ALBERT를 이용한 한국어 자연어처리: 감성분석, 개체명 인식, 기계독해

이영훈[O1], 나승훈[1], 최윤수[2], 이혜우[2], 장두성[2]

[1]전북대학교, [2]KT

Contents

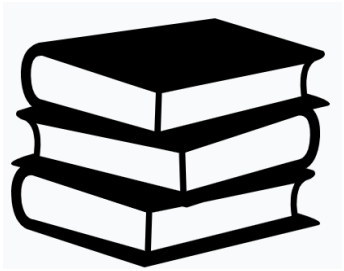
- Background
- Related Works
- AL-RoBERTa
- Experiment Result
- Conclusion

Pre-training Language Model

- Transformer Encoder 기반의 사전 학습 언어모델
- 대용량 말뭉치를 이용하여 사전 학습을 진행하고, 각 태스크에 fine-tuning 적용
- 다양한 자연어처리 태스크에서 SOTA를 달성

1. Pre-training

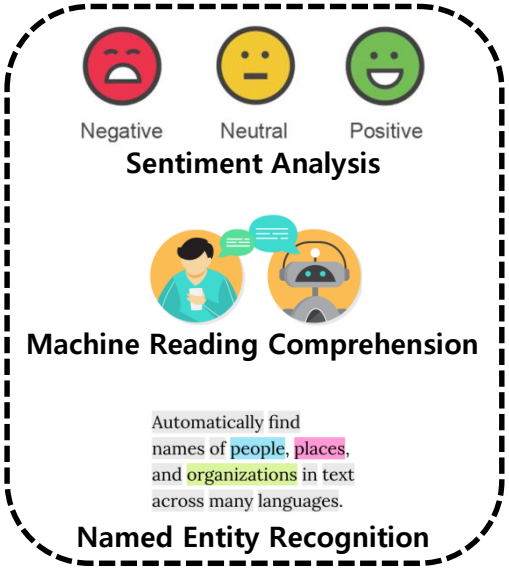
Unsupervised Learning



Corpus

2. Fine-tuning

Supervised Learning



Downstream Tasks

Pre-training Language Model

SQuAD2.0

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	**90.724**	**93.011**
3 May 04, 2020	ELECTRA+ALBERT+EntitySpanFocus (ensemble) *SRCB_DML*	90.442	92.839
4 Mar 12, 2020	ALBERT + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	90.386	92.777
5 Jan 10, 2020	Retro-Reader on ALBERT (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694v2	90.115	92.580
6 Nov 06, 2019	ALBERT + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	90.002	92.425
7 Sep 18, 2019	ALBERT (ensemble model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	89.731	92.215
7 Feb 25, 2020	Albert_Verifier_AA_Net (ensemble) *QIANXIN*	89.743	92.180

