

# 1 조 제주어 번역 프로젝트

뭐라고  
말하는지  
모르겠지요  
?

모ᄃ시ᄃ

거예ᄃ

고ᄃ라ᄃ

시ᄃ디

모ᄃ르ᄃ-

크ᄃ게

제주  
방언

무진 거예 고ᄃ 시디 불려ᄃᄃ



# *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



Welcome to  
Jeju Island



홍지음서

# contents

---

## I. 분석 배경

분석 목적  
데이터 설명

## II. 분석 과정

품사 태깅  
모델링 선정

## III. 분석 결과

분석 모형  
한계점

## IV. 결론

프로젝트 결론  
참고문헌



Welcome to  
Jeju Island



홍지음서



# 제주어 란?

“

훈민정음에 가장 가까운 언어

”

제 주 어

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



# 분석 배경



제주어는 2010년  
유네스코가 지정한  
언어 소멸 위험 정도에서  
4단계인  
사멸 직전에 위치한다.



# 최종 목표

제주어 -> 표준어  
표준어 -> 제주어  
로 번역해서 사람들에게  
접근성이 용이하게끔 인식을  
변화시켜 언어 소멸단계를 낮추는 것



Welcome to  
Jeju Island



홍저옵서

# 데이터 설명

## 제주 일상어

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

## 단어 사전

생 거, 족은말젯, 덩강, 돌코롬 등

## 제주 민요

봉지가, 질군악,  
산천초목, 오돌또기 등

## 제주 속담

친정이 가도 못 얻은 저녁드심  
바당에 가민 얻나.





# 품사 태깅 (사전 만들기)



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

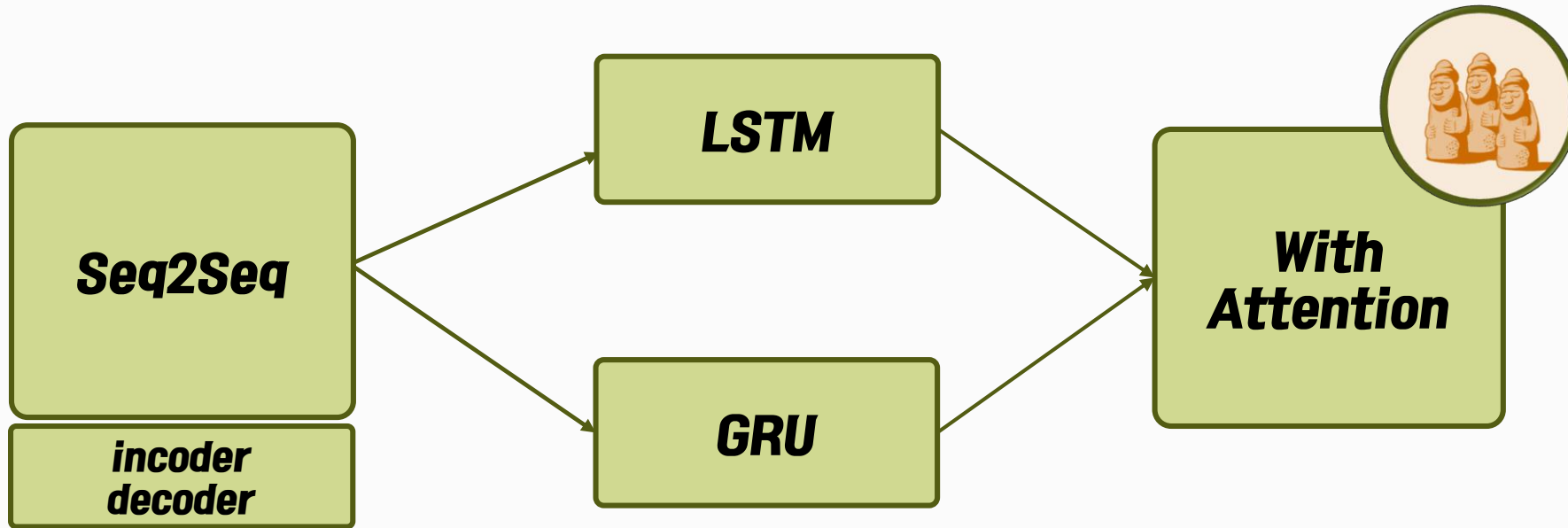
OKT	Komoran	품사
Noun	NN	명사
Adverb	MA	부사
Determiner	NP	대명사
Exclamation	IC	감탄사
Adjective	VA	형용사
Suffix	XS	접미사
Verb	VV	동사
Josa	JK	조사
number	SN	숫자

