Seq2seq with attention
한글-영어 기계번역

Neural Machine Translation

빅데이터 경영통계전공 20172865 홍지원

Contents

데이터 준비

모델의 성능을 높이기 위해 데이터 추가 수집

AI Hub에서 한글-영어 구어체 말뭉치 데이터 20만 개 중 5만개 사용

원문의 순서가 내림차순으로 되어 있기 때문에 데이터셋 랜덤 추출

	A	B	C
1	SID	원문	번역문
2	1	'Bible Coloring'은 성경의 아름다운 이야기를 체험 할 수 있는 컬러링 앱입니다.	Bible Coloring' is a coloring application that allows you to experience beautiful stories in the Bible.
3	2	씨티은행에서 일하세요?	Do you work at a City bank?
4	3	푸리토의 베스트셀러는 해외에서 입소문만으로 4차 완판을 기록하였다.	PURITO's bestseller, which recorded 4th rough -cuts by words of mouth from abroad.
5	4	11장에서는 예수님이 이번엔 나사로를 무덤에서 불러내어 죽은 자 가운데서 살리셨습니다.	In Chapter 11 Jesus called Lazarus from the tomb and raised him from the dead.
6	5	6.5, 7, 8 사이즈가 몇 개나 더 재입고 될지 제게 알려주시면 감사하겠습니다.	I would feel grateful to know how many stocks will be secured of size 6.5, 7, and 8.
7	6	F/W 겐조타이거 키즈와 그리고 이번에 주문한 키즈 중 부족한 수량에 대한 환불입니다.	18fw Kenzo Tiger Kids, and refund for lacking quantity of Kids which was ordered this time.
8	7	강아지들과 내 사진을 보낼게.	And I'll send you a picture of me and my dogs.
9	8	그 수익금 중 일부를 위안부 할머니들을 위해 쓰고 그들을 위해 여러 가지 캠페인을 벌이고 있습니다.	Part of profits are used for the comfort women, and it is holding various campaigns for them.
10	9	그들은 내가 잘하는 것을 바탕으로 별명을 사용하고 있기 때문에 나는 사람들이 치타라고 불러주면 기분이 좋아.	I feel happy when people call me cheetah because they are using a nickname based on something that I am good at.
11	10	그러므로 실제로 컴퓨터 프로그램을 만든 사람이 프로그램에 대한 저작자가 돼요.	So, a person who made a computer program actually becomes an author of that computer program.
12	11	나는 친구에게 그 철학자의 책을 선물해 주겠다고 말했습니다.	I told my friends that I will give you the philosopher's book as a gift.
13	12	나머지 사진은 내 친구들이야.	And the rest of the pictures are my friends.
14	13	나머지 시간에는 공부해요.	I study for the rest of the time.
15	14	네가 하는 일과 공부 잘하길 멀리서 응원할게.	I will cheer you on your work and your grade from far away.
16	15	다른 선수들이 몬스터를 사냥할 경우 당신은 추가 경험치를 획득해요.	If other players hunt monsters, you gain additional experience.
17	16	당신에게 영화관 티켓을 그냥 보여 주면 되나요?	Can I just show you my ticket to the movie theater?
18	17	마치 목욕탕 창구처럼 보일까말까 한 작은 구멍으로 내가 돈을 주면 그 여자가 교통카드를 충전시켜주었던 기억이 납니다.	I remember that the person recharged my transportation card when I gave her money through a tiny hole, just like the ticket office for the public bath.

데이터 준비

모델의 성능을 높이기 위해 데

AI Hub에서 한글-영어 구어체

원문의 순서가 내림차순으로

```
def preprocess_sentence(w):
    # creating a space between a word and the punctuation following it
    # eg: "he is a boy." => "he is a boy ."
    # Reference:- https://stackoverflow.com/questions/3645931/python-padding-pu
    try:
      w = w.lower().strip()
      w = re.sub(r"([?.!,¿])", r" ₩1 ", w)
      w = re.sub(r'[" "]+', " ", w)

      # replacing everything with space except (a-z, A-Z, ".", "?", "!", ",")
      w = re.sub(r"[^a-zA-Z0-9가-힣?.!,¿]+", " ", w)

      w = w.strip()

      # adding a start and an end token to the sentence
      # so that the model know when to start and stop predicting.
      w = '<start> ' + w + ' <end>'
    except:
      w = re.sub(r"([?.!,¿])", r" ₩1 ", w)
      w = re.sub(r'[" "]+', " ", w)
```