KorQuAD 1.0

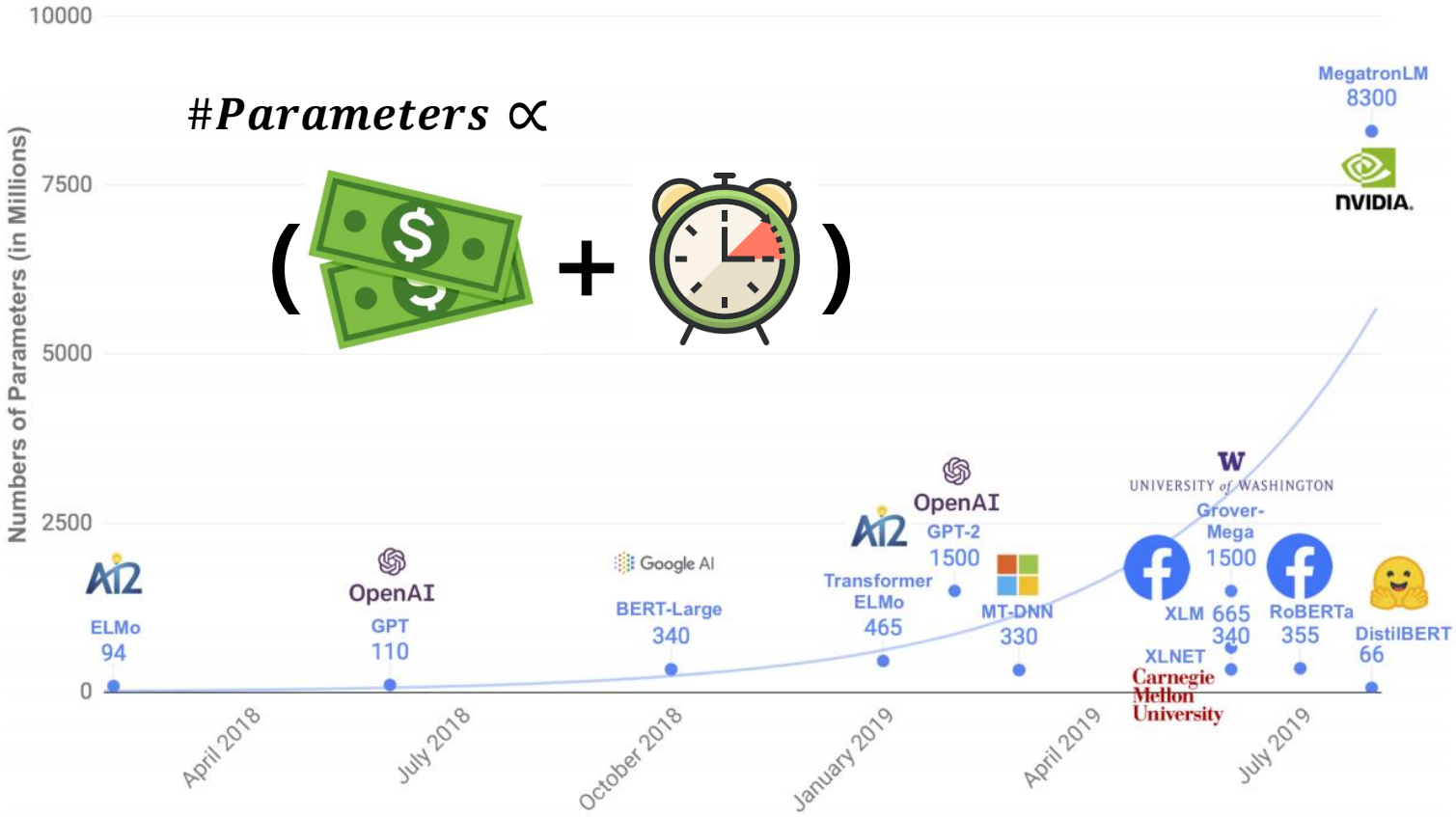
Rank	Reg. Date	Model	EM	F1
-	2018.10.17	**Human Performance**	80.17	91.20
1	2020.01.08	SkERT-Large (single model) Skelter Labs	87.66	**95.15**
2	2019.10.25	KorBERT-Large v1.0 ETRI ExoBrain Team	**87.76**	95.02
4	2019.06.26	LaRva-Kor-Large+ + CLaF (single) Clova AI LaRva Team	86.84	94.75
6	2019.06.04	BERT-CLKT-MIDDLE (single model) Anonymous	86.71	94.55

KorQuAD 2.0

Rank	Reg. Date	Model	EM	F1
-	2019.09.05	**Human Performance**	68.82	83.86
1	2020.05.03	SDS-NET v1.1 Samsung SDS AI Advanced Research	**73.87**	**86.81**
3	2020.06.15	CNS-BERT (single model) Seungyoung Lim	70.67	83.57
4	2020.04.20	LaRva (single model) NAVER Clova AI LaRva	66.95	83.54
5	2020.02.21	CNS-BERT (single model) Seungyoung Lim	68.39	82.62

Pre-training Language Model

- 학습 데이터가 많을수록, 모델의 파라미터가 많을수록 성능 증가
- 파라미터의 수가 많아질수록, 자원의 한계로 현실적으로 큰 모델의 학습은 어려움



Sanh, Victor, et al. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.", 2019.

