2020

CAN YOU IDENTIFY
QUESTION PAIRS
THAT HAVE
THE SAME INTENTS?

Identifying similar questions using maLSTM

Contents

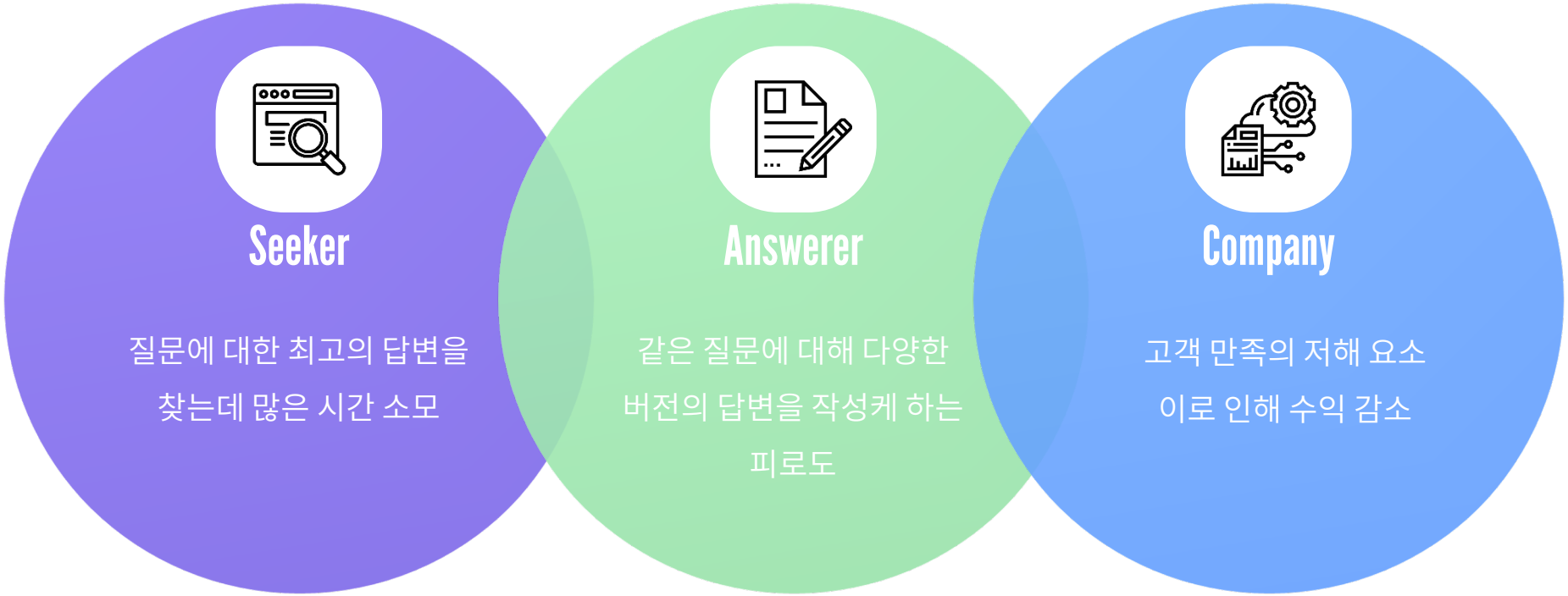
Introduction

Modeling

Result

Motivation

Quora는 2010년 Adam D’angelo가 설립한 웹사이트로 질문을 올리면 유저들이 답변을 해주는 방식으로 운영되는 웹 사이트이다.
기본적으로 실명제인만큼 답변의 전문성이 높은 경우가 많으며 버락 오바마, 힐러리 클린턴, 마크 주커버그 등 많은 유명인들도 사용 중에 있다.
그렇다면 Quora에서 중복된 질문을 찾는게 왜 중요할까?



“the best answer to any question”라는 슬로건의 달성과 함께 고객 만족 달성을 통한 사용자 편의 증진

Data

Train data

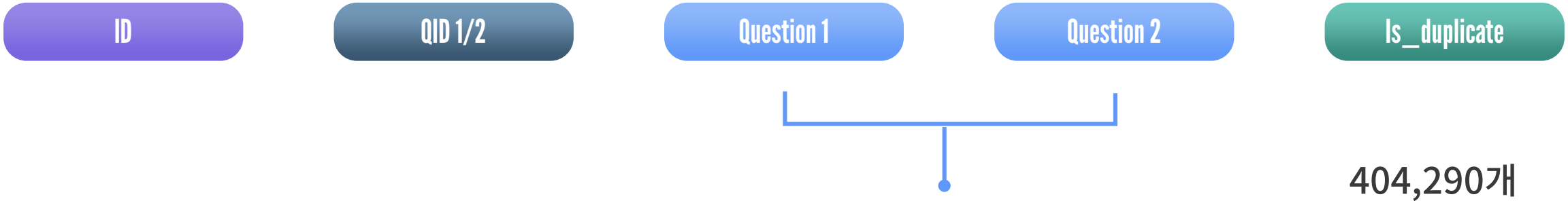
`qid1, qid2`

각 질문의 고유번호

`is_duplicate`

1 – question1&2 같은 의미

0 - question1&2 다른 의미



`question1, question2`

각 질문의 구체적인 내용

404,290개

example

Id	qid1	qid2	question1	question2	“target variable” is_duplicate
0	0	1	What is the step by step guide to invest in share market in india?	What is the step by step guide to invest in share market?	0
11	23	24	How do I read and find my YouTube comments?	How can I see all my Youtube comments?	1