홍지음서

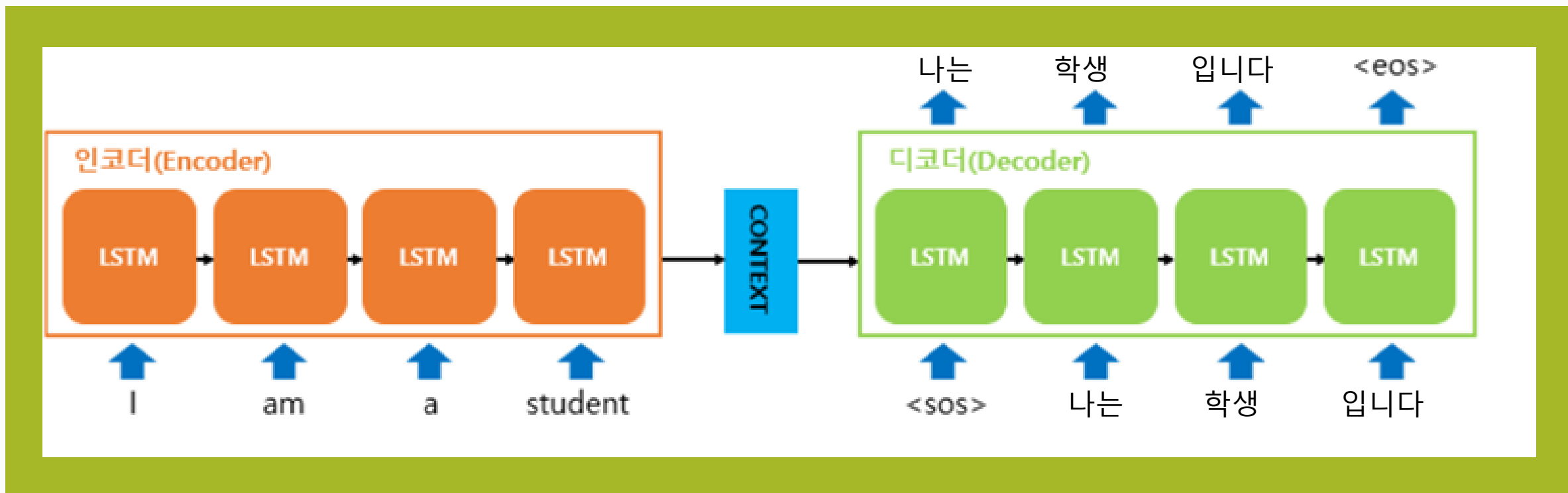


# Model Selection

---



# 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



