

# Deep Learning ChatBot Module 개발

도전!



# 도전!



# 도전!

Copy (Many)Code Line by Line

while(understand==false) Read reference, article, paper..

print(Log), print(Log), print(Log) ...

완벽한 이해보다는 구현을 먼저..



# 개발!

## Intend Inference

사용자의 발화 의도를 파악하고, 의미있는 정보를 추려낸다.

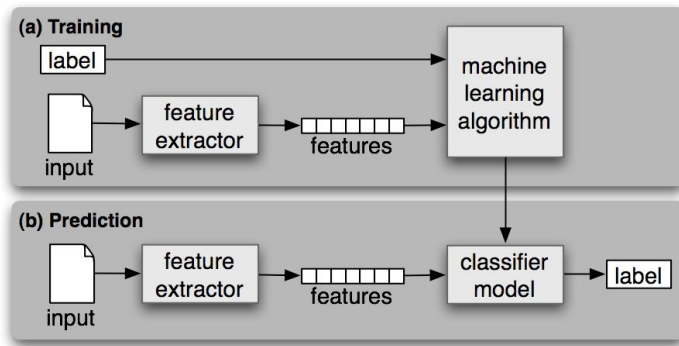
## Seq2Seq

일상에서 사용하는 대화를 이해하고 자연스러운 응답을 생성한다.

# 개발!

## Intend Inference

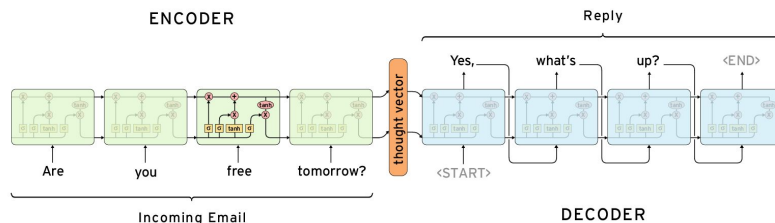
- Gensim
- NaiveBayesClassifier
- Term-existence



딥러닝  
(머신러닝)

## Seq2Seq

- TensorFlow
- English to French
- RNN, Attention, GRU, Bucketing,



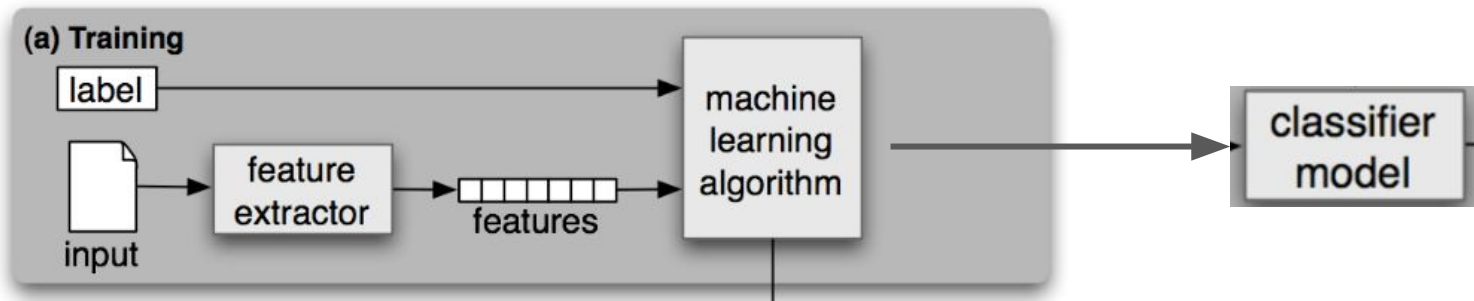
# 개발!