Data

Train data

`qid1, qid2`

각 질문의 고유번호

`is_duplicate`

1 – question1&2 같은 의미

0 - question1&2 다른 의미

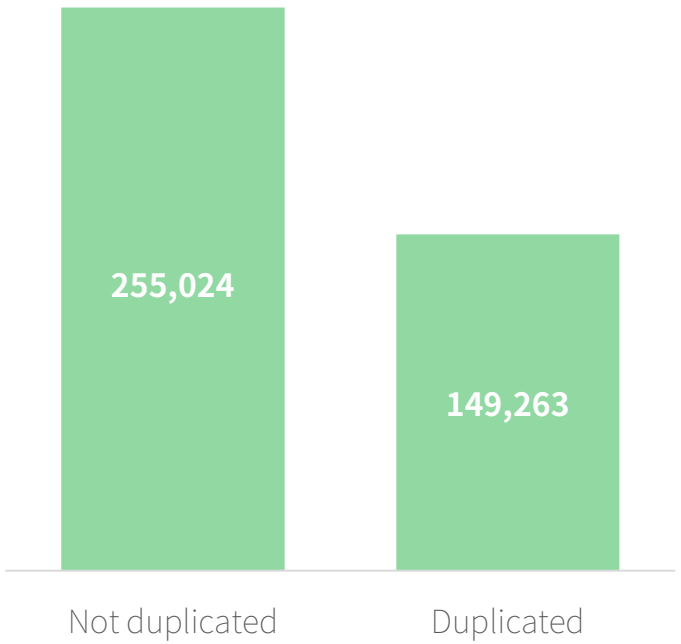
ID	QID 1/2	Question 1	Question 2	Is_duplicate

`question1, question2`

각 질문의 구체적인 내용

404,290개

데이터 확인



동일하지 않은 질문 / 전체 질문 :

255,024 / 404,287 ≒ **0.63**

"Accuracy 비교의 기준점 역할"

Data

Train data

`qid1, qid2`

각 질문의 고유번호

`is_duplicate`

1 – question1&2 같은 의미

0 - question1&2 다른 의미

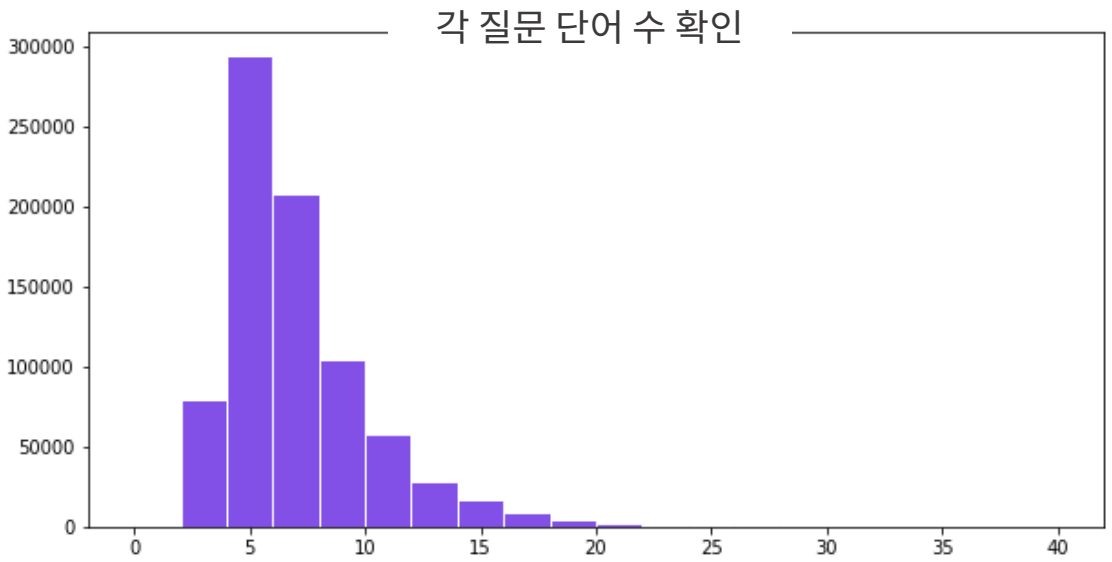
ID | QID 1/2 | Question 1 | Question 2 | Is_duplicate

404,290개

`question1, question2`

각 질문의 구체적인 내용

데이터 확인



Min length : 0

Max Length : 110

Mean Length : 6.57

Median Length : 6

Data

Train data

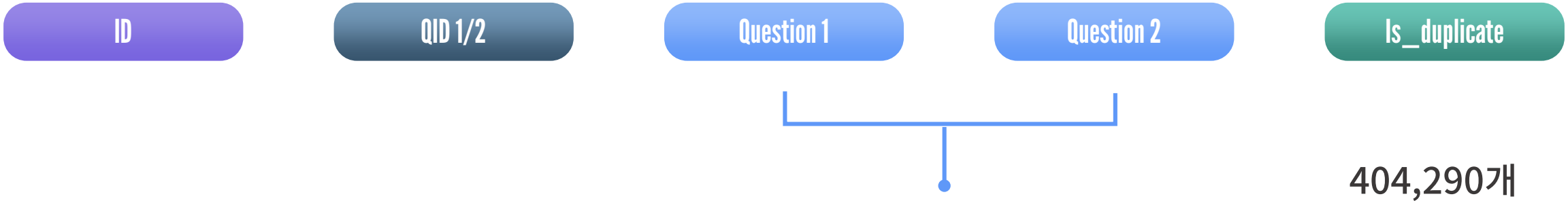
`qid1, qid2`

각 질문의 고유번호

`is_duplicate`

1 – question1&2 같은 의미

0 - question1&2 다른 의미



`question1, question2`

각 질문의 구체적인 내용

404,290개

example

Id	qid1	qid2	question1	question2	“target variable” is_duplicate
0	0	1	What is the step by step guide to invest in share market in india?	What is the step by step guide to invest in share market?	0
11	23	24	How do I read and find my YouTube comments?	How can I see all my Youtube comments?	1