Welcome to  
Jeju Island

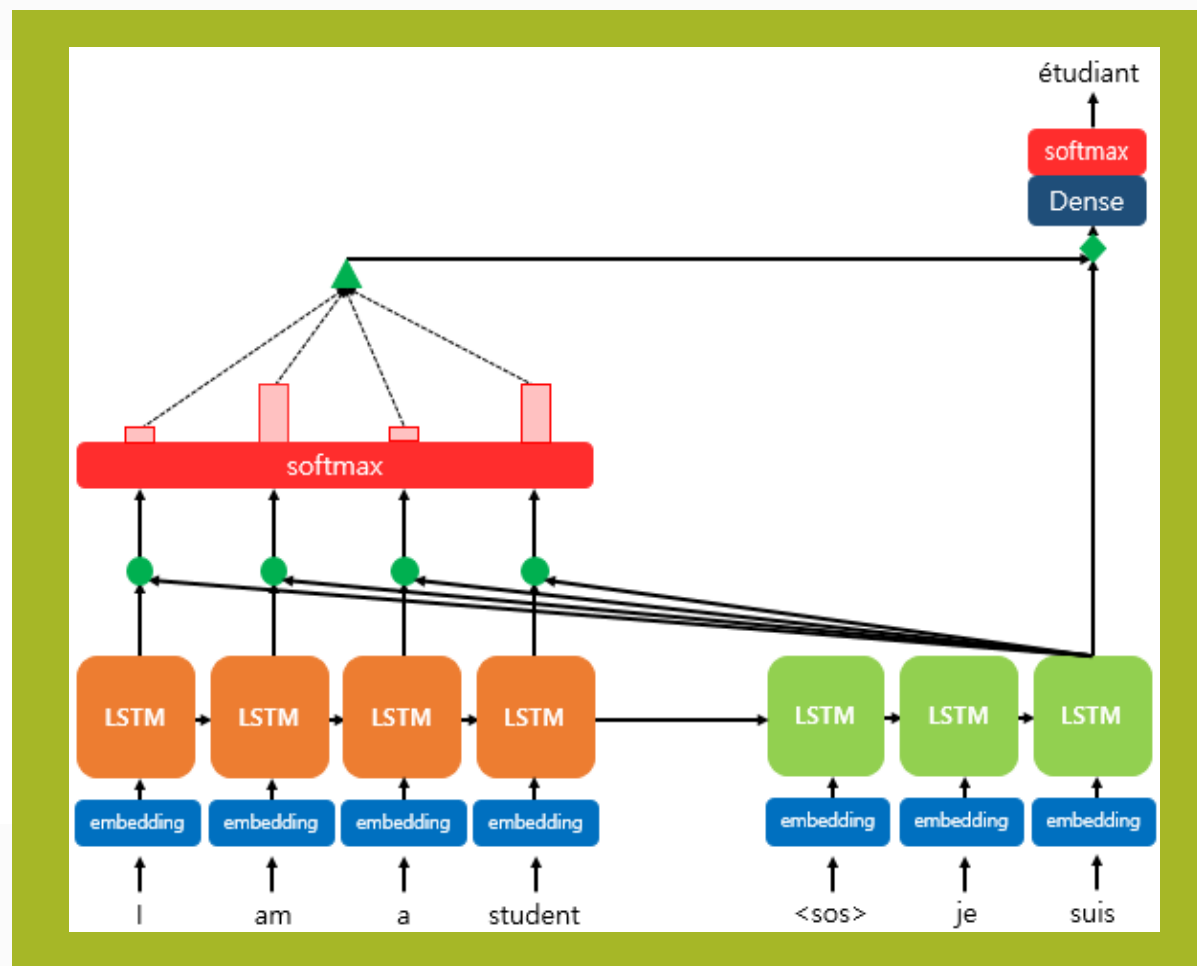
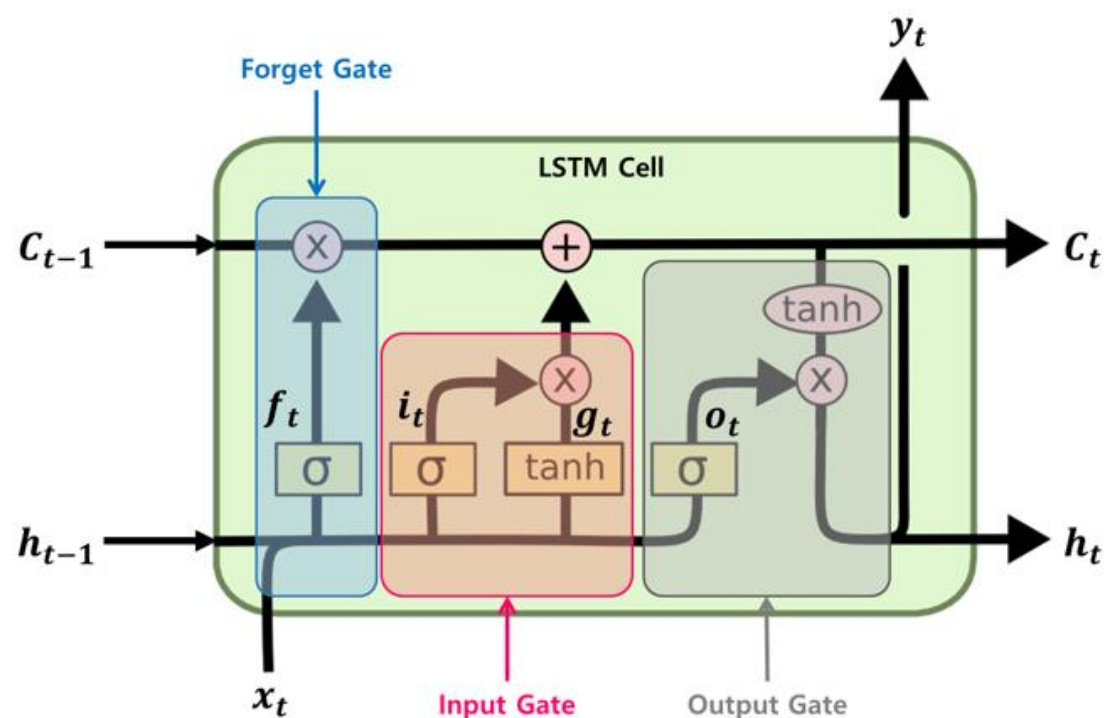


