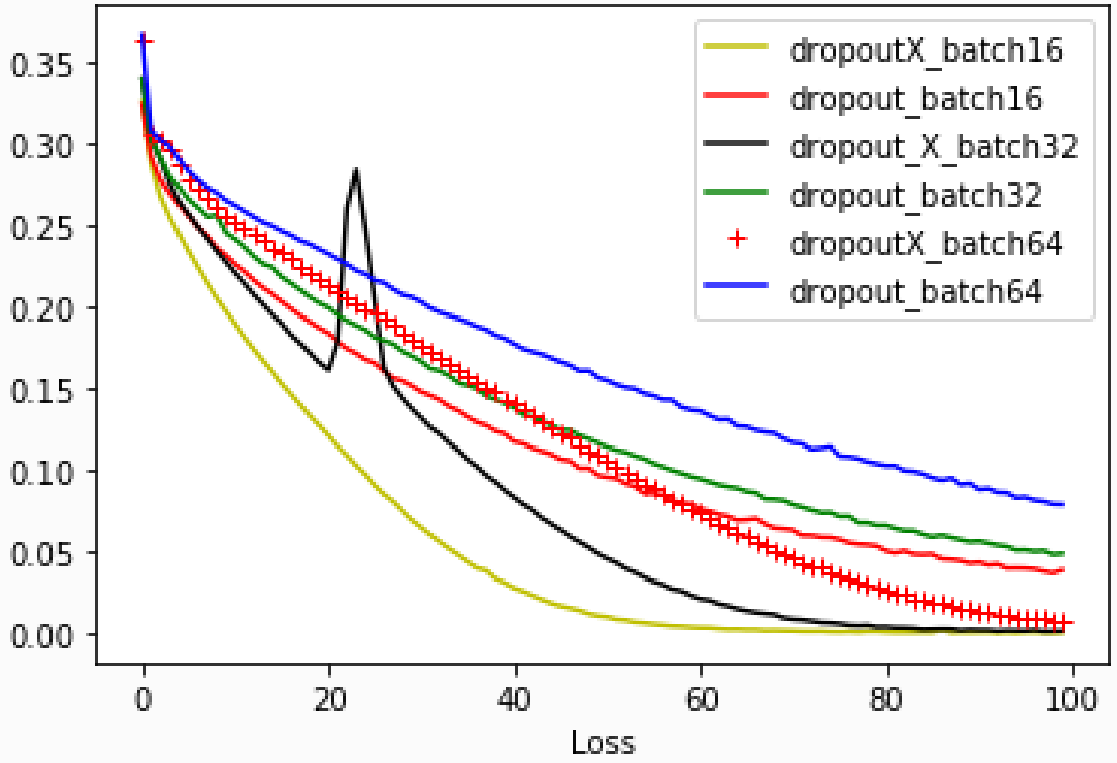
교수님은 교통 체증으로 인해 연착되었다. — #1 Very low BLEU score
선생님이 지각한 것은 혼잡 때문이었다. — #2 Slightly higher but low BLEU
선생님은 교통체증 때문에 늦으셨다. — #3 Higher BLEU than #1 and #2
교수는 교통체증 때문에 늦게 도착했다. — #4 Higher BLEU than #3

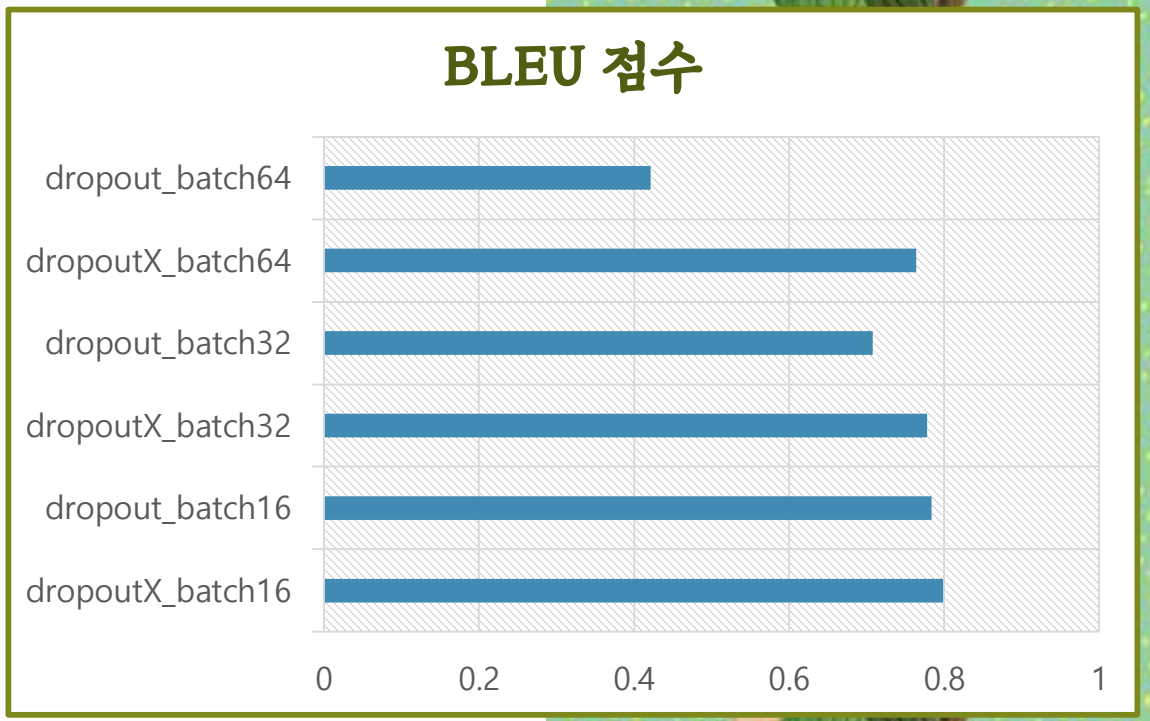
Best : 선생님은 교통체증 때문에 늦게 도착 하셨다. — #5 *Best BLEU Score*

Many accurate and correct translations can score lower
Simply because they use different words

green = 4-gram match (very good!)
turquoise = 3-gram match (good)
red = word not matched (bad!)