		C
		번역문
		Bible Coloring' is a coloring application that allows you to experience beautiful stories in the Bible.
		Do you work at a City bank?
		PURITO's bestseller, which recorded 4th rough -cuts by words of mouth from abroad.
		In Chapter 11 Jesus called Lazarus from the tomb and raised him from the dead.
		I would feel grateful to know how many stocks will be secured of size 6.5, 7, and 8.
		18fw Kenzo Tiger Kids, and refund for lacking quantity of Kids which was ordered this time.
		And I'll send you a picture of me and my dogs.
		Part of profits are used for the comfort women, and it is holding various campaigns for them.
		I feel happy when people call me cheetah because they are using a nickname based on something that I am good at.
		So, a person who made a computer program actually becomes an author of that computer program.
		I told my friends that I will give you the philosopher's book as a gift.
		And the rest of the pictures are my friends.
		I study for the rest of the time.
		I will cheer you on your work and your grade from far away.
		If other players hunt monsters, you gain additional experience.
17	16	당신에게 영화관 티켓을 그냥 보여 주면 되나요?	Can I just show you my ticket to the movie theater?
18	17	마치 목욕탕 창구처럼 보일까말까 한 작은 구멍으로 내가 돈을 주면 그 여자가 교통카드를 충전시켜주었던 기억이 납니다.	I remember that the person recharged my transportation card when I gave her money through a tiny hole, just like the ticket office for the public bath.

Subword Tokenizer

에이미는 남자 친구가 찍어 준

사진을 나에게 보여 주었습니다.

에이미 V 는 V 남자 V 친구 V 가 V 찍어 V 준 V

사진 V 을 V 나에게 V 보여 V 주었 V 습니다 V .

신조어 같은 새로운 단어가 문장에 등장해도 유연하게 대처하기 때문에
SentencePiece를 Subword tokenizer로 선택

SentencePiece

Vocab size

vocab size가 클수록 불필요한 단어 생성,
작을수록 음절 단위로 분리

20,000 단어

Model Type

bpe : 완성되지 않은 문장이 많음
unigram : 문장이 비교적 매끄러움

Unigram(Defult)

Piece 정의

User symbol : <start>, <end>
전처리 과정에서 생긴 토큰 정의

```
input_tensor
```

```
array([[   2,  112,   59, ...,    0,    0,    0],
       [   2,   84, 1325, ...,    0,    0,    0],
       [   2,    8,   53, ...,    0,    0,    0],
       ...,
       [   2,   12,   99, ...,    0,    0,    0],
       [   2,   65,  402, ...,    0,    0,    0],
       [   2,   42,   12, ...,    0,    0,    0]], dtype=int32)
```

```
--pad_id=0 --pad_piece=<pad> --unk_id=1 --unk_piece=<unk> --bos_piece=<start> --bos_id=2 --eos_piece=<end> --eos_id=3 --user_defined_symbols=<start>,<end>
```