홍제읍서





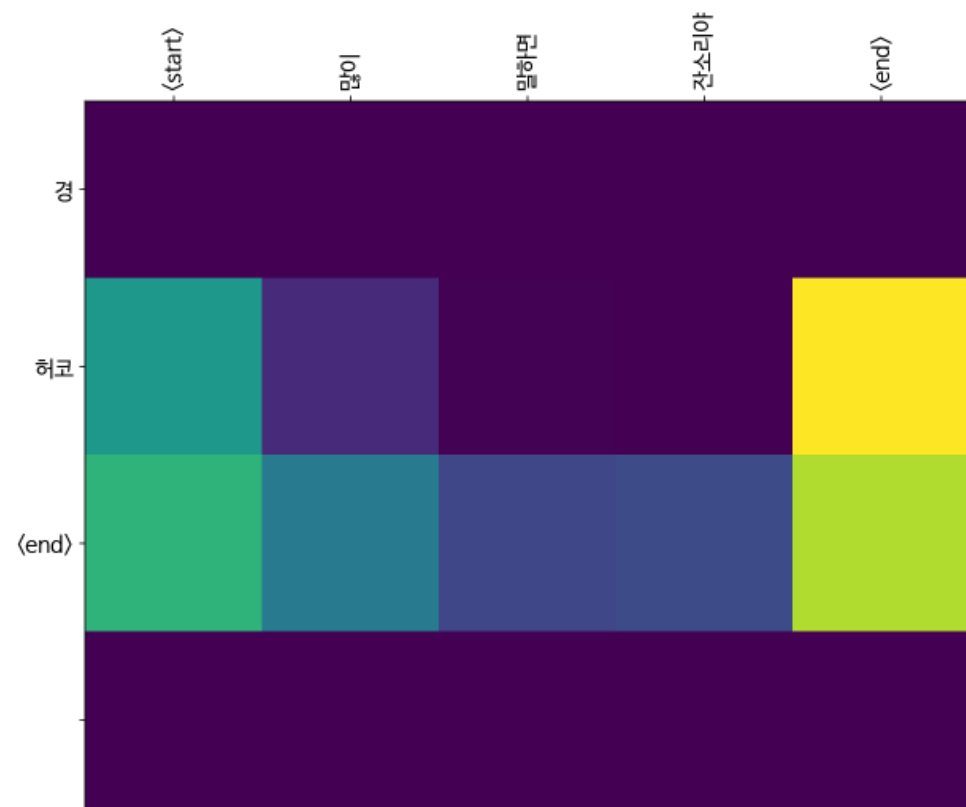
# ***LSTM with Attention***



# LSTM with Attention 결과



Input: <start> 많이 말하면 잔소리야 <end>  
Predicted translation: 경 허코 <end>



`translate(lines[1])`

Input: 너와 함께 하면 즐거움이 열배야 <end>

Predicted translation: 느영 고찌 글민 지꺼집이 열배여 <end>

`translate(lines[2])`

Input: 같이 가요 함께 해요 <end>

Predicted translation: 춤말로 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>

