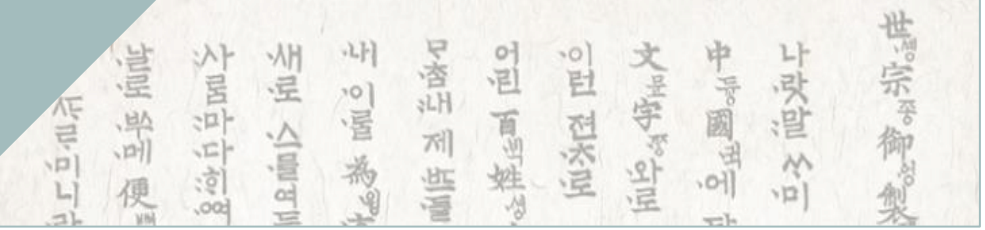
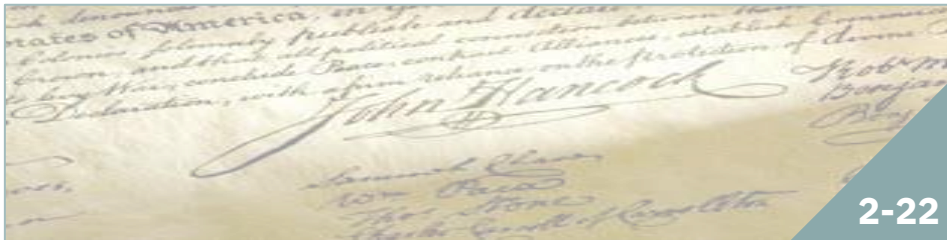
딥러닝을 이용한 한글-영문 번역기

RNN model : Seq2seq with attention

김용재 국승용 박혜정 한민재 황성연



목 차



요 약

모델 핸들링

- LSTM & GRU
  (+ ATTENTION, BI LSTM)
- PARAMS
  (hidden units, dropout, embedding dim)
- Going deeper
- Uniform Distribution
- Transformer

데이터 핸들링

- 형태소 분석기 (5가지)
- 데이터 추가
- Reverse
- Sentences max length
- 불용어 사전

Dataset

AI Hub 한국어-영어 번역(병렬) 말뭉치

SID	원문	번역문
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.
2	씨티은행에서 일하세요?	Do you work at a City bank?
3	푸리토의 베스트셀러는 해외에서 입소문만으로 4차 완판을 기록하였다.	PURITO's bestseller, which recorded 4th rough -cuts by words of mouth from abroad.
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.
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.
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.
7	강아지들과 내 사진을 보낼게.	And I'll send you a picture of me and my dogs.

◆ 문장 70만개

- 구어체 : 자연스러운 구어체 문장 → 400,000 문장
- 대화체 : 상황/시나리오 기반 대화 세트 → 100,000 문장 (추가)
- 문어체(뉴스) : 뉴스 텍스트 → 200,000 문장 (추가)

Best 형태소 분석기 : Mecab (with LSTM)

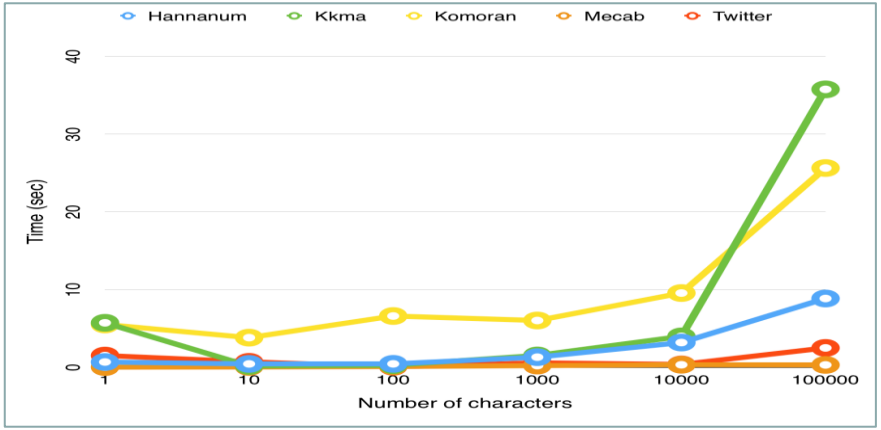
- batch_size : 64
- hidden_units : 256 / Layer 개수 1
- embedding_dim : 64
- optimizer : Adam(learning rate 0.01)