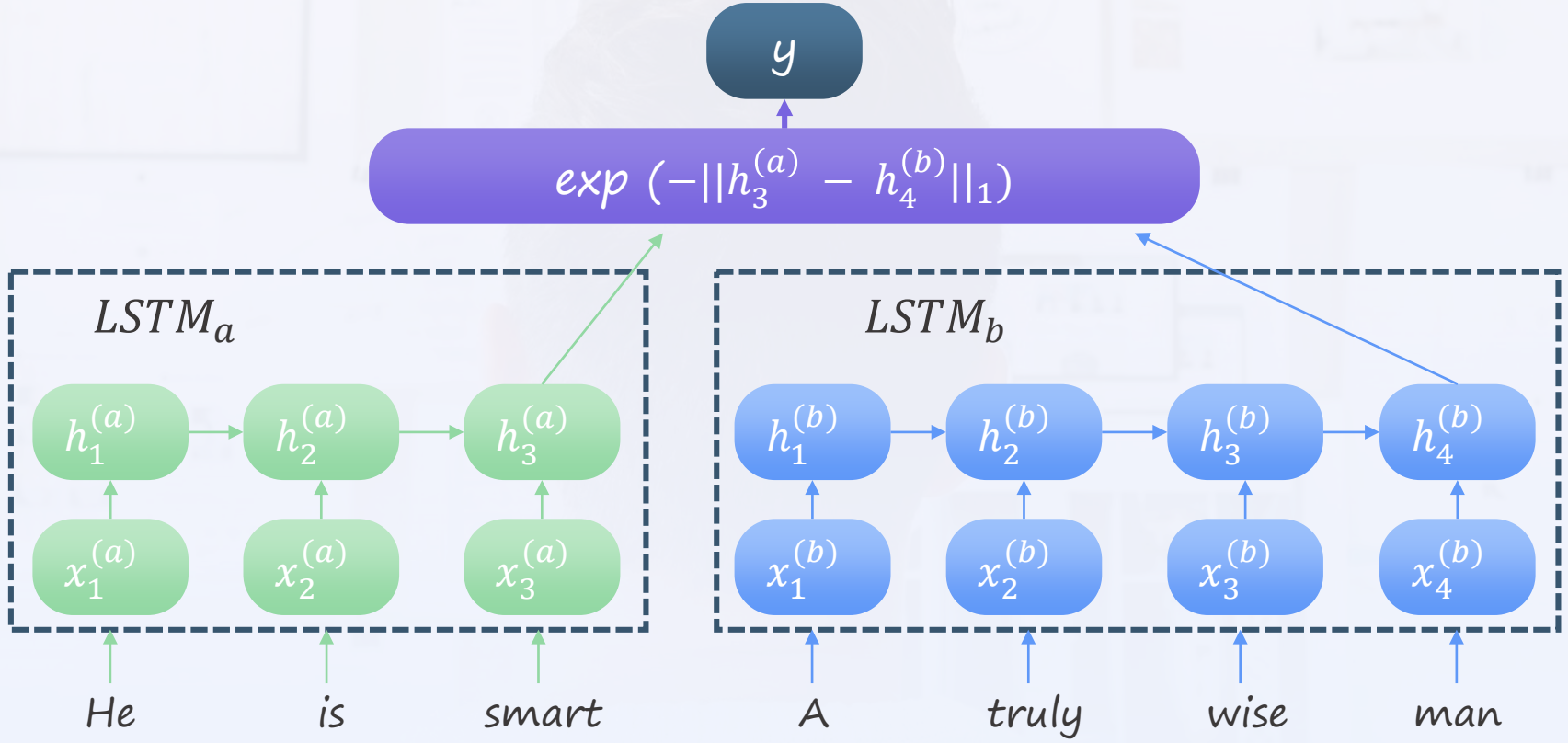
Introduction

Modeling

Result

MaLSTM

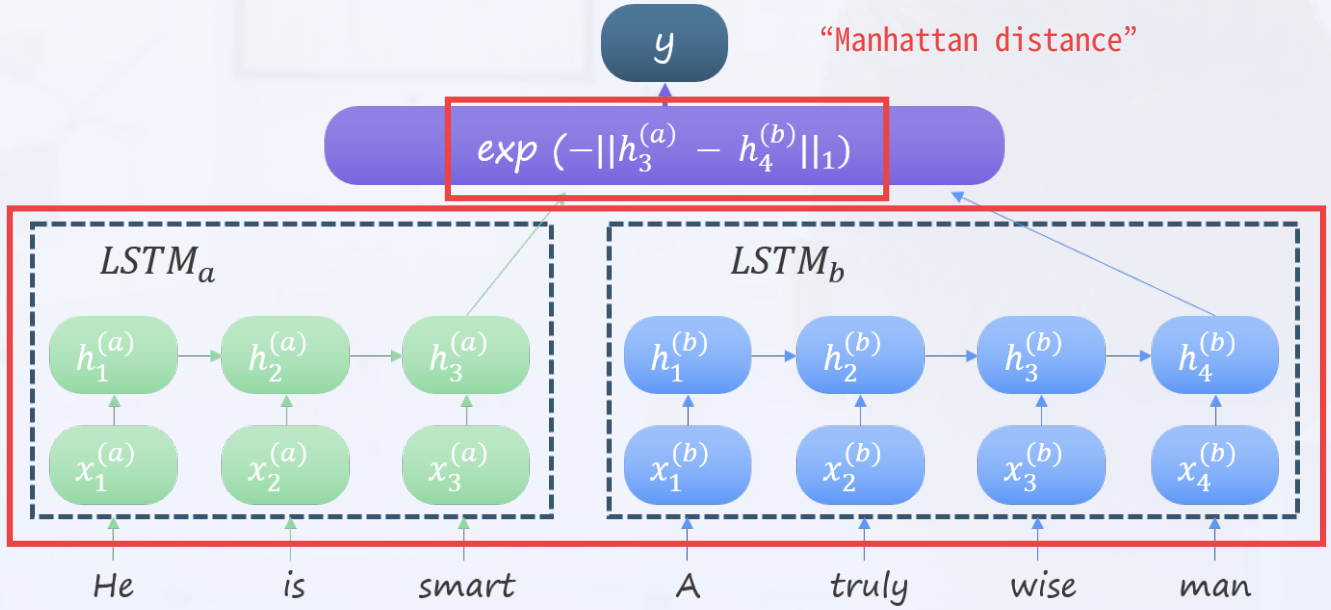
Siamese networks + LSTM with Manhattan distance