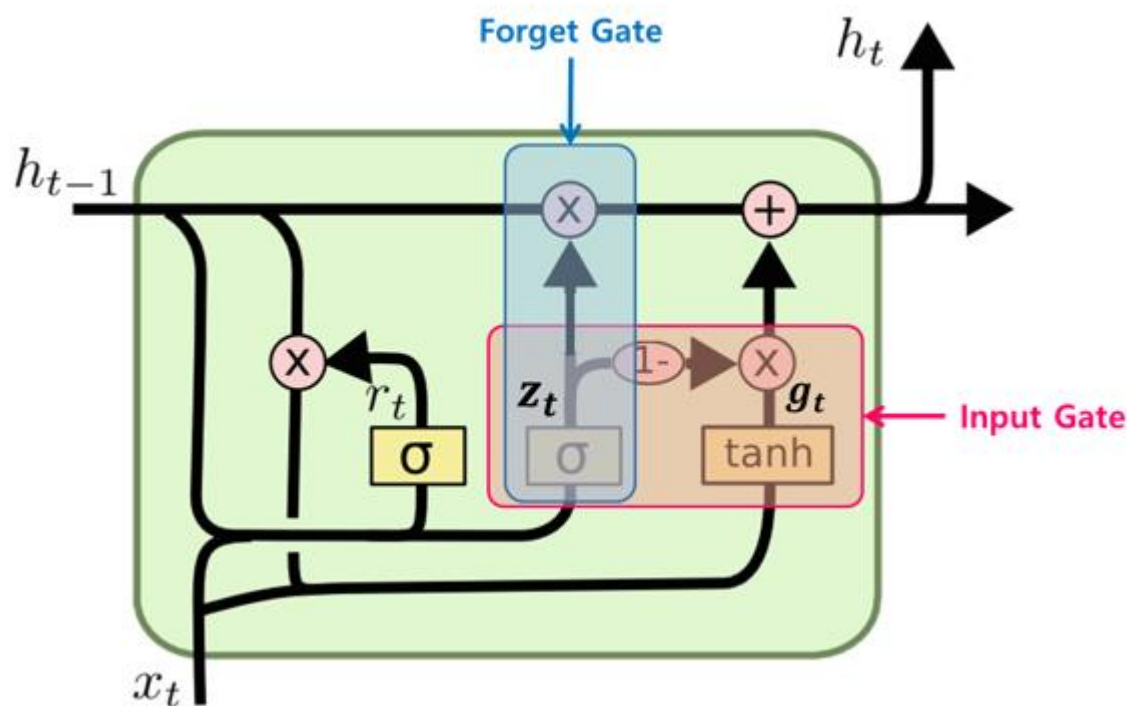


# GRU with Attention

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

$$\text{Luong score}(h_t, \bar{h}_s) = \begin{cases} h_t^\top \bar{h}_s & \text{dot} \\ h_t^\top W_a \bar{h}_s & \text{general} \\ v_a^\top \tanh(W_a [h_t; \bar{h}_s]) & \text{concat} \end{cases}$$