Training Efficiency

- 기존에 학습된 Teacher Network를 확장된 Student Network의 초기 파라미터로 사용하여 학습 속도 및 성능에 영향 (Knowledge Transfer without training)

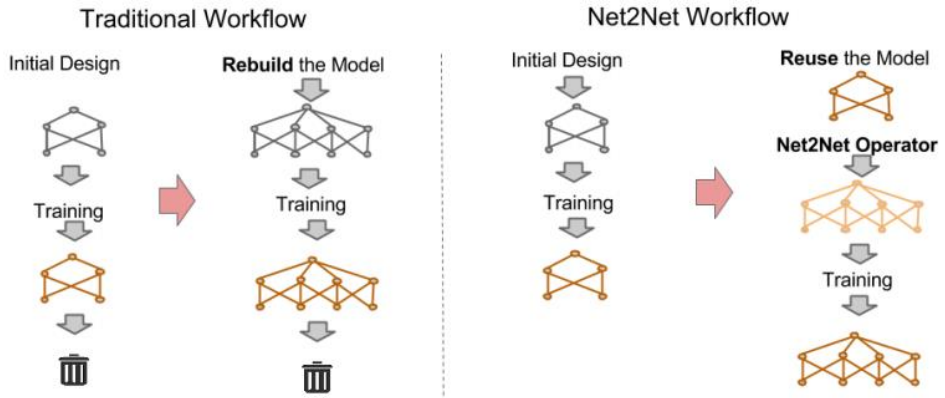
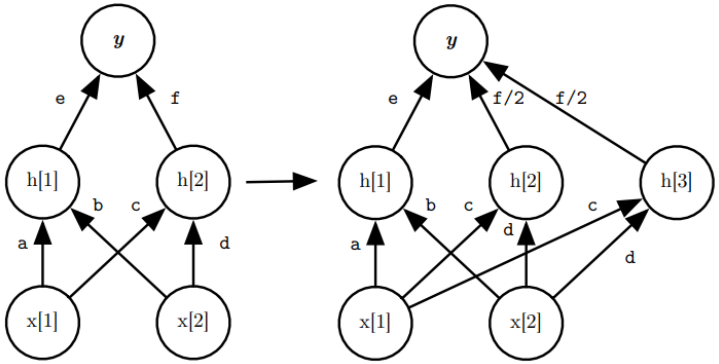
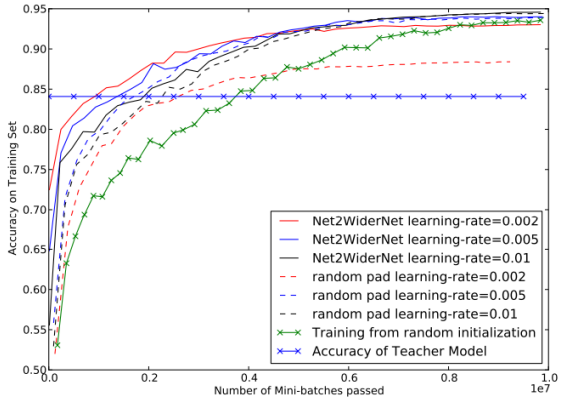


Figure 1: Comparison between a traditional workflow and the `Net2Net` Workflow; `Net2Net` reuses information from an already trained model to speed up the training of a new model.

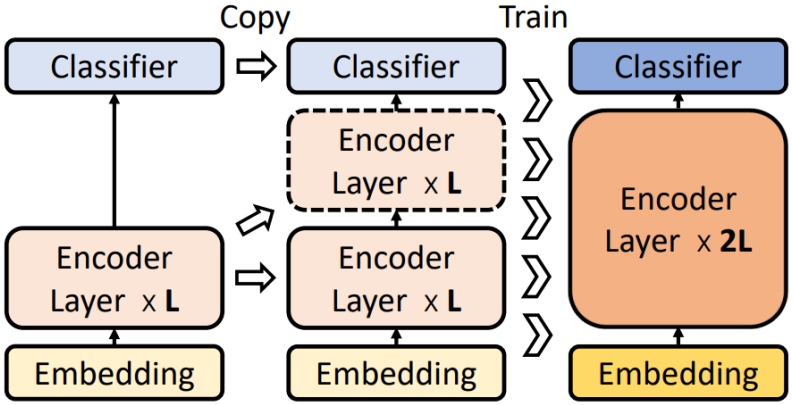




Chen, Tianqi, Ian Goodfellow, and Jonathon Shlens. "Net2net: Accelerating learning via knowledge transfer.", 2015.

