Text Generation And QA Model With Pruning Using KorQuAD

배윤성, 김경민, 장준우, 권순완, 전동규, 김욱
Samsung Electronics Co., Ltd.

Outline

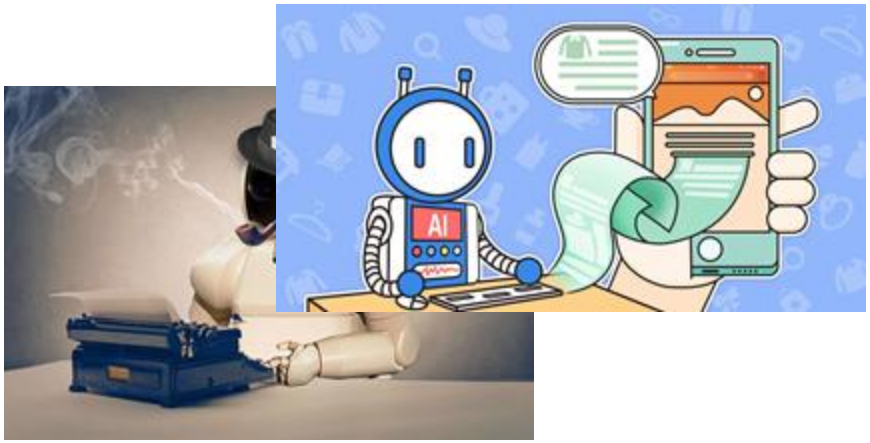
- **Motivation**
- **Proposed Method**
  - Overall System Diagram
  - Discrete VAE
  - BiDAF based QA Model
  - BERT based QA Model
  - Lightweight BERT by Pruning
- **Experimental Results**
- **Summary**

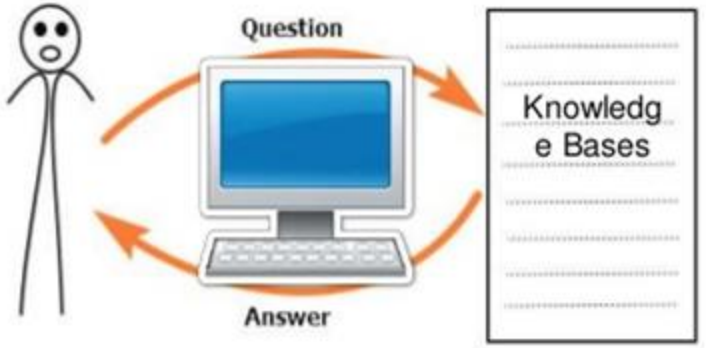
Motivation and Goal

- 한국어 **NLP: 영어에 비해 상대적으로 data의 숫자가 작은 편**
  - 언어 연구의 중요한 문제점 중 하나

- **Solution: data augmentation by text generation**
  - Text generation model을 통해 학습 성과를 향상시키는 것이 가능

- **Goal: QA model improving by data augmentation**
  - KorQuAD data를 augmentation 하여 한국어 QA model의 성과를 향상
  - 성공할 경우 향후 data가 부족한 다른 application에도 적용 가능
    E.g.) FAQ generation model



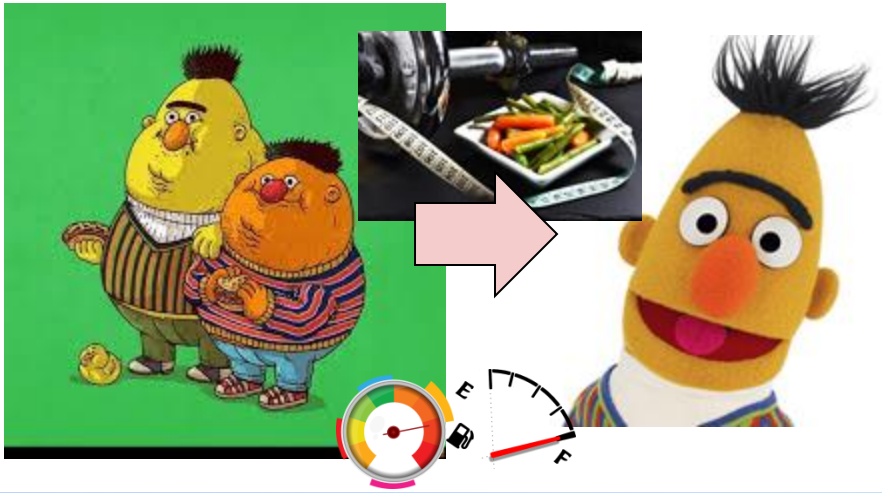


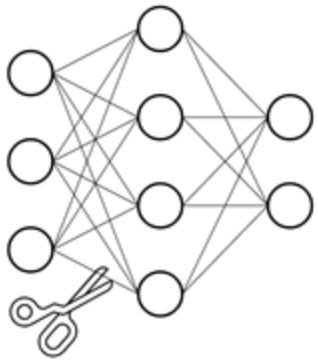
Motivation and Goal

- **Another issue: heavy model size**
  - 최근 대중화된 BERT의 경우 model의 크기가 매우 크다는 단점을 지님
  - Memory size, computation speed, power consumption 등에 약점을 보임
  - Edge device computing 등에는 사용하기 어려움