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

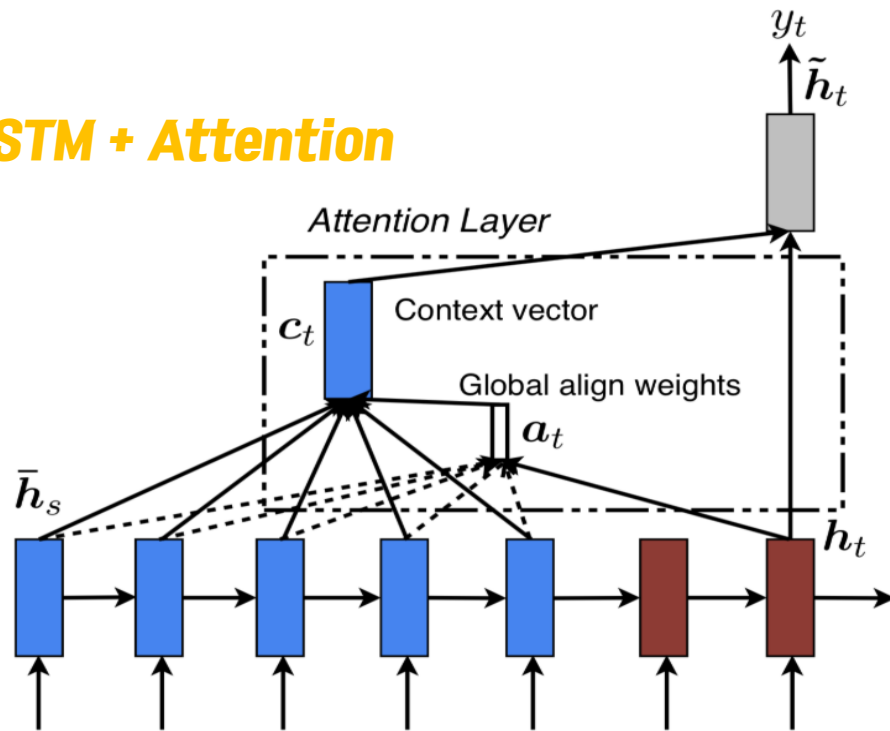






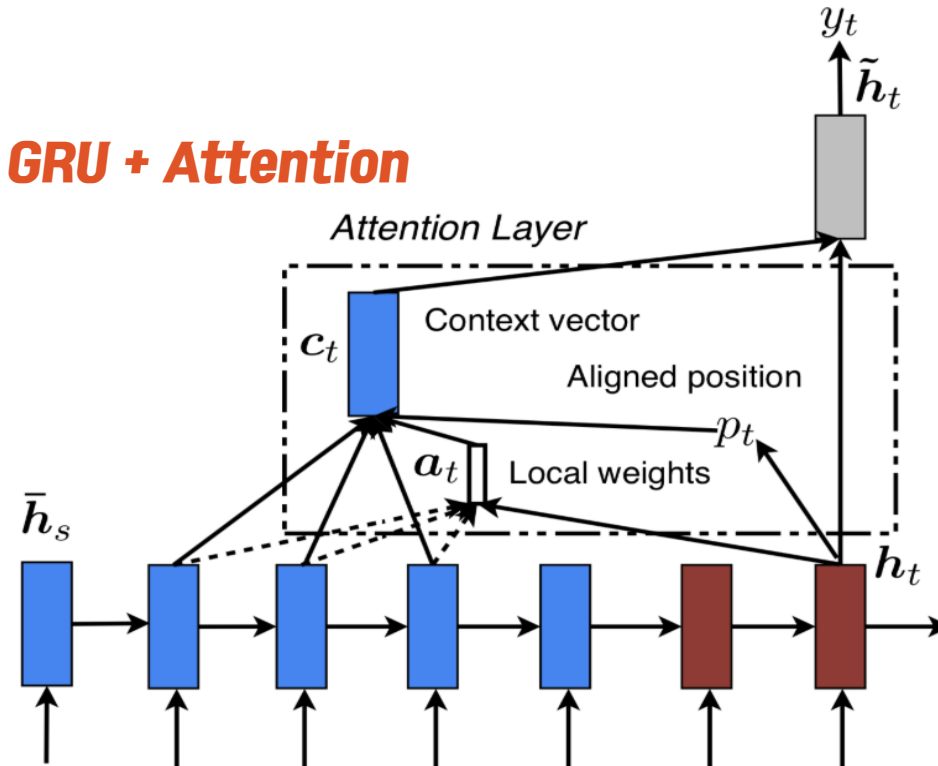
# GRU with Attention

## LSTM + Attention



Global Attention Model

## GRU + Attention



Local Attention Model

"어멍 " -> "어머니"  
 "또꼬망 " -> "똥구멍"  
 "똥괴기 " -> "돼지고기"  
 "단취 " -> "단추"  
 "시미웃 " -> "손자용상복"  
 "요령 " -> "방울"  
 "작박 " -> "작은바가지"  
 "잠데 " -> "쟁기"  
 "물꾸럭 " -> "문어"  
 "감저 " -> "고구마"  
 "멘도롱호다 " -> "따뜻하다"  
 "얼랍지다 " -> "당황하다"  
 "데싸지다 " -> "자빠지다"  
 "데꺼불다 " -> "던져버리다"  
 "일고 " -> "일곱"  
 "통시 " -> "화장실"  
 "먹엄직이 살암직이 시상" -> "먹어 볼만한 살아 볼만한 세상"  
 "모덜고루 풀엉 알룸게 지정" -> "메밀가루 반죽하여 얇게 지저"  
 "고루삭삭 빼어지국" -> "사방으로 흩어지고"  
 "경해그네 어뎡 되연" -> "그래서 어떻게 되었어"  
 "정 골아도 빙세기 웃곡" -> "저렇게 말해도 빙그레 웃고"  
 "말 골암쩌" -> "말을 하네"  
 "이래 도라앗작 저래 도라앗작" -> "이리 흔들리고 저리 흔들리고"  
 "아고 삼춘 물꾸럭 나 얼마마씩?" -> "아고 삼춘 문어 한마리 얼마쥬?"

# \* 평가방법 BLEU 란?

**입력 값 :** The teacher arrived late because of the traffic.

(Source Original)

**정답 값 :** 선생님은 교통체증 때문에 늦게 도착 하셨습니다.

(Reference Translation)

**예측 값 :** 교수님은 교통 체증으로 인해 연착되었다.  
선생님이 지각한 것은 혼잡 때문이었다.  
선생님은 교통체증 때문에 늦으셨다.  
교수는 교통체증 때문에 늦게 도착했다.

#1 Very low BLEU score  
#2 Slightly higher but low BLEU  
#3 Higher BLEU than #1 and #2  
#4 Higher BLEU than #3