Training Efficiency

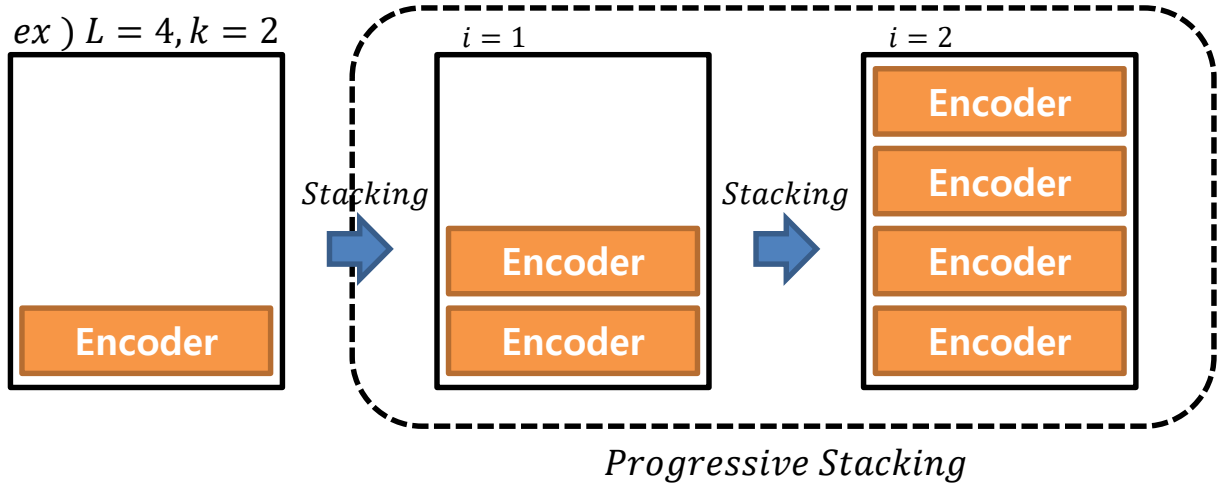
- Shallow 모델을 쌓아 Deep 모델을 만듦으로 Training efficiency 증가



Algorithm 1 Progressive stacking

$M_0' \leftarrow \text{InitBERT}(L/2^k)$
$M_0 \leftarrow \text{Train}(M_0')$ {Train from scratch.}
for $i \leftarrow 1$ to k **do**
  $M_i' \leftarrow \text{Stack}(M_i)$ {Doubles the number of layers.}
  $M_i \leftarrow \text{Train}(M_i')$ {M_i has $L/2^{k-i}$ layers.}
end for
return M_k

Figure 3. The diagram of the *stacking* algorithm.