Data 20만

batch(128)	Okt	Mecab	Kkma	Komoran	Hannanum
Accuracy	82.53%	82.26%	82.65%	82.67%	82.56%

Data 40만

batch(64)	Okt	Mecab	Kkma	Komoran	Hannanum
Accuracy	84.32%	84.08%	84.29%	84.17%	83.91%



· 모든 형태소 분석기에서
Data 20만 < Data 40만

· Acc 및 시간 자원 고려
Mecab 선택

사진 출처 : https://konlpy.org/ko/v0.6.0/morph/

Seq2seq 모델 비교 : LSTM vs GRU (with Mecab)

LSTM

구 분	batchsize	learning rate	data set	형태소 분석기	Val_Acc
Base	64	0.01	400,000	Mecab	84.08%
Attention	64	0.01	400,000	Mecab	86.60%

↳ Baseline Model !

GRU

구 분	batchsize	learning rate	data set	형태소 분석기	Val_Acc
Base	64	0.01	400,000	Mecab	84.32%
Attention	64	0.01	400,000	Mecab	82.65%

· Base 및 Attention 모두 LSTM > GRU → LSTM 선택

· LSTM Attention > Base → Attention 선택

모델 성능 향상 : Hidden units control

구 분	Hidden units	Embedding dim	Val_Acc
#1	32	64	85.57%
#2	64	64	86.01%
#3	128	64	84.45%
#4	256	64	86.60%

· Hidden units ↑ → Val_Acc ↑ ∴ hidden units = 256

모델 성능 향상 : Embedding dim control

구 분	learning rate	hidden units	Embedding dim	Val_Acc
#1	0.01	256	32	87.69%
#2	0.01	256	64	86.60%
#3	0.7	256	64	79.50%
#4	0.01	256	128	85.13%

· Embedding dim ↓ → Val_Acc ↑ ∴ Embedding dim = 32

· Learning rate ↓ → Val_Acc ↑ ∴ learning rate = 0.01

모델 성능 향상 : Going Deeper (Depth)

	hidden units	Embedding dim	Depth	Dropout	Val_Acc
Base	256	32	1	X	84.08%
	256	32	2	X	88.37%
	256	32	2	0.3	88.09%
	256	32	4	X	88.67%
	256	32	4	0.3	89.31%

· Depth ↑ → Val_Acc ↑ ∴ depth = 4

· Dropout O → Val_Acc ↑ ∴ dropout OK

모델 성능 향상 : Going Deeper (Dropout)

	hidden units	Embedding dim	Depth	Dropout	Val_Acc
Base	256	32	1	X	84.08%
	256	32	4	X	88.67%
	256	32	4	0.3	89.31%
	256	32	4	0.5	86.48%
	256	32	4	0.7	88.52%

· Dropout ↓ → Val_Acc ↑ ∴ dropout = 0.3

모델 성능 향상 : Going Deeper (hidden units, embedding dim)

	hidden units	Embedding dim	Depth	Dropout	Val_Acc
Base	256	32	1	X	84.08%
	64	32	4	0.3	86.78%
	128	32	4	0.3	88.37%
	256	64	4	0.3	88.19%
	256	32	4	0.3	89.31%

· Hidden units ↑ → Val_Acc ↑ ∴ Hidden units = 256

· Embedding dim ↓ → Val_Acc ↑ ∴ Embedding dim = 32

모델 성능 향상 : Going Deeper (Data preprocessing)

	hidden units	Embedding dim	Data Preprocessing	Learning rate	Depth	Dropout	Val_Acc
Base	256	32	X	0.01	1	X	84.08%
	256	32	X	0.01	4	0.3	89.31%
	256	32	O	0.01	4	0.3	79.92%
Pick	256	32	X	0.001	4	0.3	97.94%
	256	32	O	0.001	4	0.3	96.43%