BLEU 모델 검증


dropoutX_batch16
dropout_batch16
dropout_X_batch32
dropout_batch32
dropoutX_batch64
dropout_batch64
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
20
40
60
80
100
Loss


BLEU 점수
dropout_batch64
dropoutX_batch64
dropout_batch32
dropoutX_batch32
dropout_batch16
dropoutX_batch16
0
0.2
0.4
0.6
0.8
1

훈저옵서

프로젝트 결론

제주어 데이터 수집 부족
제주어 표기, 컴퓨터에 인식 부족 (고전어)
제주어 전문지식 부족 (품사 태깅 미흡)
제주어 번역에 더 적합한 모델을 찾지 못함
Ex) self.Attention, Bert

참고 문헌 및 출처

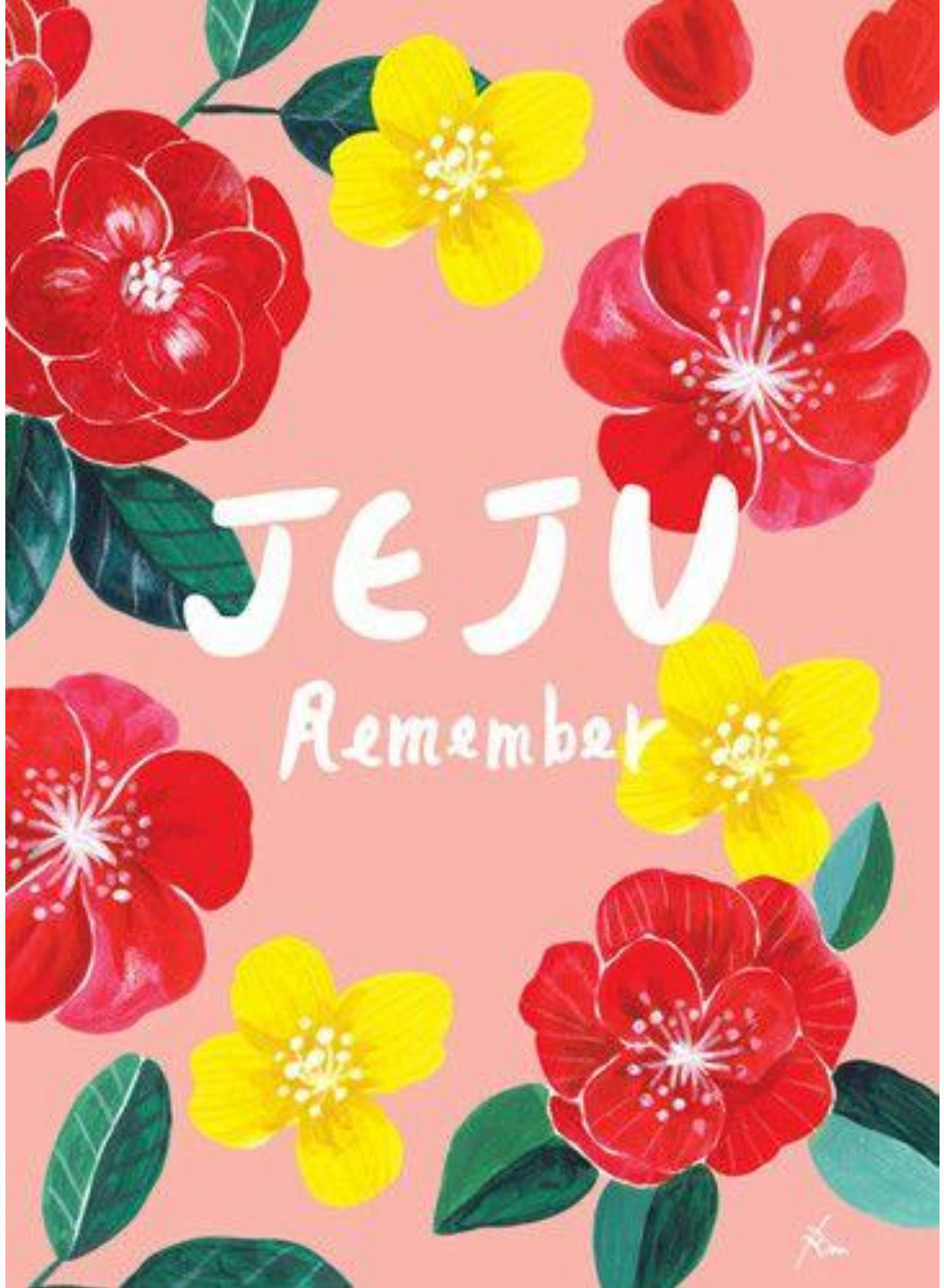
- *"Attention is all you need" (Submitted on 12 Jun 2017 (v1), last revised 6 Dec 2017 (this version, v5))*
  Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin

- *"Effective Approaches to Attention-based Neural Machine Translation (Luong Attention)"*
  Luong et al.(2015)

- *"LSTM(Long Short Term Memory)" 2014* & *"BiLSTM with attention" 2016* & *"Gated Recurrent Unit(GRU) with attention"*
  Cho et al.

- *Neural Machine Translation by Jointly Learning to Align and Translate (Bahdanau Attention)*
  Bahdanau et al.(2014)

- 제주특별자치도
  http://www.jeju.go.kr/



Thank you