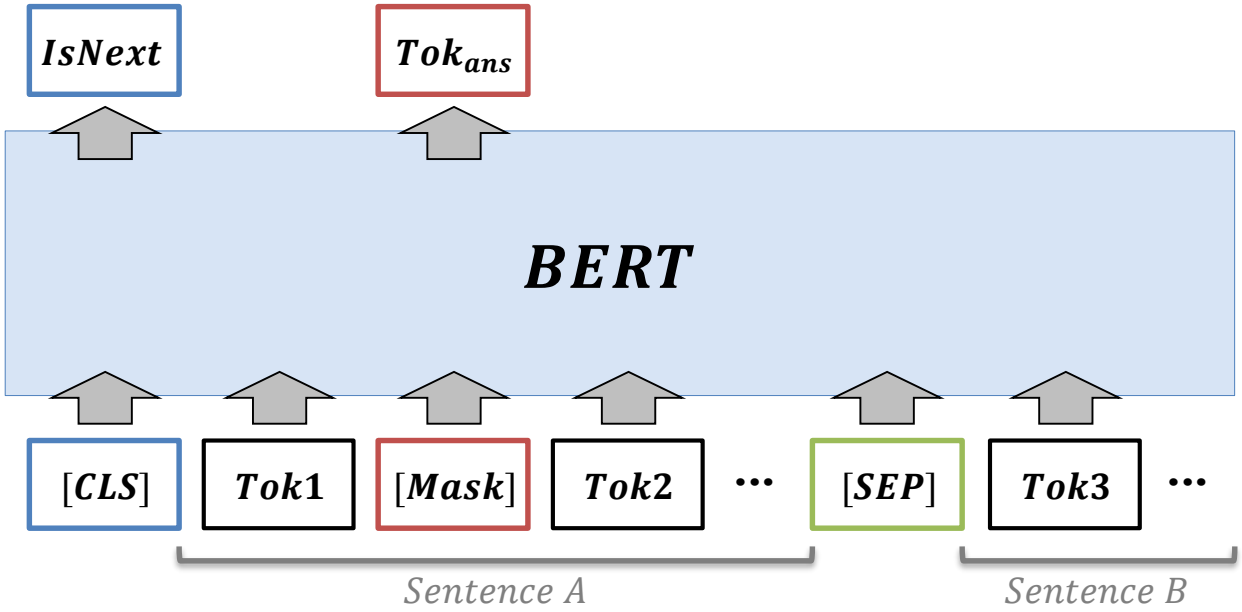
Gong, Linyuan, et al. "Efficient training of bert by progressively stacking." 2019.

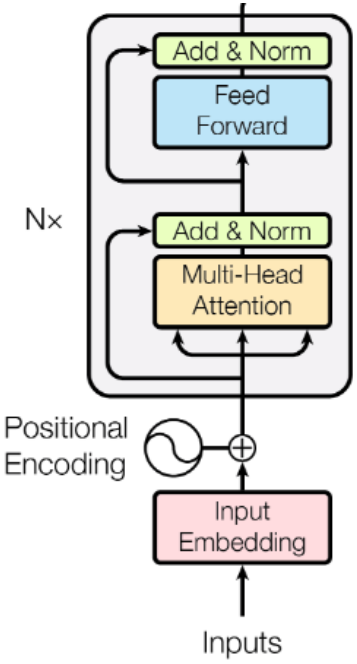
BERT:

Pre-training of Deep Bidirectional Transformers for Language Understanding

- Transformer Encoder 기반 모델
- 두 가지 태스크를 이용하여 사전학습

1. **Masked LM** : 문자 토큰을 랜덤으로 마스킹하고 마스킹 된 위치의 토큰 예측
2. **NSP(Next Sentence Prediction)** : 두 문장의 순서가 적절한지 예측





RoBERTa:

A Robustly Optimized BERT Pre-training Approach

- **NSP(Next Sentence Prediction) 제거**
  - BERT 모델의 학습에 NSP의 효용성에 의문을 제기하며 **NSP Loss 제거**
  - 하나 이상의 문서를 이용하여 최대 토큰 길이에 가깝게 구성하는 **FULL-SENTENCES**
- **Dynamic Masking**
  - 기존 고정된 마스크 위치를 학습하는 BERT와 달리 **마스크의 위치를 동적으로 결정**
- **Whole Word Masking**
  - Google AI에서 word piece 단위의 마스킹을 개선한 방법 제안
  - 한 단어가 여러 word piece로 구성 될 경우, 한 단어에 해당하는 **word piece를 모두 마스킹**하여 성능 향상

ALBERT:

A Lite BERT for Self-supervised Learning of Language Representations

- **SOP(Sentence Order Prediction)**
  - RoBERTa와 같이 NSP의 문제점을 인식하고 이를 개선한 **SOP(Sentence Order Prediction) 제안**
  - 연속되는 두 문장(Positive)과 문장 순서를 앞뒤로 바꾼 문장 (Negative)으로 구성되어 문장의 순서가 옳은지 여부를 판단