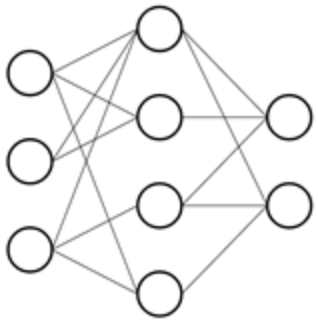
- **Solution: model pruning**
  - 영향이 크게 없는 weight / node를 제거함으로써 경량화 가능

- **Goal: 수업에 배운 pruning 기법을 BERT에 적용하여 경량화 진행**





Before pruning



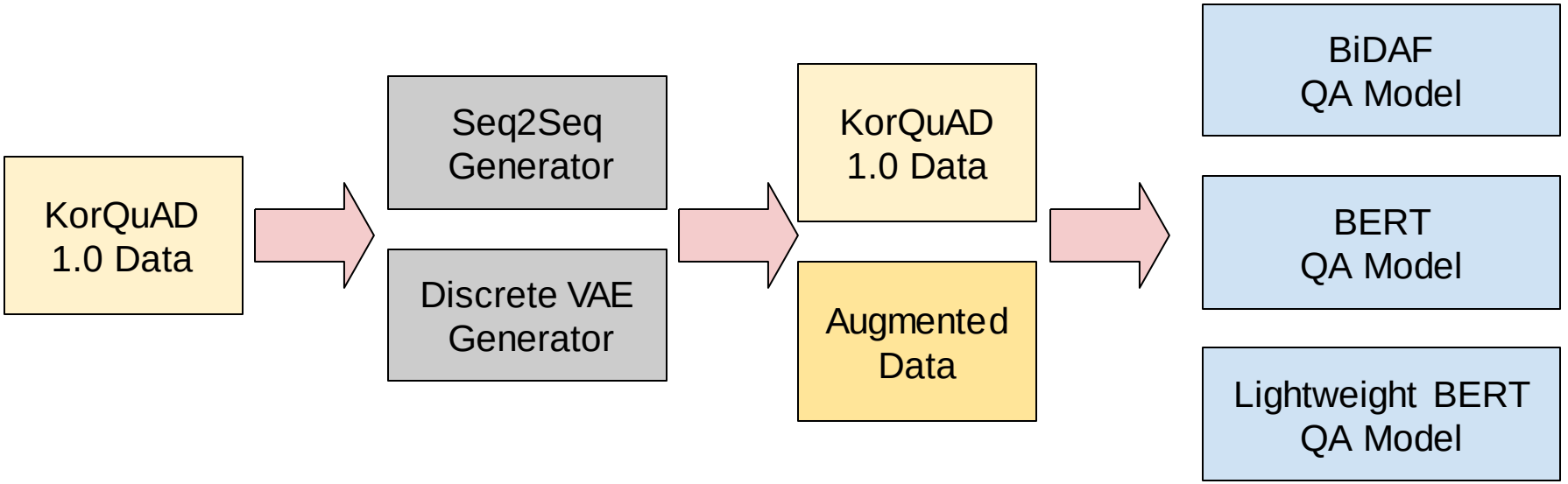
After pruning

Proposed Method

- **Overall system diagram**

Text Generation
- Seq2Seq Generator
- Discrete VAE Generator

QA Model
- BiDAF
- BERT
- Bert with Pruning

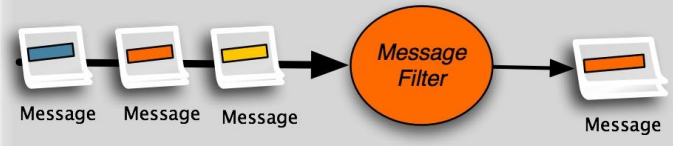


Proposed Method – QA Text Generator

- **VAE based QA text generator**
  - 장점: Answer / question 을 동시에 생성 가능, 더 다채로운 문장 생성 가능
  - 단점: Training이 어려움, Hyper-parameter tuning에 시간 소요