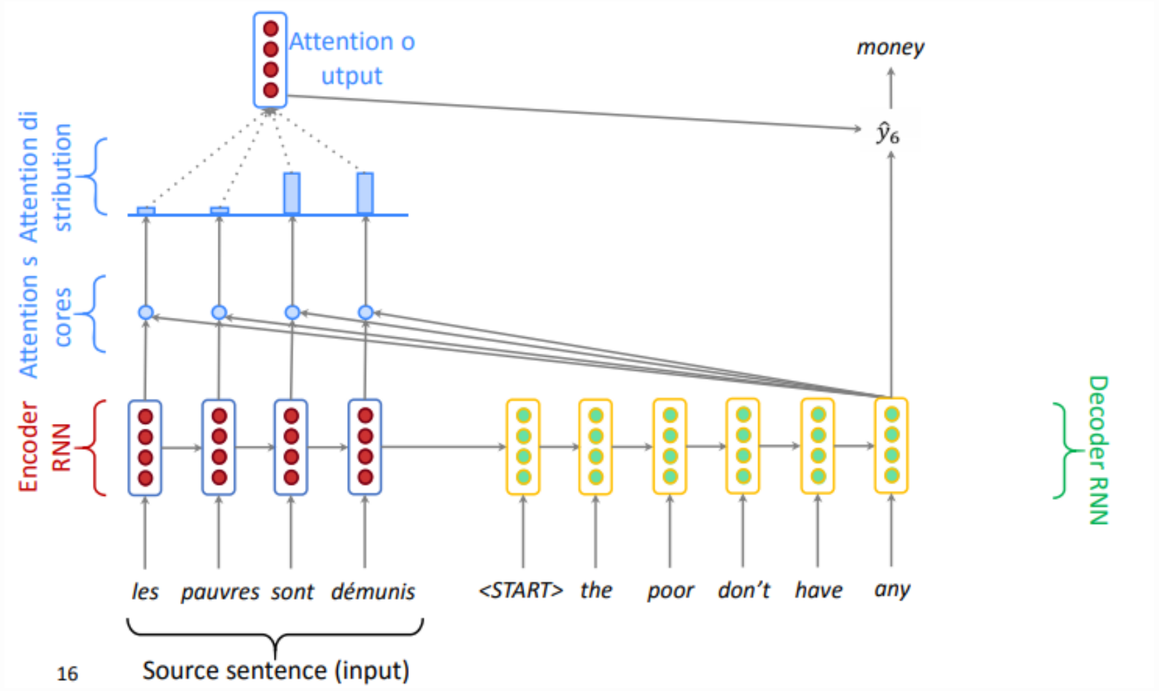
Train

```
BUFFER_SIZE = len(input_tensor_train)
BATCH_SIZE = 64
steps_per_epoch = len(input_tensor_train)//BATCH_SIZE
embedding_dim = 512
units = 1024
vocab_inp_size = len(word2idx_i)+1
vocab_tar_size = len(word2idx_o)+1
```

Epoch = 15

Optimizer = Adam

Seq2seq with attention



Train

```
BUFFER_SIZE = len(input_tensor_train)
BATCH_SIZE = 64
steps_per_epoch = len(input_tensor_train)//BATCH_SIZE
embedding_dim = 512
units = 1024
vocab_inp_size = len(word2idx_i)+1
vocab_tar_size = len(word2idx_o)+1
```

Epoch = 15

Optimizer = Adam

Epoch 1 Loss 1.4085

Epoch 2 Loss 1.1755

Epoch 3 Loss 1.0730

Epoch 4 Loss 0.9877

Epoch 5 Loss 0.9097

Epoch 6 Loss 0.8347

Epoch 7 Loss 0.7614

Epoch 8 Loss 0.6888

Epoch 9 Loss 0.6176

Epoch 10 Loss 0.5493

Epoch 11 Loss 0.4862

Epoch 12 Loss 0.4289

Epoch 13 Loss 0.3777

Epoch 14 Loss 0.3316

Epoch 15 Loss 0.2907

모델평가 - BLEU score

Train Set mean

0.66

Input: <start> have a good day , everyone . <end>
Predicted translation: 모두 좋은 하루 보내세요 . <end>

Input: <start> i m thinking of putting my house on sale . <end>
Predicted translation: 우리 집을 팔려고 내놓을까 생각 중이에요 . <end>

Input: <start> your profile photo is too sexy . <end>
Predicted translation: 당신의 프로필 사진은 한 사진을 찍었습니다 . <end>

Valid Set mean

0.62

Input: <start> i will pick you up when you arrive . <end>
Predicted translation: 당신이 호텔에 도착하면 내가 데리러 갈게 . <end>

Input: <start> we parted after having dinner today . <end>
Predicted translation: 우리는 오늘 저녁을 먹고 헤어졌어요 . <end>

Input: <start> i don t want to go to crowded places . <end>
Predicted translation: 저는 술에 취하고 싶지 않아 . <end>

모델평가

잘 된 번역

Input: <start> i wish my family to be happy with no hardships . <end>
Predicted translation: 엄마가 행복했으면 좋겠어 . <end>

Input: <start> how long does it take to go downtown ? <end>
Predicted translation: 시내로 가는 데 얼마나 걸려요 ? <end>

잘못된 번역

Input: <start> attach the fixing hook to the strap of the shoe back . <end>
Predicted translation: 이 번호로 파일을 참조해 해당 ok 버튼을 눌러서를 입력하세요 . <end>