- **Factorized Embedding Parameterization**
  - Embedding Size(E)와 Hidden Layer Size(H)를 따로 적용
  - $O(V \times H)$에서 $\boldsymbol{O(V \times E + E \times H)}$로 파라미터 수 감소

- **Cross-Layer Parameter Sharing**
  - Transformer layer의 attention layer와 FFN(Feed-forward Network) **파라미터를 공유**하여 전체 파라미터 수를 줄임

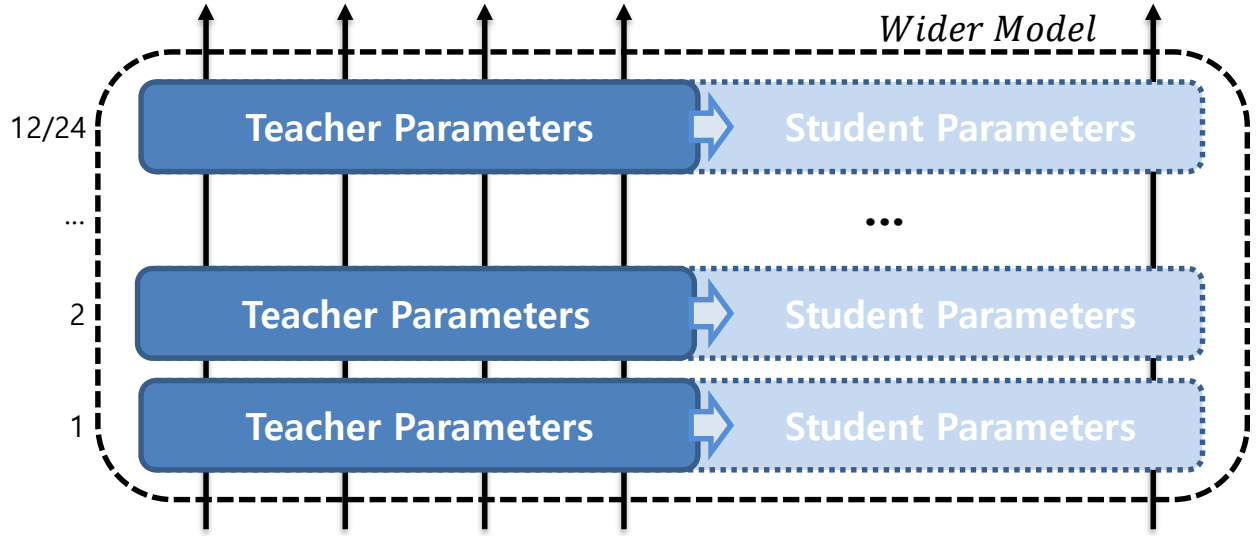
제안모델: AL-RoBERTa

- 기존의 RoBERTa 모델에 ALBERT의 **(1) Factorized Embedding Parameterization**, **(2) Cross-Layer Parameter Sharing**를 적용하여 파라미터 수 감소
- large 모델에서 RoBERTa 모델에 비교하여 파라미터 개수 약 20배 감소
- ALBERT xlarge 모델을 이용하여 실험 진행

Model		#Param	#Layer	Hidden	Embedding	Param sharing
RoBERTa	**base**	110M	12	768		False
	large	340M	24	1024		
	xlarge	1278M	24	2048		
AL-RoBERTa	base	11M	12	768	128	True
	large	17M	24	1024		
	xlarge	55M	24	2048		
	xxlarge	207M	12	4096		

제안모델: AL-RoBERTa

- **Wider ALBERT 적용**
  - 랜덤하게 파라미터를 초기화 하는 것이 아닌 ALBERT 모델의 확장에 따라 이전 모델(Teacher)의 파라미터를 이용하여 모델(Student) 초기화함 (Knowledge Transfer)
  - 동일한 Gradient로 학습되는 것을 방지하기 위해 각 Weight의 Scale을 고려하여 확장된 파라미터에 Gaussian Distribution을 이용하여 Noise 추가