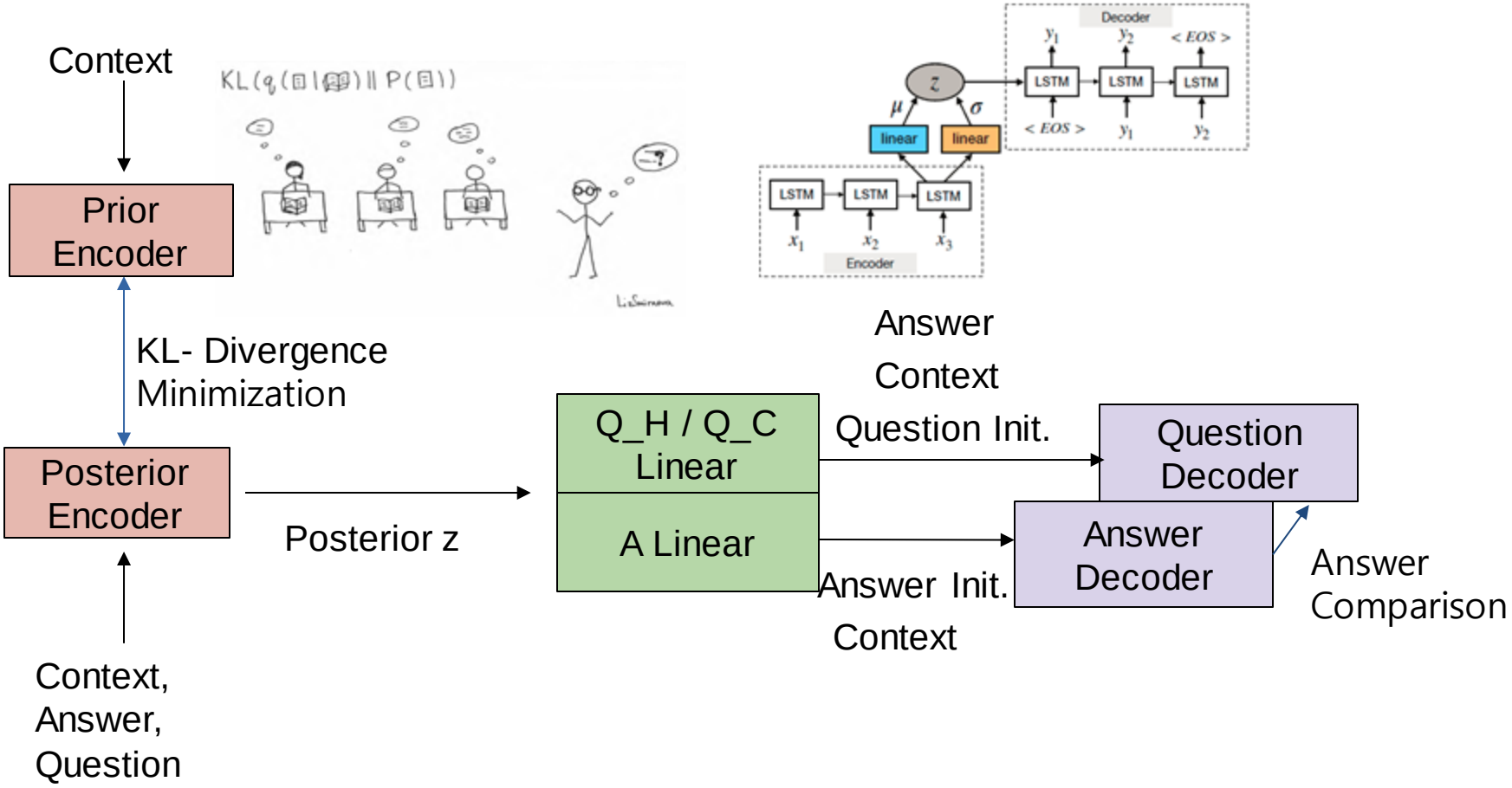
- **Seq2Seq based QA text generator**
  - 장점: 단순한 구조, Training이 쉬운 편, Question이 안정적으로 생성됨
  - 단점: Answer는 생성할 수 없음, 다채로운 문장이 나오지 않음

- **전략: 두 generator를 모두 사용하여 text augmentation 진행**
  - Seq2Seq generator: 안정적인 문장 생성
  - VAE generator: 다양한 형태의 문장 생성
  - Filtering: quality가 현저히 낮은 문장 제외