## Intend Inference - 학습

사진 검색	여름에 찍은 사진만 골라서 보여줘	SEARCH	30개
	어제 찍은 사진 뭐 있어?	SEARCH	
	꽃이 나온 사진만 보여줘	SEARCH	
	바다가 나온 사진	SEARCH	
	내가 나온 사진들만 보자	SEARCH	

사진 전송	철리에게 카카오톡으로 보내줘	SEND	30개
	소피한테 방금 분류한 사진들 전송	SEND	
	이 사진들 파이한테 보내줘	SEND	
	앞으로 이런 사진들 매주마다 닉한테 전송해줘	SEND	

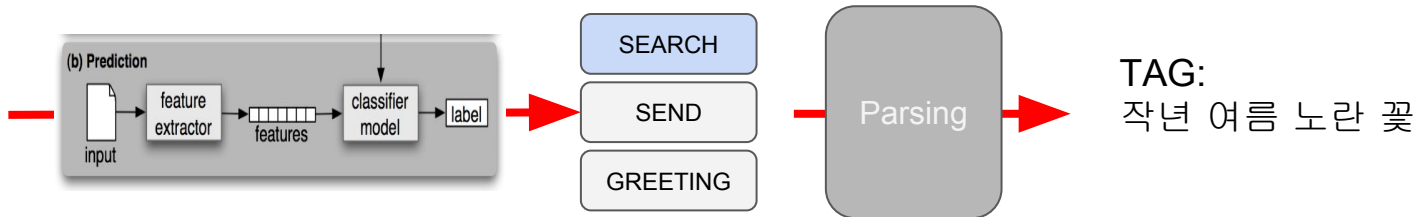
### (a) Training



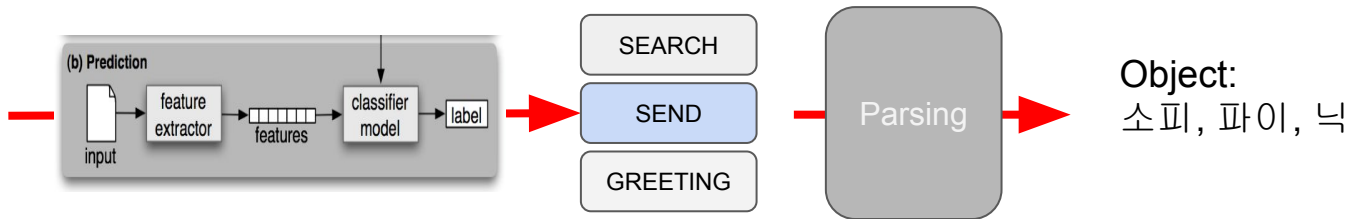
# 개발!

## Intend Inference - 예측

Q: “작년 여름에 찍은 노란  
꽃  
사진 보여줘”



Q: “소피하고 파이하고  
닉한테 보내줘”

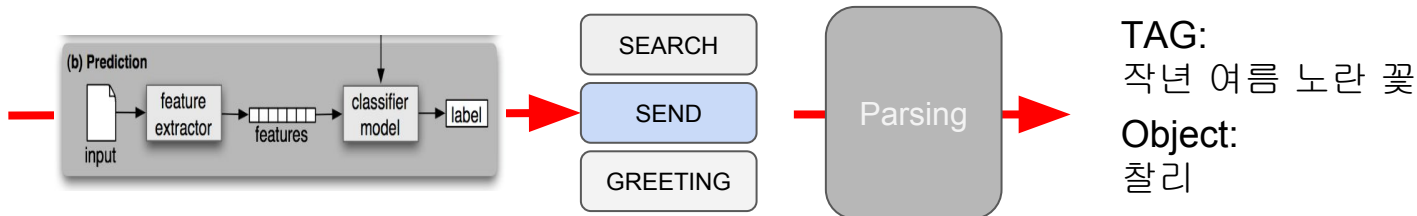




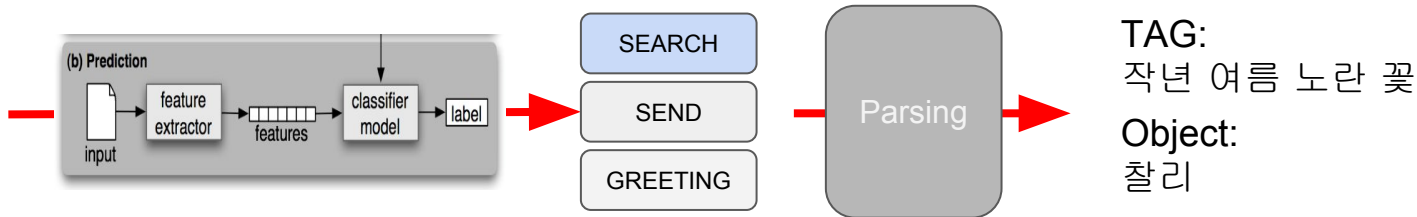
# 개발!