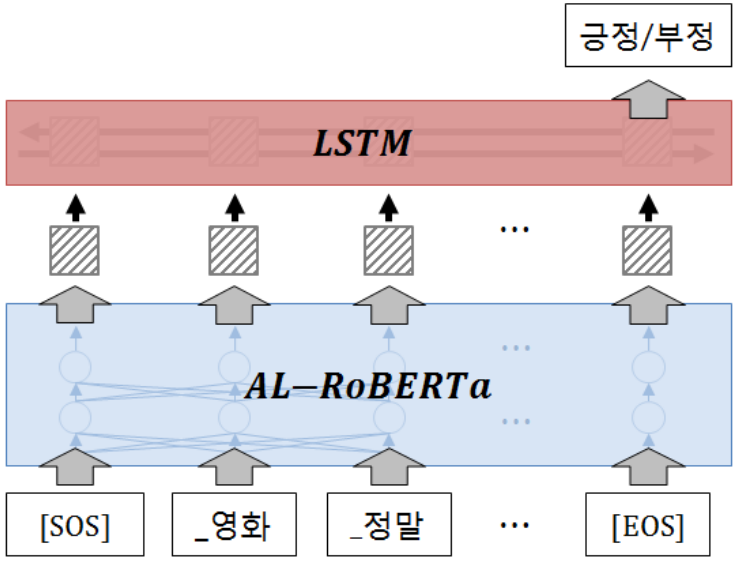


- **자소 단위 BPE Tokenizer**
  - 형태소 단위 : 형태소 분석기 오류 전파, OOV(Out of Vocabulary)
  - 음절 단위 : 의미 구분의 한계

입력 문장: 난 프랑스 영화가 이래서 좋다..
결과 토큰: _난 _프랑스 _영화 가 _이래 서 _좋다 . .

Experiment Result

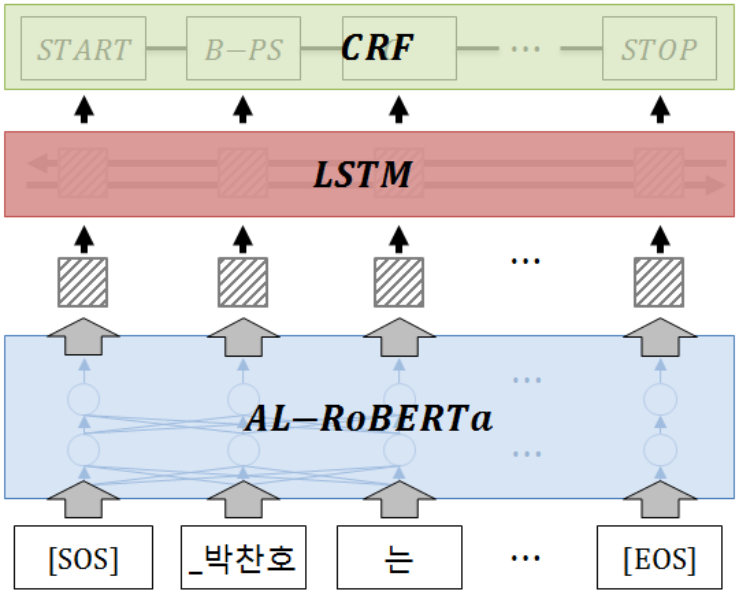
- **Sentiment Analysis**
  - 네이버 영화리뷰 감성분석 데이터 (NSMC, Naver Sentiment Movie Corpus) 사용
  - 마지막 레이어 출력 값에 양방향 LSTM 적용
  - 입력 토큰 : [SOS] Context [EOS]
  - 다른 모델들과 비교하여 적은 수의 파라미터를 가짐에도 비슷하거나 높은 성능을 보임
  - 특히 동일한 데이터로 학습한 RoBERTa 모델보다 높은 성능을 보임