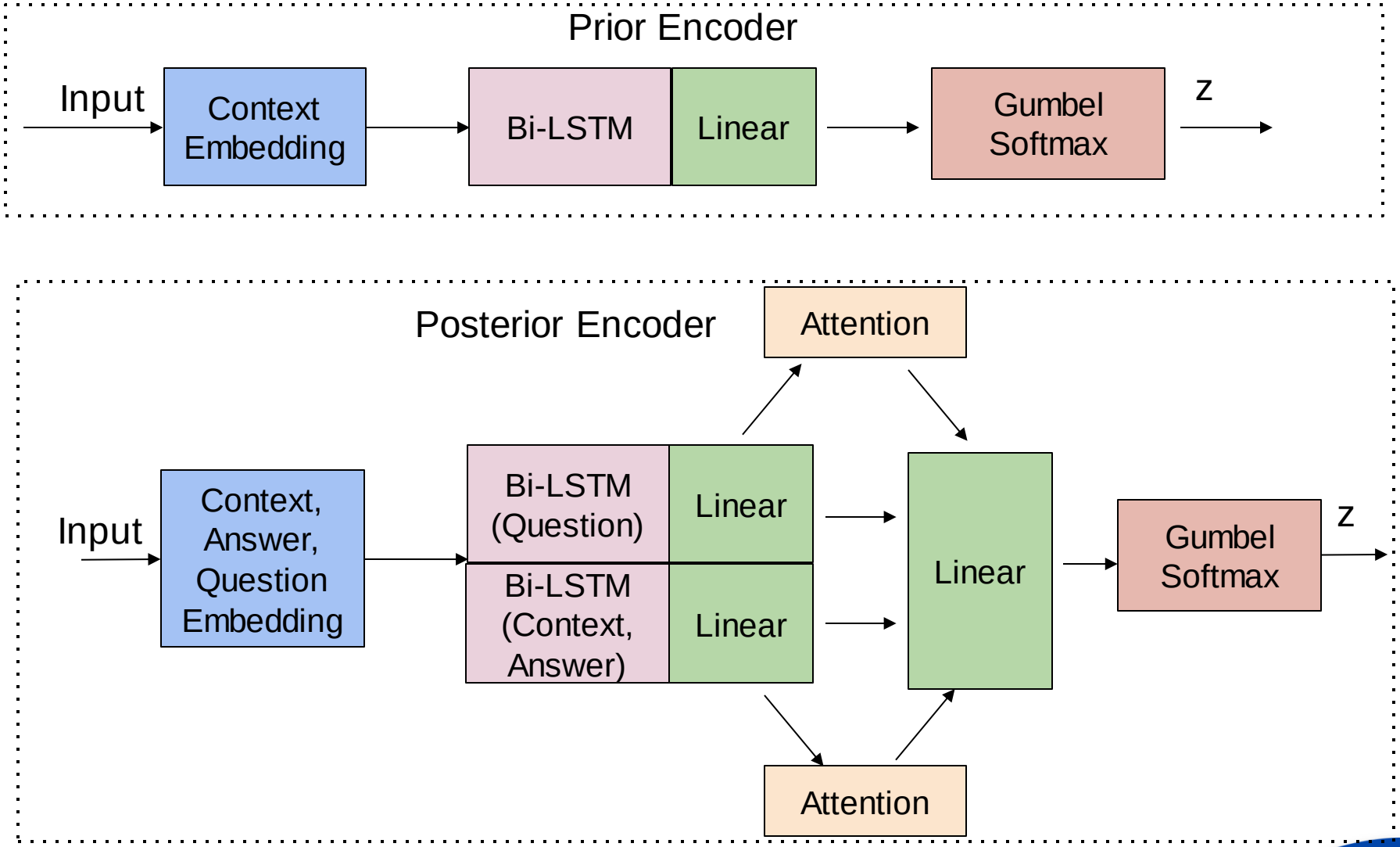
Proposed Method – QA Text Generator

- **VAE based QA text generator: Overall Diagram**



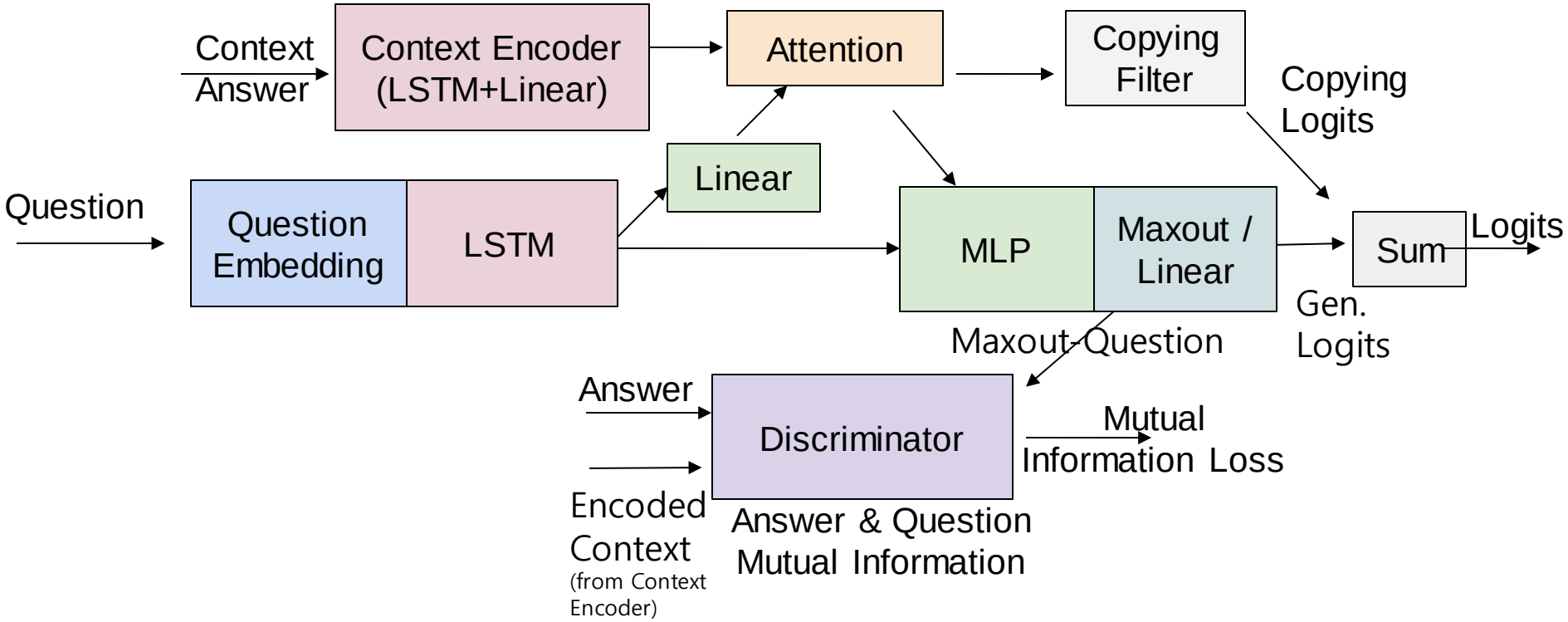
Proposed Method – QA Text Generator

- **VAE based QA text generator: Encoders**