## Intend Inference - 예측(예외)

“작년 여름에 찍은 노란 꽃  
사진 찰리한테 보내줘”



“지난 겨울에 찍은 노란 꽃  
사진 찰리한테 전송해줘”

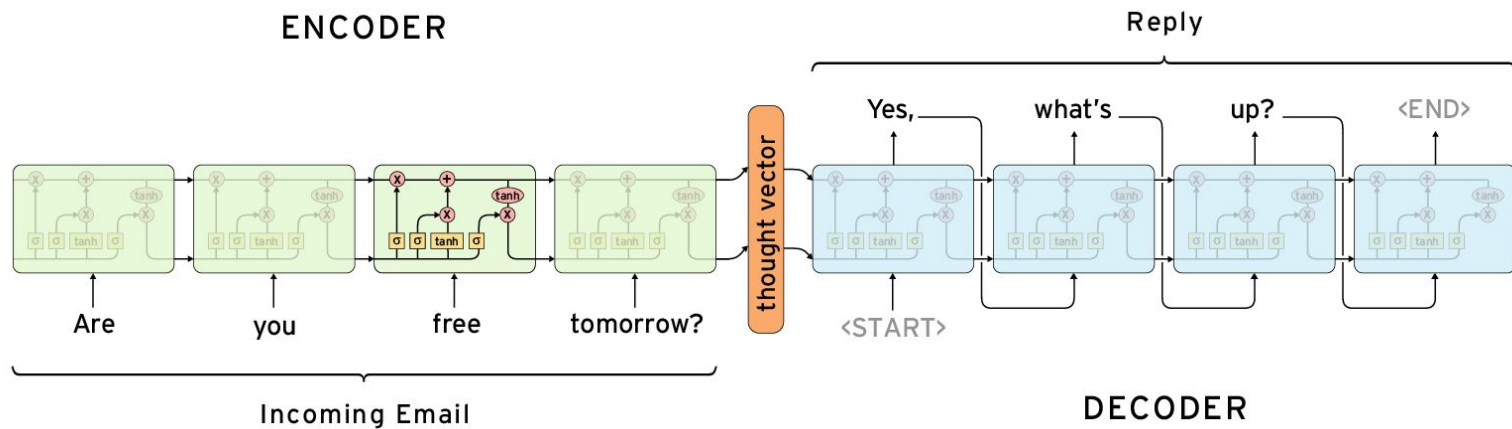


“안녕”