◆ Data preprocessing

· 해석에 영향을 미치지 않는다고 생각되는 불용어 추가 → 주격조사(은, 는, 이, 가) 삭제

· Data preprocessing O → Val_Acc ↓ ∴ data preprocessing X
· Learning rate ↓ → Val_Acc ↑ ∴ learning rate = 0.001

중간 점검 및 피드백

학습 후 Train Data로 예측한 결과

입력문장 : 11장에서는 예수님이 이번에 나사로를 무덤에서 불러내어 죽은 자 가운데서 살리셨습니다.
정답문장 : In Chapter 11 Jesus called Lazarus from the tomb and raised him from the dead.
번역문장 : bowed paste, Yes, funeral cupcake defendant Kyungpook defendant Kyungpook defendant Kyungpook .....

입력문장 : "공식 초청 레터"를 받는 데로 비자를 받을 것입니다.
정답문장 : I will get my vias as soon as I receive the " official invitation letter."
번역문장 : twitter topic. noun names? twitter complaint pre-formed twitter complaint pre-formed .....

입력문장 : 간다 양이 암이라고 저는 들었어요.
정답문장 : I heard that Miss Kanda has cancer.
번역문장 : bone grave grave grave grave grave grave grave .....

◆ 높은 Val_Acc에도 불구하고 대부분의 문장 오역 및 낮은 BLEU Score

· 오역 · 같은 단어 반복 · 이전 문장의 번역 단어 재등장

▶ 위 문제를 해결하기 위해 추가적인 핸들링 필요 판단

① 데이터 추가 학습
② 문장 별 토큰 수 기준 긴 문장 삭제
③ 추가 모델 핸들링 (논문 참고)

중간 점검 및 피드백

① 데이터 추가 (40만개 → 70만개)

문장 40만개

- 구어체 400,000 문장

→

문장 70만개

- 구어체 400,000 문장
- 대화체 100,000 문장
- 문어체 (뉴스) 200,000 문장

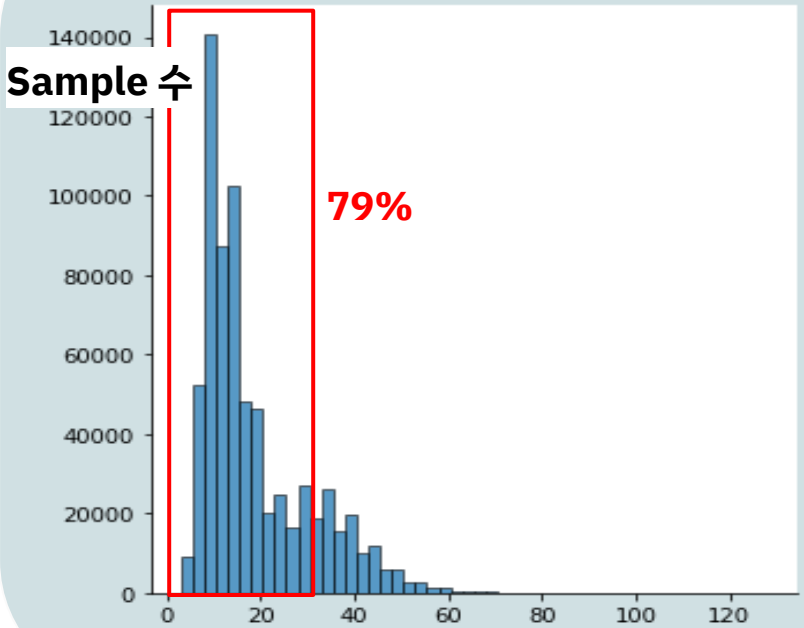
▶ 데이터 추가 시 기대 효과

① Val_Acc 향상
② Overfit 일부 해소
③ 데이터 다양화 (대화체, 문어체)