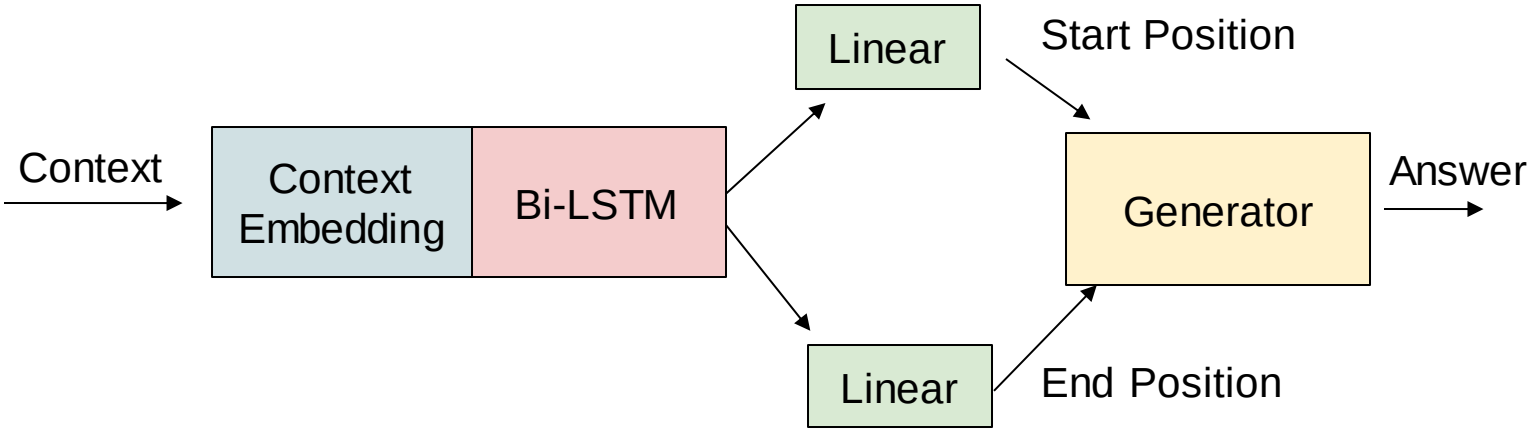
Proposed Method – QA Text Generator

- **VAE based QA text generator: Question Decoder**



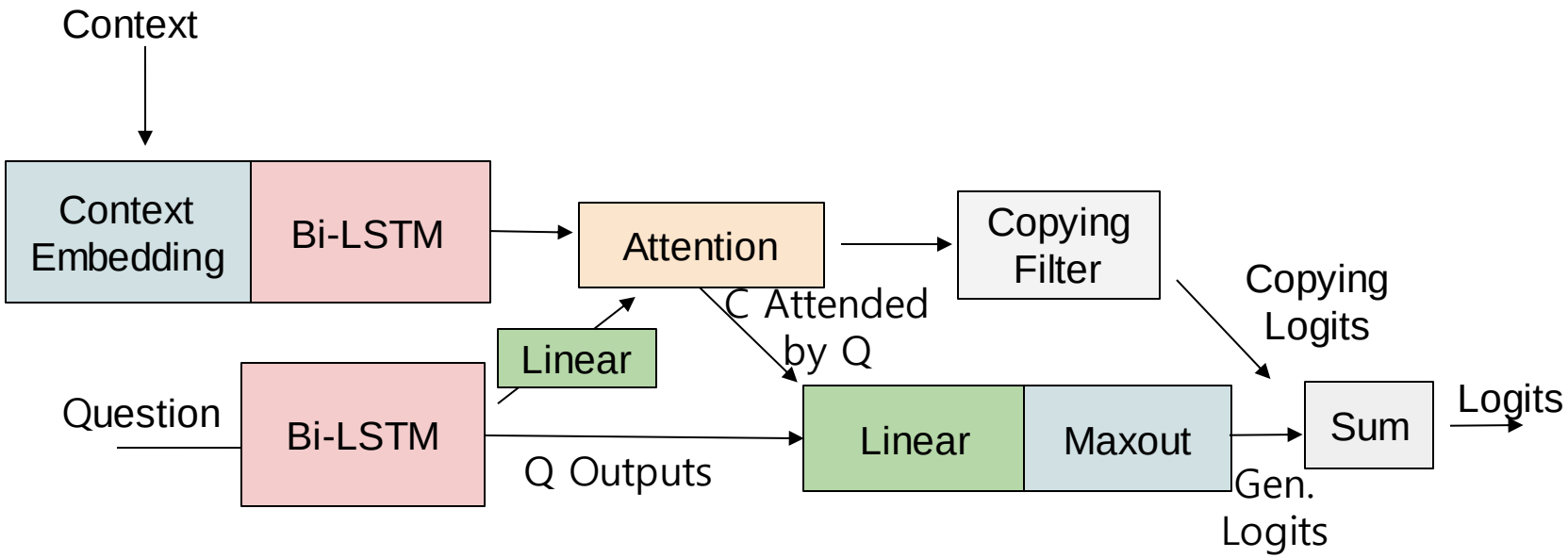
Proposed Method – QA Text Generator