*출처 : Siamese Recurrent Architectures for Learning Sentence Similarity

MaLSTM

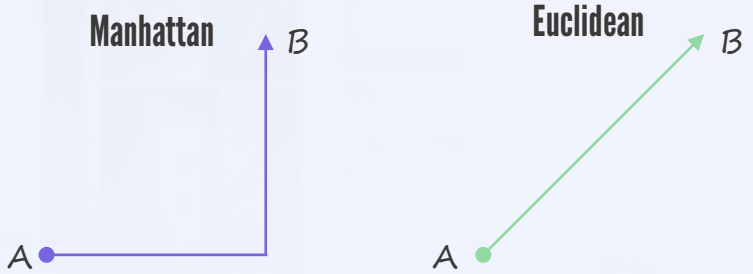
Siamese networks + LSTM with Manhattan distance



- input(question1 / question2)이 두 개로 구성
  Siamese형태를 가진 두 개의 LSTM model을 사용

- 유사도 측정에는 Manhattan distance를 이용

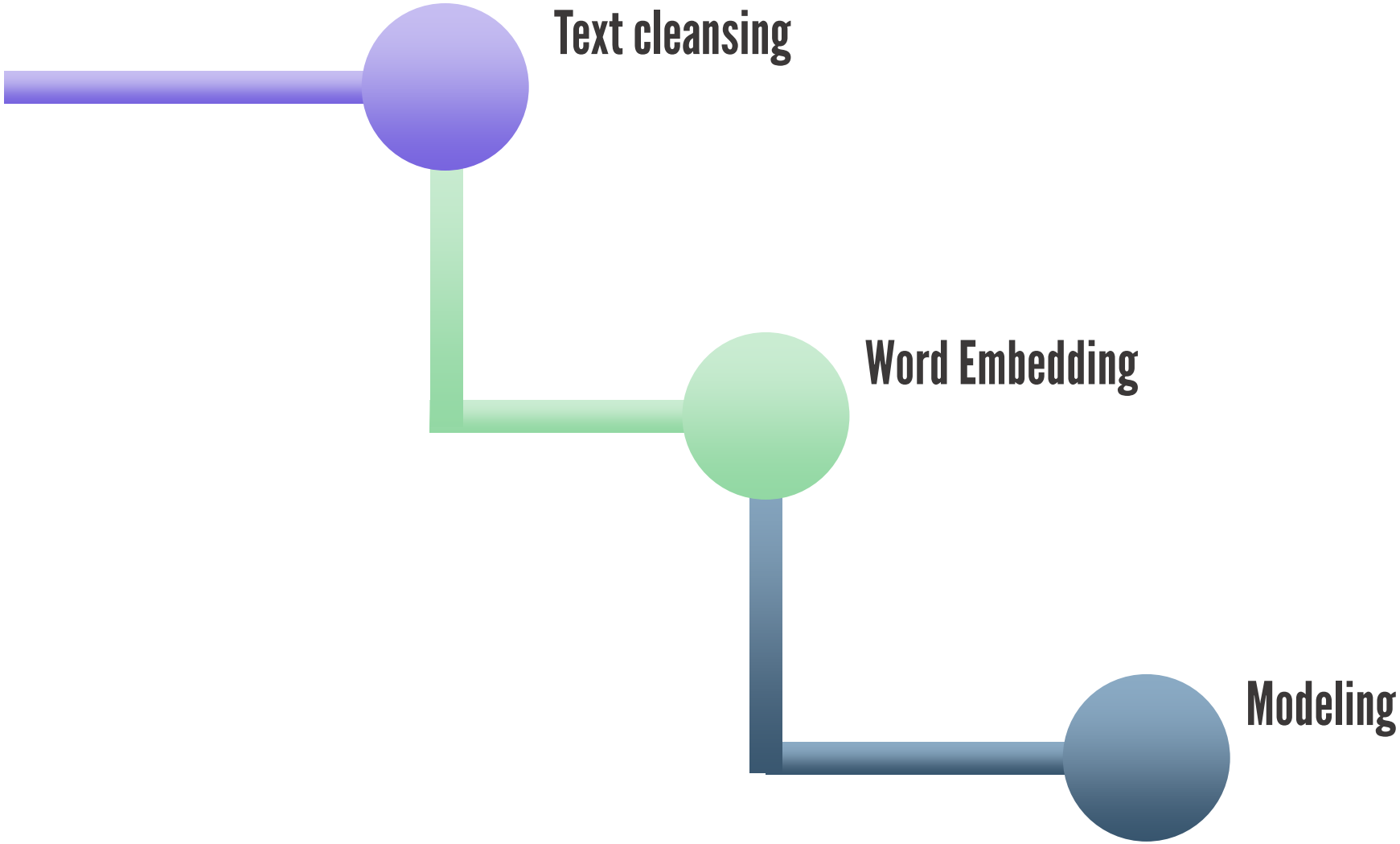
$$L_1 = \sum_{i=1}^{n} |a_i - b_i|$$



*출처 : Siamese Recurrent Architectures for Learning Sentence Similarity

Modeling

step by step



Modeling

step by step

Text cleansing

Word Embedding

Modeling

데이터 전처리

(1) Text cleansing & Normalization

특수문자 제거, apostrophe(‘) 제거 (축약어를 원형으로)