Model	#Param	학습데이터 크기	정확도
BERT(형태소태그)[9]	110M	500M	86.57
RoBERTa(형태소태그)[10]	110M	15G	89.88
RoBERTa	110M	500M	85.17
AL-RoBERTa	**55M**	**500M**	**86.05**

Experiment Result

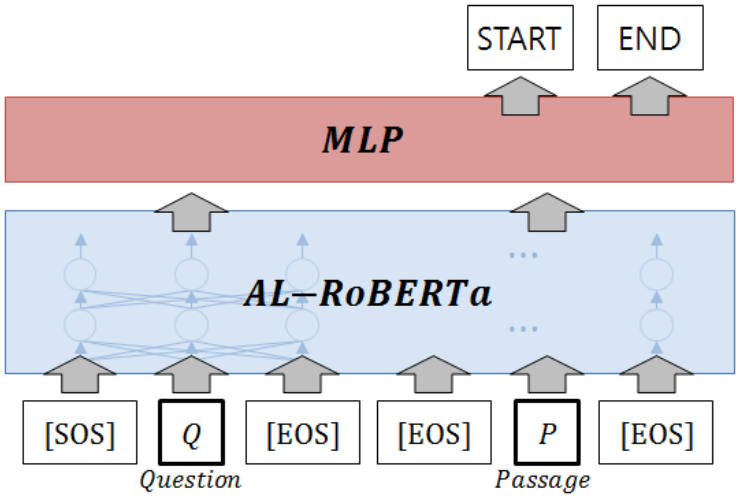
- **NER (Named Entity Recognition)**
  - ETRI 엑소브레인 언어분석 말뭉치 데이터 사용
  - 마지막 레이어 출력 값에 양방향 LSTM을 이용하여 인코딩, CRF를 이용하여 출력 층 구성
  - 입력 토큰 : [SOS] Context [EOS]
  - 감성분석의 결과와 동일하게 적은 수의 파라미터를 가짐에도 비슷하거나 높은 성능을 보임



Model	#Param	학습데이터 크기	정확도
BERT(형태소태그)[9]	110M	500M	91.58
RoBERTa(형태소태그)[10]	110M	15G	94.79
RoBERTa	110M	500M	91.94
AL-RoBERTa	**55M**	**500M**	**91.87**

Experiment Result

- **MRC (Machine Reading Comprehension)**
  - 한국어 질의응답 데이터 KorQuAD(The Korean Question Answering Dataset) 사용
  - 마지막 레이어 출력 값에 MLP Layer를 추가하여 start, end point를 각각 얻음
  - 입력 토큰 : [SOS] Question [EOS] [EOS] Passage [EOS]
  - Google-multilingual 모델의 경우 상위 100개 언어의 위키피디아 말뭉치의 대용량 데이터로 학습



Model	#Param	학습데이터 크기	EM	F1
BERT+MCAF (google-multilingual) [13]	110M	N/A	83.01	91.43
BERT-ETRI [14]	110M	23.5G	84.82	92.74
KT RoBERTa [8]	110M	18G	87.11	94.47
RoBERTa	110M	500M	78.63	88.25
AL-RoBERTa	**55M**	**500M**	**82.98**	**91.44**

결론 및 향후 연구

- **결론**
  - 큰 규모의 사전학습 언어모델 학습에서 파라미터에 따른 자원의 한계를 해결하기 위해 RoBERTa 모델에 파라미터 수를 줄인 AL-RoBERTa를 제안하고 여러 한국어 태스크에 적용
  - 학습의 효율성을 위해 Wider-Net을 이용하여 모델의 파라미터 초기화
  - 모델은 다른 모델과 비교하였을 때 적은 수의 파라미터와 데이터로 학습하였음에도 비슷하거나 태스크에 따라서는 높은 성능을 보여주었음
  - 대용량 데이터에 대해서 학습을 진행하였을 때 성능 향상을 기대할 수 있음

- **향후 계획**
  - 큰 규모의 모델과 대용량 데이터를 이용하여 학습
  - 학습 효율성(Training Efficiency)를 향상시키기 위한 연구 진행

감사합니다