- **VAE based QA text generator: Answer Decoder**



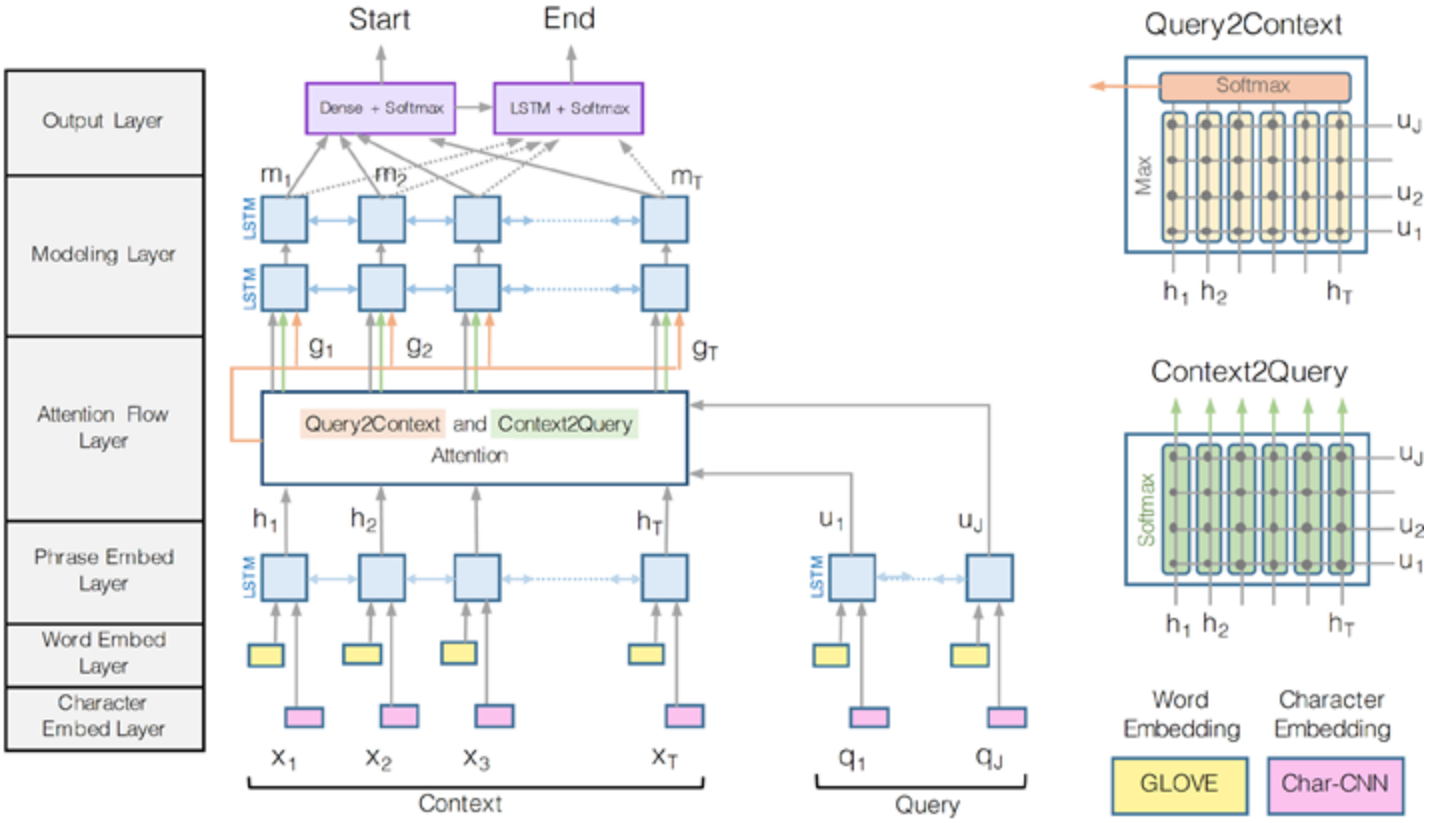
Proposed Method – QA Text Generator

- **Seq2Seq text generator**



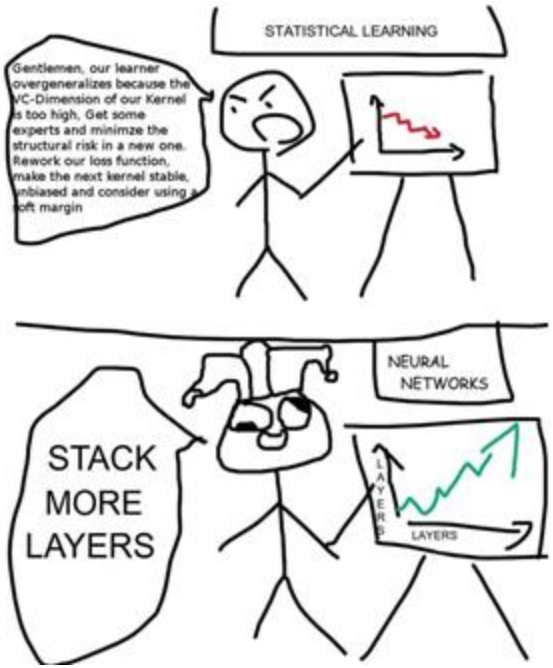
Proposed Method – QA Model

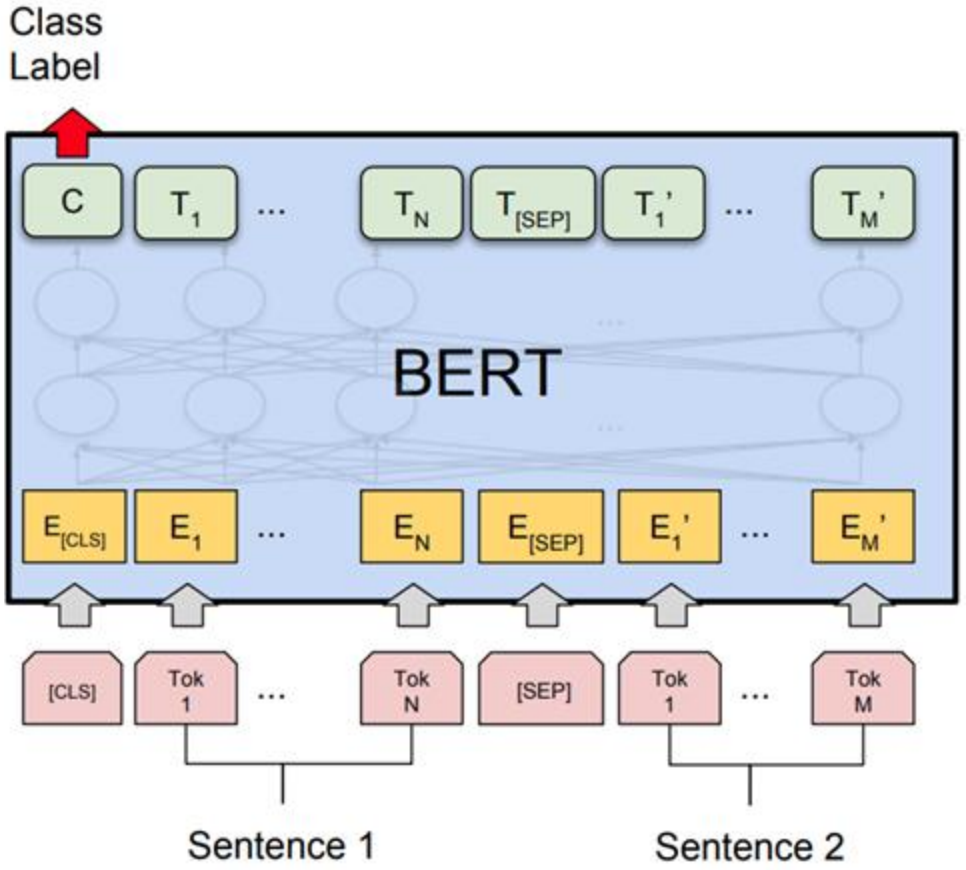
- **BiDAF Model**
  - Training 속도가 빠름
  - 성능은 BERT에 비해 낮으나 오히려 data augmentation의 효과를 확인하기 용이할 것으로 판단함



Proposed Method – QA Model

- **BERT Model**
  - Training 속도가 매우 느림
  - 성능이 높으나 data augmentation의 효과를 볼 수 있을지 불확실하였음