(2) Tokenization

text_to_word_sequence를 이용

(3) Stopwords 제거

nltk stopwords 영어 불용어 리스트를 이용

*“no, not, nor”은 제거 X – 부정어 중요할 가능성 존재

*”불용어 처리시 의문사가 제거되는 문제가 발생하여 데이터를 의문사 제거 / 의문사 제거 X 로 나눠서 저장

(4) Lemmatisation

WordNetLemmatizer이용

Modeling

step by step

Text cleansing

Word Embedding

Modeling

Word Embedding

(1) Word2Vec

1) my_data를 통한 학습 (300 차원)

2) 교차 학습 : my_data + *GoogleNews-vector-negative300

(2) GloVe

1) my_data를 통한 학습 (300 차원)

2) Wikipedia 2014 + Gigaword 5를 통한 pre-trained (300 차원)

Train-Test Split 실시

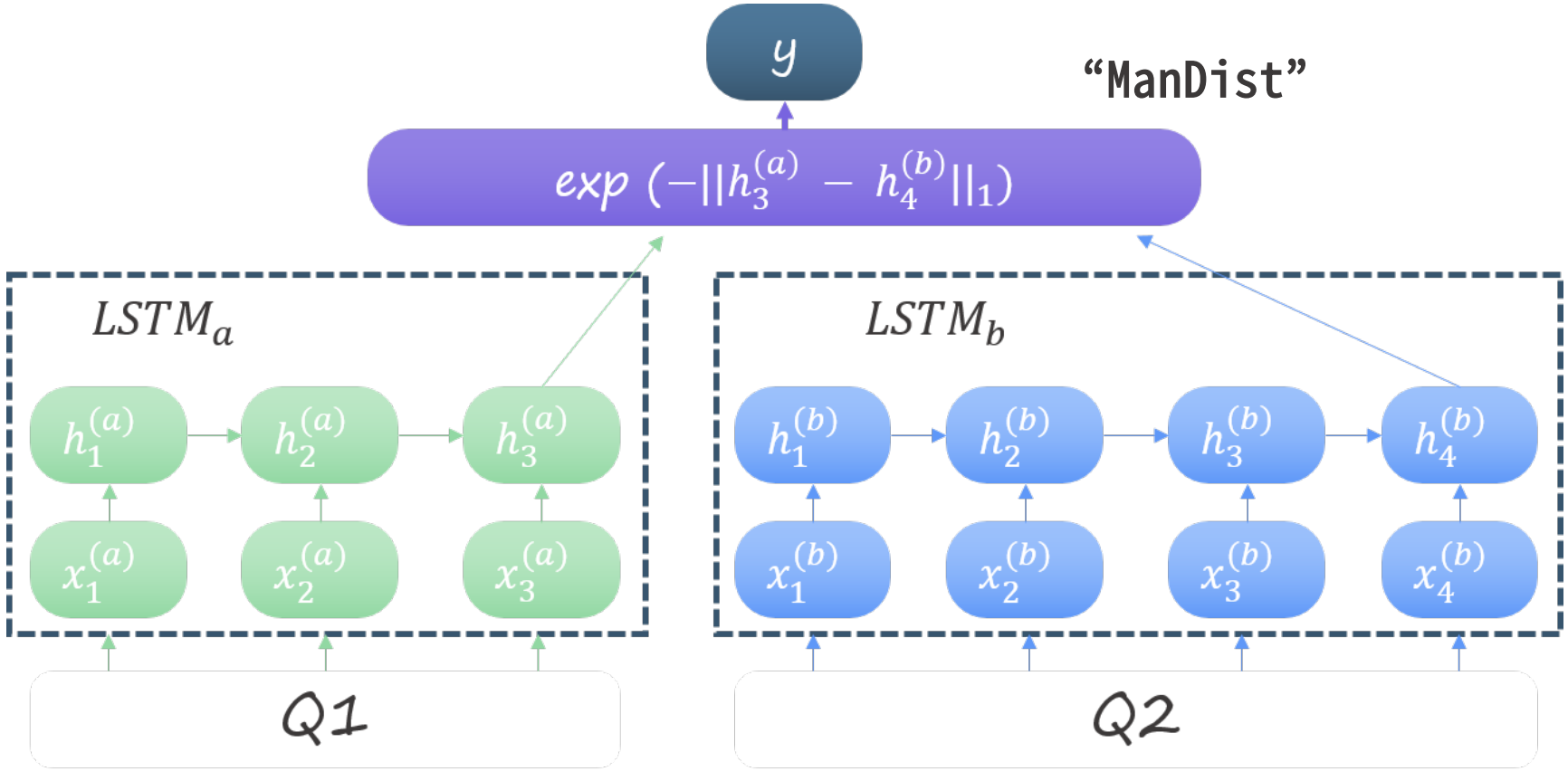
Train : Test = 0.7 : 0.3

*GoogleNews-vector-negative300 : 1,000억 단어 규모의 구글 뉴스 데이터를 이용하여 300만 개의 단어를 300차원 벡터로 embedding