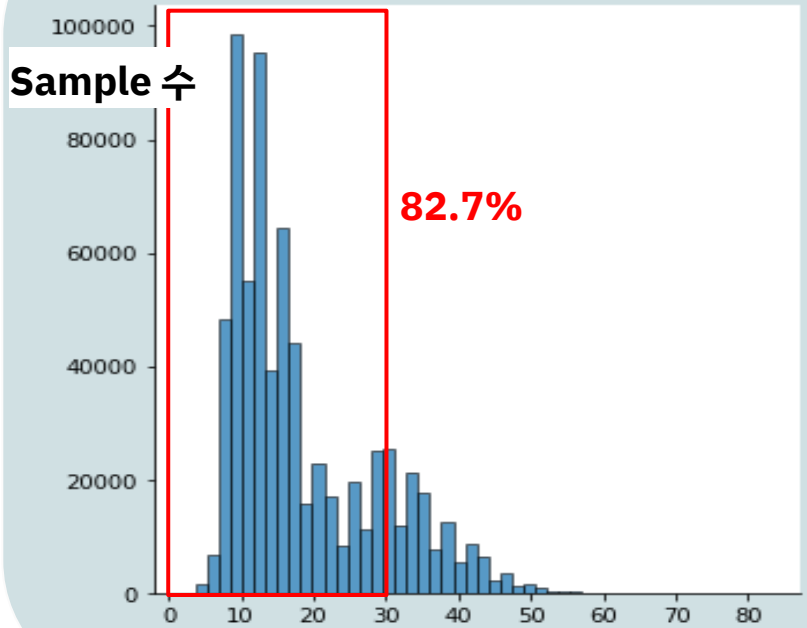
중간 점검 및 피드백

② 문장 별 토큰 수 기준 긴 문장 삭제

한글 문장 길이



영어 문장 길이



・대부분의 문장이 30개 이하의 토큰으로 이루어짐 → 30개 초과 샘플 삭제 (패딩 효율↑)

중간 점검 및 피드백 ③ 추가 모델 핸들링 1 (Layer)

Pick

구 분	hidden units	Embedding dim	# of Data	Learning rate	Depth	ETC	Dropout	Val_Acc
#1	256	32	400,000	0.001	4	LSTM	0.3	97.94%
#2	256	32	700,000	0.001	4	BI LSTM	0.3	97.29%
#3	256	32	700,000	0.001	4	BI directional GRU	0.3	95.62%

- **BI LSTM** : 비슷한 Val_Acc
- **BI directional GRU** : 데이터 ↑, But Val_Acc ↓

입력문장 : BBC에서는 지금 뭘 하려나?
정답문장 : I'm wondering what the BBC are doing at the moment?
번역문장 : Something From We went gone went ribs gone BBC BBC BBC hairstyle Since BBC BBC BBC BBC BBC BBC BBC BBC BBC BBC BBC BBC BBC unusual unusual hairstyle hair

▶ 몹시 **형편없는 Test 결과** → 주요 단어는 어느정도 짚어내지만, 이전과 동일한 결과 지속 (오역, 같은 단어 반복)

중간 점검 및 피드백

③ 추가 모델 핸들링 2 (Params)

Pick

구 분	hidden units	Embedding dim	# of Data	Learning rate	Depth	ETC	Dropout	Val_Acc
#1	256	32	400,000	0.001	4	LSTM	0.3	97.94%
#2	256	32	700,000	0.001	4	Uniform Distribution	0.3	85.54%
#3	256	32	700,000	0.001 Nadam	4	Optimizer	0.3	83.29%
#4	256	32	700,000	0.005 Nadam	4	Optimizer + Uniform Distribution	0.3	85.86%

- **Uniform Distribution (weight 초기값 설정)** : 성능 향상에 영향 X
- **N Adam (Optimizer 변경)** : 성능 향상에 영향 X

입력문장 : 네가 하는 일과 공부 잘하길 멀리서 응원할게. → **MISMATCH**
정답문장 : I will cheer you on your work and your grade from far away.
번역문장 : Myeongsoo. What's miserable, What's momentous awaiting superheroes stiffened Bake .....