# 개발!

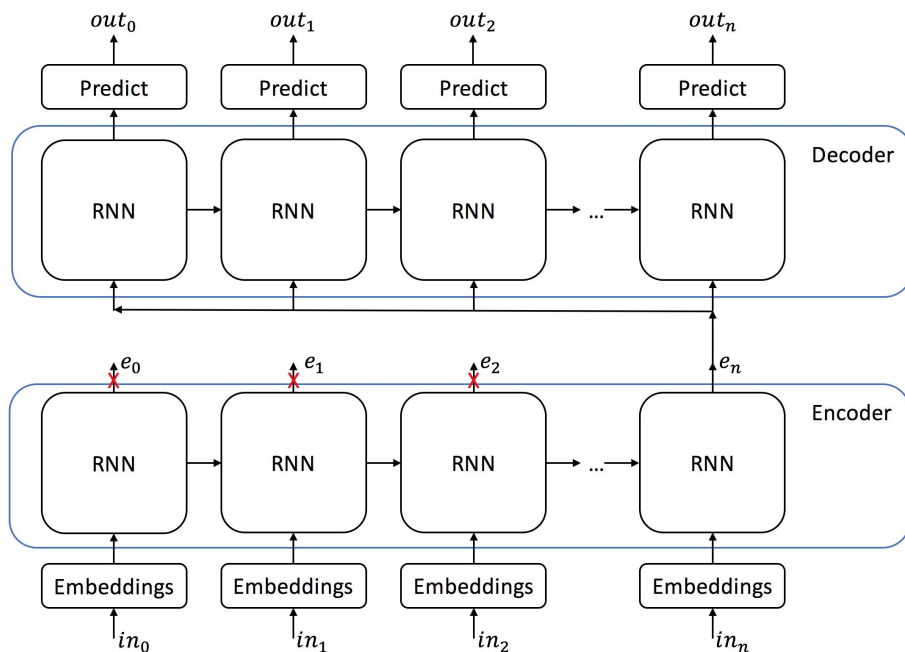
## Seq2Seq - 설명



# 개발!

## Seq2Seq - 설명

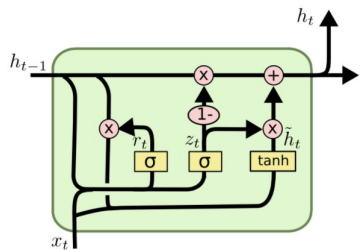
### With Attention:



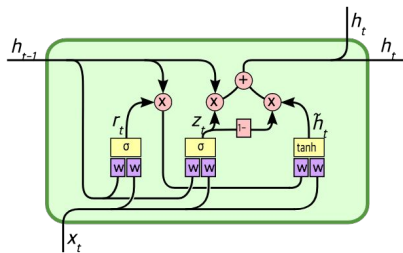
# 개발!

## Seq2Seq - 설명

Cell:



LSTM



GRU

Bucket:

$[(8, 10), (12, 14), (16, 19)]$

Sampled Softmax, Projection ... 저도 잘 이해가..

# 개발!

## Seq2Seq - 학습

일반적으로

- Closed Model에서도 최소 10000쌍 이상..

### Cornell Movie-Dialogs Corpus

- **220,579** conversational exchanges between
- **10,292** pairs of movie characters
- **9,035** characters from 617 movies
- **304,713** total utterances
- Very well-formatted (almost perfect)

### 3.2 데이터 수집

시나리오의 데이터를 수집하기 위해서 간단한 대화를 할 수 있는 모델을 만들 필요가 있었다. 이를 위해서 30명 이상의 사람으로부터 10,000여개의 대화 쌍을 수집하였다. 최대한 중복되는 대화 쌍을 줄이고, 나올 수 있는 대화를 다양하게 하기 위해 각 사람들은 6개의 세부 시나리오 중 하나에 할당되어 300쌍, 전체 시나리오에서 300쌍을 만들도록 했다. 각 대화 쌍은 아이가 먼저 질문이나 이야기를 꺼내고 모델 및 로봇이 그에 대해 대답하도록 구성되어 있다. 이 데이터를 가지고 기초 모델을 만든 후, 랩 투어, 학회 등에서 시연을 하며 기록된 로그 데이터를 통해 계속해서 새로운 데이터를 수집하였다.

현재는 우리는?

- 데이터가 없음.. 이상한 데이터 100쌍정도가 다임 ㅠㅠ

AI 란 무엇입니까?	인공 지능은 생각하는 기계를 만드는 데 전념하는 공학 및 과학의 한 분야입니다.			
어떤 언어를 쓰셨습니까?	파이썬			
너는 데이터처럼 들린다.	예, 나는 지휘관 데이터의 인격적 인성에 영감을 받았습니다.			
다시요 이고 어떤 식별이 나	그게 내 이름이야			