Modeling

step by step

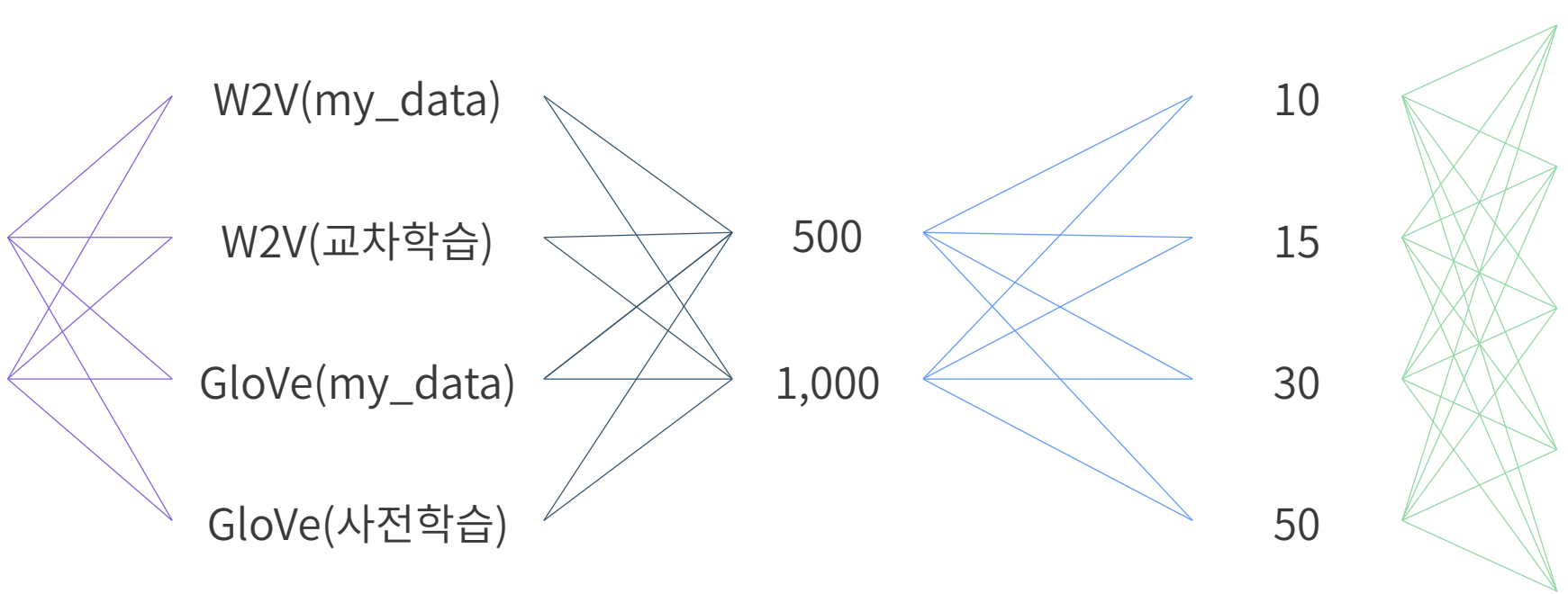
- Text cleansing
- Word Embedding
- **Modeling**



Modeling

step by step

의문사 사용 여부	Embedding	Batch size	Hidden layer	Optimizer
True	W2V(my_data)	500	10	Adadelta
False	W2V(교차학습)	1,000	15	Adam
	GloVe(my_data)		30	Momentum
	GloVe(사전학습)		50	RAdam
				SGD



2 X 4 X 2 X 4 X 5

" 320 "

*bias initialization = 2.0

Introduction

Modeling

Result

Result

Siamese LSTM과 두 개의 독립된 LSTM

[Data parameter settings]
Maximum length of questions:
20 (w2v), 30 (Glove)

[Model parameter settings]
Dimension of hidden layer: 50
Initial value of bias: 2.0
Optimizer: Adam
Batch size: 500
Number of epochs: 50

	학습	의문사 사용	Precision		Recall		F1 score		Accuracy		Loss	
			siamese	each	siamese	each	siamese	each	siamese	each	siamese	each
W2V	pretrained	O	**0.7770**	**0.7500**	**0.6862**	**0.6306**	**0.7288**	**0.6851**	**0.8113**	**0.7859**	**0.4882**	**0.4554**
		X	0.7619	0.7476	0.6068	0.6308	0.7333	0.6843	0.8101	0.7850	0.5518	0.4562
	my_data	O	0.7592	0.7444	0.6779	0.6211	0.7163	0.6772	0.8016	0.7812	0.4992	0.4671
		X	0.7602	0.7273	0.6797	0.6395	0.7177	0.6805	0.8025	0.7782	0.5605	0.4680
GloVe	pretrained	O	0.7479	0.7318	0.7199	0.6450	0.7336	0.6857	0.8069	0.7815	0.5664	0.4696
		X	0.7534	0.7467	0.7107	0.6210	0.7314	0.6781	0.8072	0.7822	0.6369	0.471
	my_data	O	0.7359	0.6671	0.5963	0.54	0.6588	0.5969	0.7718	0.7305	0.5483	0.5353
		X	0.7243	0.6313	0.5956	0.5646	0.6536	0.5961	0.7668	0.7173	0.6198	0.5554

모든 측면에서 Siamese LSTM이 우수

Result

사전훈련과 Quora데이터로 훈련된 단어 임베딩

[Data parameter settings]
Maximum length of questions:
20 (w2v), 30 (Glove)

[Model parameter settings]
Dimension of hidden layer: 50
Initial value of bias: 2.0
Optimizer: Adam
Batch size: 500
Number of epochs: 50

	학습	의문사 사용	Precision		Recall		F1 score		Accuracy		Loss	
			siamese	each	siamese	each	siamese	each	siamese	each	siamese	each
W2V	pretrained	O	**0.7770**	**0.7500**	**0.6862**	**0.6306**	**0.7288**	**0.6851**	**0.8113**	**0.7859**	**0.4882**	**0.4554**
		X	0.7619	0.7476	0.6068	0.6308	0.7333	0.6843	0.8101	0.7850	0.5518	0.4562
	my_data	O	**0.7592**	**0.7444**	**0.6779**	**0.6211**	**0.7163**	**0.6772**	**0.8016**	**0.7812**	**0.4992**	**0.4671**
		X	0.7602	0.7273	0.6797	0.6395	0.7177	0.6805	0.8025	0.7782	0.5605	0.4680
GloVe	pretrained	O	0.7479	0.7318	0.7199	0.6450	0.7336	0.6857	0.8069	0.7815	0.5664	0.4696
		X	0.7534	0.7467	0.7107	0.6210	0.7314	0.6781	0.8072	0.7822	0.6369	0.471
	my_data	O	0.7359	0.6671	0.5963	0.54	0.6588	0.5969	0.7718	0.7305	0.5483	0.5353
		X	0.7243	0.6313	0.5956	0.5646	0.6536	0.5961	0.7668	0.7173	0.6198	0.5554