▶ 마찬가지로 미흡한 수준의 번역 결과, 이전과 동일한 결과 지속 (오역, 같은 단어 반복)

결과 1 (평가지표 : Val_Acc)

모델 검증 (Test data)	모델 실 적용 (Predict)	BLEU score
입력 : 다른 선수들이 몬스터를 사냥할 경우 당신은 추가 경험치를 획득해요. 정답 : If other players hunt monsters, you gain additional experience. 번역 : artists' That. CSR. paycheck Mechanics \ Mecha	입력 : 아기자기 곰 펀칭이 사랑스러운 파자마! 번역 : deodorization deodorization uninsured uninsured balance balance $100. $100. Mechanics Mechanics Mechanics Mechanics 입력 : 2PM의 닉쿤을 소개합니다. 번역 : u673 u673 u673 IMS Gulbi fix-ups. $100. $100. agitated? agitated?	9.577466191714931 × e-232

· Val_Acc가 높다고 만족스러운 번역 결과를 얻지 못했다.
· 따라서 평가 지표를 1) 사람에 의한 평가 및 2) BLEU score로 변경하였다.

결과 2 (평가지표 변경 : 사람에 의한 평가 & BLEU score)

	모델 검증 (Test data)	모델 실 적용 (Predict)	BLEU score
LSTM depth1	입력 : 당신 의 음악 취향 마음 에 들 어요 . 정답 : I like your taste in music . 번역 : I love the taste of the song you gave me .	입력 : 정리가 필요하겠네요. 번역 : I need a quick order . 입력 : 처음 만났을 때를 떠올려 보세요 번역 : Think about the first time when you are in the first meeting .	1.3591678403181048 × e-231
LSTM depth1 with Attention	입력 : 우리 영화 감상 후 에 집 으로 가요 . 정답 : We go home after having a simple dinner . 번역 : Lets go to the movie theater together .	입력 : 정리가 필요하겠네요. 번역 : I need to organize it . 입력 : 처음 만났을 때를 떠올려 보세요 번역 : When you see the first time I met you .	1.361077635708999 × e-231

- 지금까지 학습한 모델들로부터 사람에 의한 평가 및 BLEU score 확인
- 번역 결과가 괜찮다고 생각하는 2가지 모델 선정 (① LSTM depth1, ② LSTM depth1 with Attention)

결과 3 (추가 모델핸들링 : Transformer)

Transformer

- Positional Encoding 사용 (RNN 사용 X)
- 잔여학습(Residual Learning) 사용 (for 성능향상)
- Attention(어텐션)과 Normalization(정규화) 과정 반복
- 문장 전체의 Attention 값 계산, 한 번의 계산에 병렬적으로 출력값 획득
  → 계산 복잡도↓

- Seq2seq with attention 모델의 한계점 보완
- Transformer를 사용하면 더 좋은 성능을 획득 가능

결과 3 (추가 모델핸들링 : Transformer)

Transformer

모델 검증 (Test data)

입력 : 당신 의 음악 취향 마음 에 들 어요 .
정답 : I like your taste in music .
번역 : I like your taste in music .

입력 : 우리 영화 감상 후 에 집 으로 가요 .
정답 : We go home after having a simple dinner .
번역 : Let's go to the house after the text .

모델 실 적용 (Predict)

입력 : 정리가 필요하겠네요.
번역 : I think I need to organize .

입력 : 처음 만났을 때를 떠올려 보세요
번역 : Please think of me when you first met .

입력 : 2PM의 닉쿤을 소개합니다.
번역 : I introduce 2PM's Nick .

- Transformer을 이용해 학습한 모델로 사람에 의한 평가 확인
- 사람에 의한 평가를 기준으로 최종 모델 선정 (Transformer model)

결론 및 한계점

<한계점>

- layer가 깊어짐에 따라 val acc는 증가하였으나, 결과값 및 예측값에서 특정 단어가 반복되어 나오는 현상 조치
- OOV 문제 해결 필요
- BLEU score 향상을 위한 데이터 추가 필요
- 추가 데이터 전처리 과정 필요

<결론>

- 다양한 데이터 및 모델 핸들링 시도
- 사람에 의한 평가 및 BLEU score를 통해 선택된 LSTM depth1 with Attention은 나쁘지 않으나 충분히 만족스러운 결과 획득 X
- Transformer를 이용하여 한글-영문을 번역기 구축

['<SOS>', '감사', '합니다']

['thank', 'you', '<EOS>']