**Best :** 선생님은 교통체증 때문에 늦게 도착 하셨습니다.

#5 *Best BLEU Score*

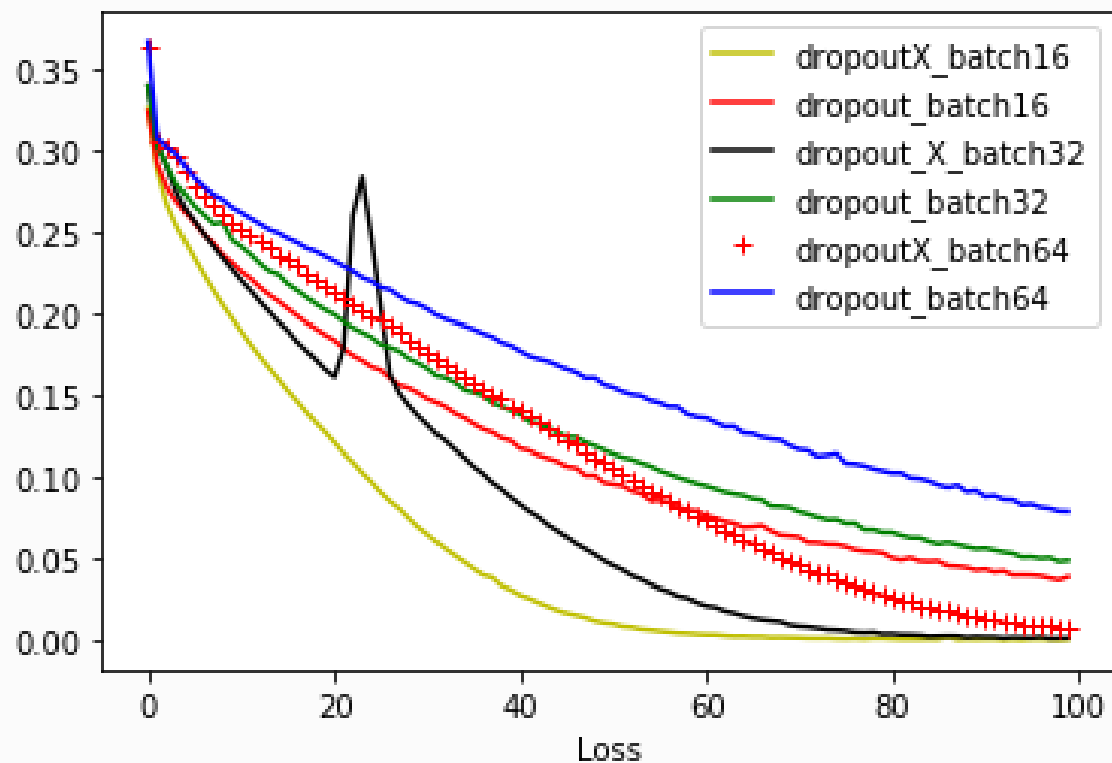
Many accurate and correct translations can score lower  
Simply because they use different words

green = 4-gram match  
turquoise = 3-gram match  
red = word not matched

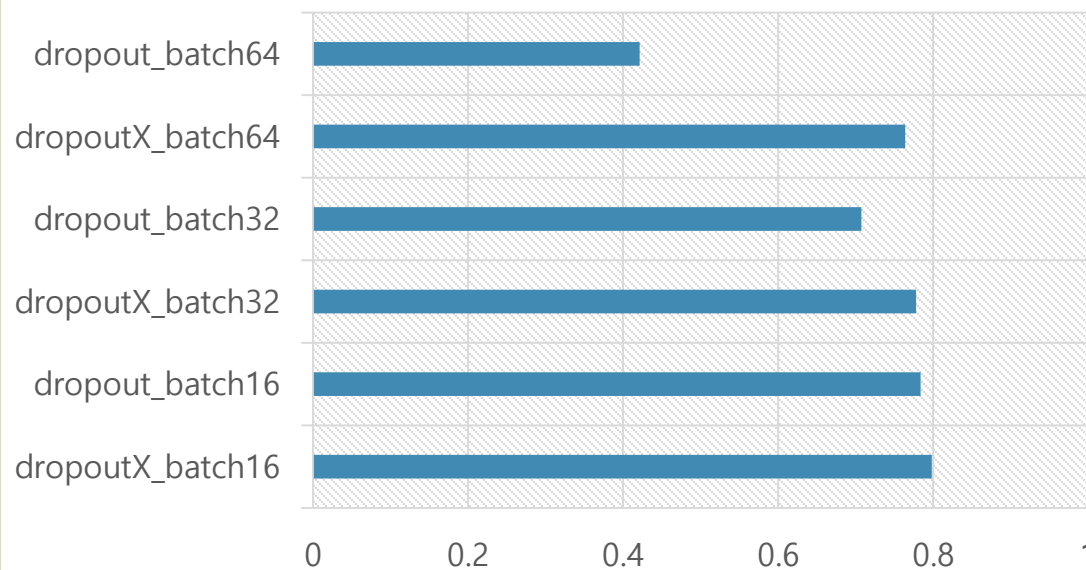
(very good!)  
(good)  
(bad!)