모든 측면에서 Siamese LSTM이 우수

데이터가 40만여 쌍, 그러나 W2V와 GloVe 모두사전 훈련된 단어 임베딩이 우수

Result

의문사 포함 여부

[Data parameter settings]
Maximum length of questions:
20 (w2v), 30 (Glove)

[Model parameter settings]
Dimension of hidden layer: 50
Initial value of bias: 2.0
Optimizer: Adam
Batch size: 500
Number of epochs: 50

	학습	의문사 사용	Precision		Recall		F1 score		Accuracy		Loss	
			siamese	each	siamese	each	siamese	each	siamese	each	siamese	each
W2V	pretrained	O	**0.7770**	**0.7500**	**0.6862**	**0.6306**	**0.7288**	**0.6851**	**0.8113**	**0.7859**	**0.4882**	**0.4554**
		X	**0.7619**	**0.7476**	**0.6068**	**0.6308**	**0.7333**	**0.6843**	**0.8101**	**0.7850**	**0.5518**	**0.4562**
	my_data	O	0.7592	0.7444	0.6779	0.6211	0.7163	0.6772	0.8016	0.7812	0.4992	0.4671
		X	0.7602	0.7273	0.6797	0.6395	0.7177	0.6805	0.8025	0.7782	0.5605	0.4680
GloVe	pretrained	O	0.7479	0.7318	0.7199	0.6450	0.7336	0.6857	0.8069	0.7815	0.5664	0.4696
		X	0.7534	0.7467	0.7107	0.6210	0.7314	0.6781	0.8072	0.7822	0.6369	0.471
	my_data	O	0.7359	0.6671	0.5963	0.54	0.6588	0.5969	0.7718	0.7305	0.5483	0.5353
		X	0.7243	0.6313	0.5956	0.5646	0.6536	0.5961	0.7668	0.7173	0.6198	0.5554

모든 측면에서 Siamese LSTM이 우수

데이터가 40만여 쌍, 그러나 W2V와 GloVe 모두사전 훈련된 단어 임베딩이 우수

사전훈련된 단어 임베딩 사용시 의문사 포함된 데이터가 우수

Result

Word2Vec과 GloVe

[Data parameter settings]
Maximum length of questions:
20 (w2v), 30 (Glove)

[Model parameter settings]
Dimension of hidden layer: 50
Initial value of bias: 2.0
Optimizer: Adam
Batch size: 500
Number of epochs: 50

	의문사 사용	train_acc	train_loss	val_acc	val_loss
W2V	O	0.8391	0.4362	0.8097	**0.4869**
	X	0.8379	0.4984	0.8098	**0.5572**
GloVe	O	0.8497	0.4795	0.8100	**0.5580**
	X	0.8494	0.5545	0.8063	**0.6446**

Word2Vec이 대체로 우수, GloVe의 높은 손실

Result

3. Optimizer에 따른 성능 비교

[Data parameter settings]
Maximum length of questions: 20 (w2v)
wh = True

[Model parameter settings]
Dimension of hidden layer: 50
Initial value of bias: 2.0
Batch size: 500
Number of epochs: 50

	Precision	Recall	F1 score	Accuracy
Adadelta	**0.598976**	**0.446669**	**0.511730**	**0.685116**
Adam	**0.775501**	**0.696730**	**0.734008**	**0.813457**
SGD	0.775094	0.696953	0.733950	0.813342
Momentum	0.774026	0.692244	0.730854	0.811652
RAdam	0.766019	0.705189	0.734347	0.811520

대부분의 데이터에서 비슷한 경향을 보임

(모두 그런 것은 아니지만 Adam을 사용했을 때 가장 성능이 좋고, Adadelta를 사용했을 때 가장 성능이 떨어짐)

Adadelta 정확도 68-69 구간에서 정체

Result

3. Optimizer에 따른 성능 비교

[Data parameter settings]

Maximum length of questions: 20 (w2v)

wh = True

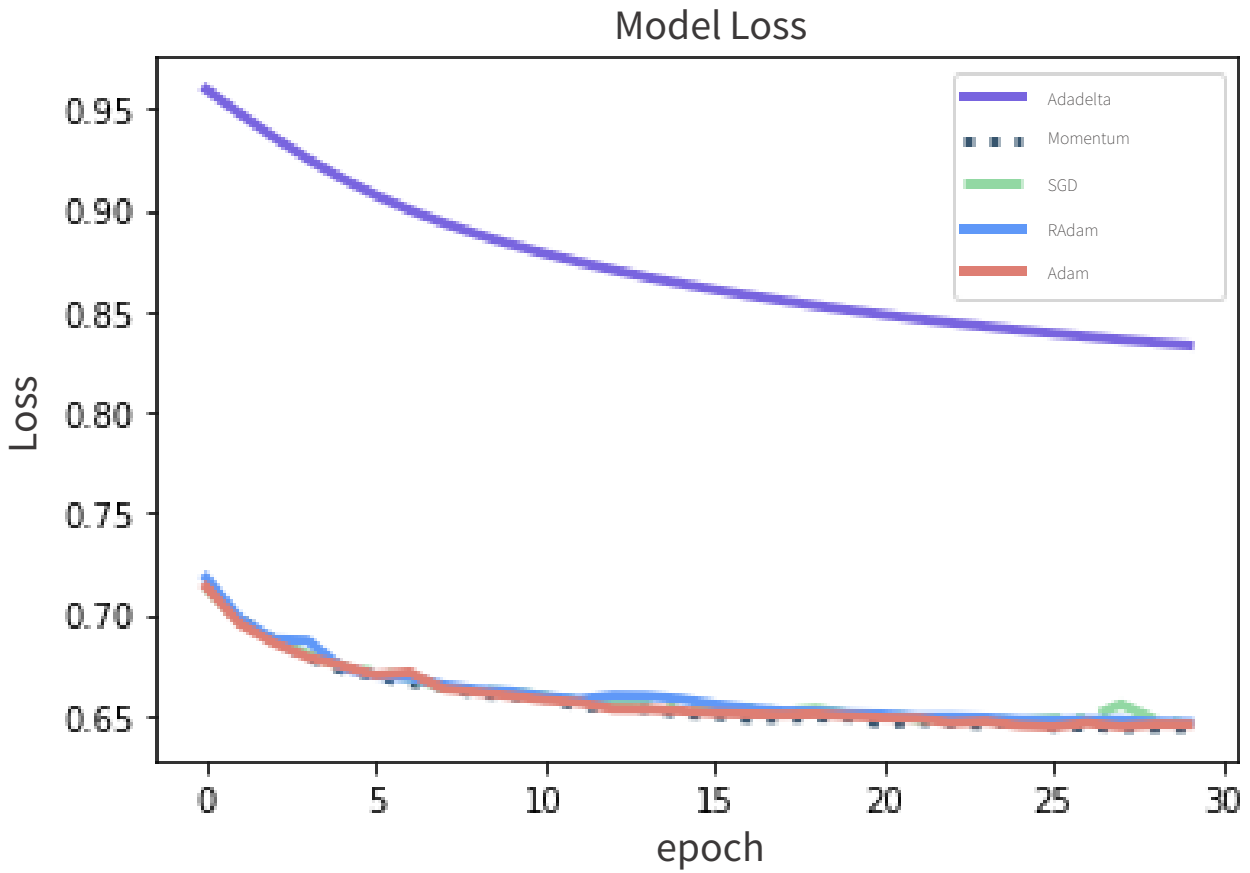
[Model parameter settings]