Input: <start> i will pack tomorrow . <end>
Predicted translation: 내일 제가 내일을 할 겁니다 . <end>

Input: <start> we will be understanding about the product you sent to the warehouse . <end>
Predicted translation: 우리는 제품을 요청한 후에 전달이 되겠습니다 . <end>

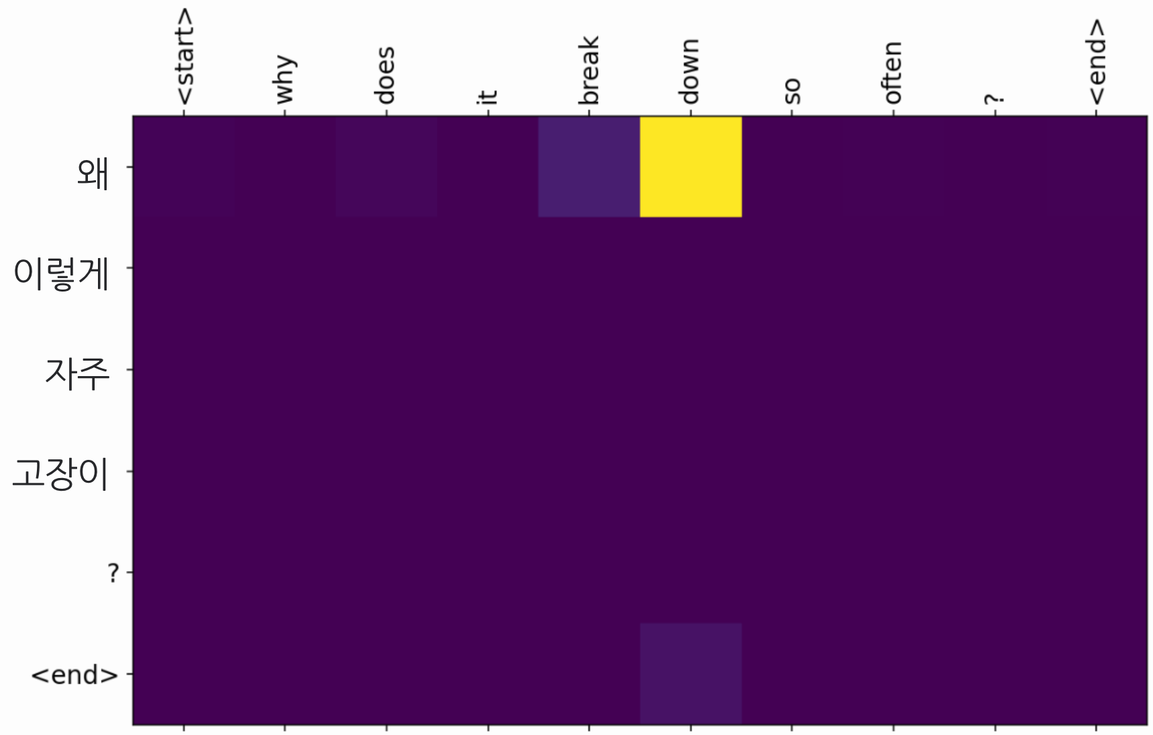
Input: <start> now you can check the number of likes of video uploaded on facebook timeline
Predicted translation: 이제 복사기에서 직접 배달음식의 한 번의 한 번의 한 번의 한 번의 한 번

일부만 맞은 번역

Input: <start> my mom bought a bag and a watch from hong kong . <end>
Predicted translation: 엄마는 어제 가방을 사서 샀습니다 . <end>

Input: <start> send me a message when you can . <end>
Predicted translation: 당신이 할 수 있는 메시지를 보내 주세요 . <end>

모델평가



attention plot을 살펴본 결과, 생각보다는 ttention이 잘 이루어지지 않았음
=> 단어의 매칭만 잘 되었고 맥락을 파악하지는 못함

보완점

발전된 tokenizer 사용

한국어 형태소 분석 후 조사를
떼어내고 SentencePiece Unigram
적용 시 성능이 향상* 된다는 연구

긴 문장의 번역 정확도 향상

attention을 적용했음에도 긴
문장의 번역에 취약한 문제.
=> 여러 개의 attention을 적용한
Transformer model도 고려.

** Taku Kudo, John Richardson, SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing, EMNLP2018

감사합니다.