# 개발!

Seq2Seq - 예측

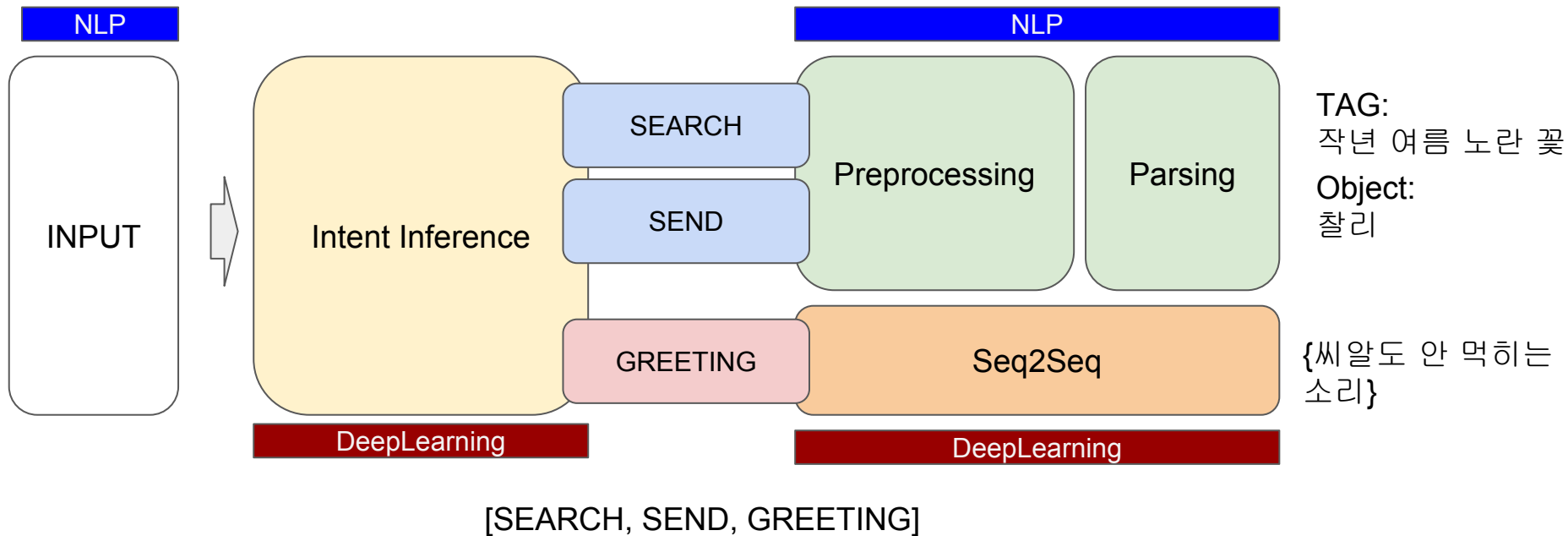


# 개발!

그 밖에도..

데이터 전처리/후처리, Word Embedding, Sentence Tokenizing 등등..

# 결과물!!

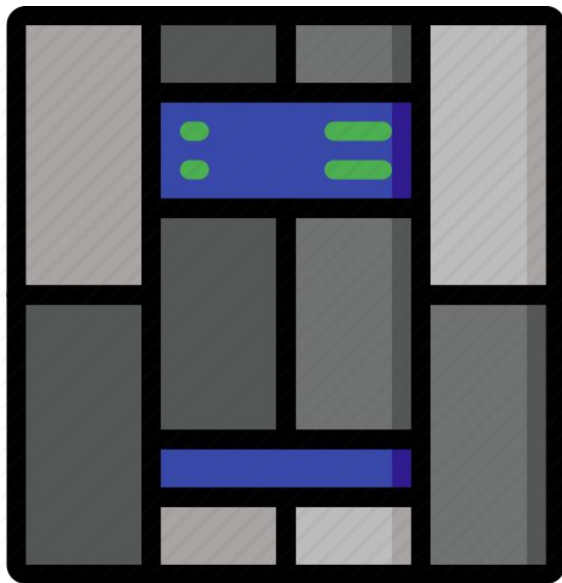




# 고찰!!

1. Deep Learning 기반의 무언가를 Gensim과 Tensorflow를 사용하여 구축 해본 것 만으로도 의미가 있었다. (신경망 모델도 (나름)최신으로!!)
2. Tensorflow는 많은 기능을 제공하는 장점이 있으나, Reference만 보고 이해하기는 힘들었다. 다른 코드 분석과 실험이 필요하다
3. 구현한 신경망 모델에 대한 이해도를 높이기 위해서는 더 많은 공부가 필요하다.
4. 사용할 만한 말뭉치(Corpus)를 구하거나 만들어야 한다.  
(국립국어원에서 제공하는 데이터로 테스트 해보았으니 우리의 챗봇 컨셉과는 맞지 않았다.)
5. Seq2Seq Model의 비중은 팀에서 논의가 필요  
(몇십만개 영어로 학습시켜도 완벽하진 않음. -> 완성도 ???, 기술성 및 트렌드 !!!)

데모!!



TARS: <http://13.124.109.145:5000/TARS/>  
(Text Analyze Recognition System)