Dimension of hidden layer: 50

Initial value of bias: 2.0

Batch size: 500

Number of epochs: 50



대부분의 데이터에서 비슷한 경향을 보임

(모두 그런 것은 아니지만 Adam을 사용했을 때 가장 성능이 좋고, Adadelta를 사용했을 때 가장 성능이 떨어짐)

Adadelta 정확도 68-69 구간에서 정체

Result

4. 지표별 Highest model

“Highest Precision”

[Data parameter settings]
Maximum length of questions: 20 (w2v)
wh = True
[Model parameter settings]
Opt: Adam
Dimension of hidden layer: 100
Initial value of bias: 2.0
Batch size: 500
Number of epochs: 50

MaLSTM with Dense layer(1) output dimension: 10

	Precision	Recall	F1 score	Accuracy	Loss
Test	**0.8117**	0.6772	0.7384	0.8227	0.4869

“Lowest loss”

[Data parameter settings]
Maximum length of questions: 20 (w2v)
wh = True
[Model parameter settings]
Opt: Adam
Dimension of hidden layer: 50
Initial value of bias: 2.0
Batch size: 500
Number of epochs: 50

독립된 LSTM

	Precision	Recall	F1 score	Accuracy	Loss
Test	0.7500	0.6301	0.6851	0.7859	**0.4554**

Result

4. 지표별 Highest model

“Highest F1-score”

[Data parameter settings]
Maximum length of questions: 30 (w2v)
wh = False
[Model parameter settings]
Opt: Adam
Dimension of hidden layer: 100
Initial value of bias: 2.0
Batch size: 1000
Number of epochs: 50

MaLSTM with Dense layer(1) output dimension: 30

	Precision	Recall	F1 score	Accuracy	Loss
Test	0.808028	0.810209	**0.807108**	0.810209	0.5844

“Highest Accuracy”

[Data parameter settings]
Maximum length of questions: 20 (w2v)
wh = False
[Model parameter settings]
Opt: Adam
Dimension of hidden layer: 15
Initial value of bias: 2.0
Batch size: 500
Number of epochs: 50

MaLSTM with Dense layer(2) output dimension: 100

	Precision	Recall	F1 score	Accuracy	Loss
Test	0.799041	0.71376	0.753996	**0.8279**	0.5151

Result

+α layer 추가하기

[Data parameter settings]
Maximum length of questions: 20 (w2v)
wh = False

[Model parameter settings]
Dimension of hidden layer: 50
Initial value of bias: 2.0
Batch size: 500
Number of epochs: 15

	X	1-layer	3-layer
train Accuracy	0.8044	0.8166	0.8239
validation Accuracy	**0.7948**	**0.7976**	**0.7993**

대부분의 경우에 레이어를 더 쌓으면 모델은 복잡해지지만 성능이 개선됨

[Data parameter settings]
Maximum length of questions: 30 (GloVe)
wh = False

[Model parameter settings]
Dimension of hidden layer: 50
Initial value of bias: 2.0
Batch size: 500
Number of epochs: 50

	X	2-layer
train Accuracy	0.8405	0.9023
train Loss	0.5696	0.4607
validation Accuracy	0.8046	0.8013
validation loss	0.6431	0.6833

그러나 항상 개선되는 것은 아니고 over-fitting의 우려 존재

Thank You

2020 CAN YOU IDENTIFY QUESTION PAIRS THAT HAVE THE SAME INTENTS?

QnA

참고 문헌

GloVe 자료 :

https://github.com/stanfordnlp/GloVe

MaLSTM 자료 :

Mueller, J., and Thyagarajan, A . 2016, Siamese Recurrent Architectures for Learning Sentence Similarity, AAAI-16

Google 사전학습 W2V :

https://code.google.com/archive/p/word2vec/

Siamese Network 자료 :

http://yann.lecun.com/exdb/publis/pdf/chopra-05.pdf

Quora(kaggle data set) :

https://www.kaggle.com/c/quora-question-pairs

RAdam :

S. Chopra and R. Hadsell and Y. LeCun, 2005 Learning a Similarity Metric Discriminatively, with Application to Face, CVPR'05

Siamese-LSTM 자료 :

https://github.com/likejazz/Siamese-LSTM