# BLEU 모델 검증



## BLEU 점수



훈제음서

# 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



Welcome to  
Jeju Island



홍지음씨

# 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/>

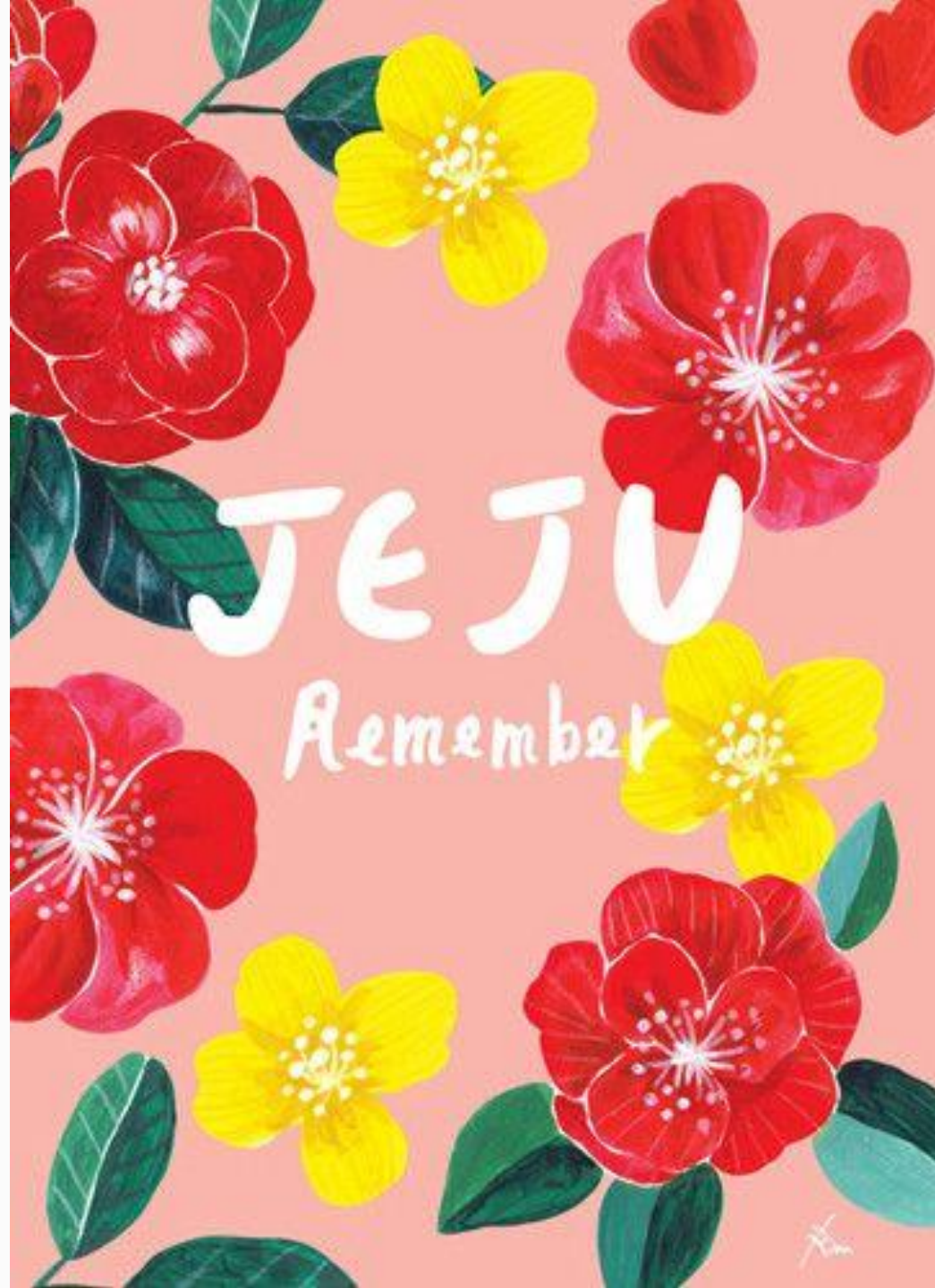


Welcome to  
Jeju Island



홍제읍서





***Thank you***