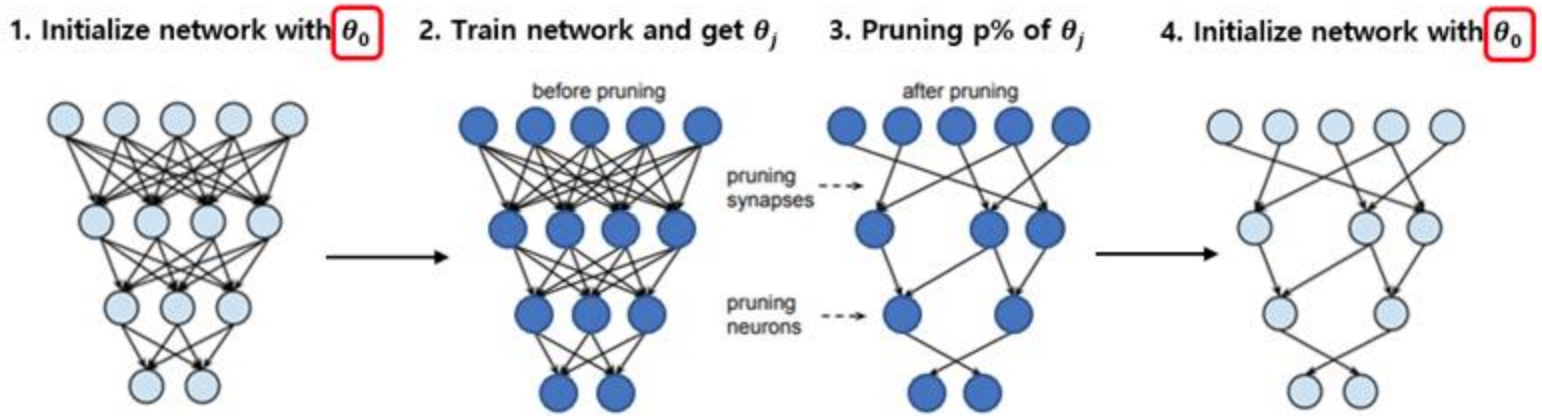




Proposed Method – QA Model

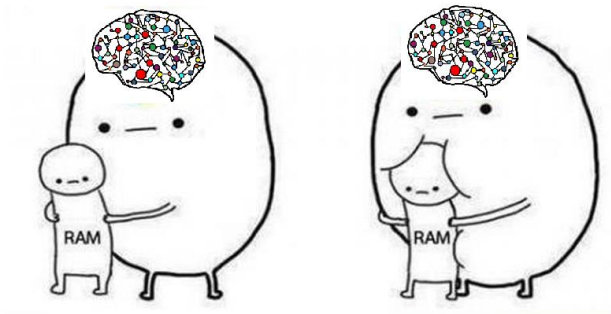
- **Lightweight BERT - Pruning**
  - BERT의 weight / node를 pruning함으로써 sparsity 증가
    → 효과적으로 수행할 경우 memory size 감소 가능
  - Low power가 중요한 edge device computing에서는 매우 중요한 기법이며 현업에서는 필수적으로 사용 됨

- **Iterative pruning 기법 사용**
  - Pruning → re-training → pruning을 반복적으로 진행



Experimental Results – Environment / Issue

- **Environments: Pytorch 기반 프로그래밍**
  - **BERT / BiDAF: nVidia V100 GPU**
    - BERT의 fine-tuning에 시간이 많이 소요됨
    - Memory 사용량 매우 큼



  - **Discrete-VAE / Seq2Seq: nVidia RTX2070 GPU**
    - VAE의 경우 decoder-encoder간 power balance에 중점적으로 tuning을 진행함
      → Architecture 개선 / hyper parameter tuning 진행
  - **Pruning: need too much time --> computing / memory hunger process**

- **Issues**
  - **Text generator의 generated data는 인간이 보기에 부족한 점이 많음**
    → Quality가 부족한 data를 추가로 사용하였을 때에도 효과가 있는지?
  - **BERT의 성능은 이미 QA 문제에 대해 overfit 수준의 성능을 보인다는 평이 있음**
    → Data augmentation이 효과가 있을 지?
  - **한국어 처리가 원활하게 될 것인지?**

Experimental Results – QA Generation

- **Seq2Seq Generator**
  - **Successful generation**

Successful Generation Example

- 김보름은 2012년 12월 22일 3000m 경기에서 몇 초의 대회신기록을 세웠는가 ?
- 18세기에 칠레 사회를 주도하고 있는 이민자들은 무엇에 들어갔는가 ?
- 맨체스터 시티가 현재의 구단 명칭으로 바뀐 년도는 ?
- 데이비드 베컴의 국가대표팀에서 최다 출전 기록을 보유한 선수는 ?
- 바이에른은 몇 년도에 프랑크 왕국에 점령되었나요 ?
- 아우슈비츠의 소장이었던 사람은 누구인가 ?
- 알로사우루스가 조상으로 여겨지는 과는 ?
- 샤이니가 MKMF에 이어 2연속 신인상을 수상한 것은 무엇인가 ?
- Carlill v . Carbolic Smoke Ball Company 판결에 대해 광고 전략이었을 것이라고 주장한 회사는 ?

Experimental Results – QA Generation

- **Seq2Seq Generator**
  - **Failed generation: mainly by repeated words**

Failed Generation Example

- 채무불이행으로 인하여 채무불이행에 대한 침해를 방지하는 것을 방지하거나 그러한 것은?
- 이영애가 처음 방송활동에 복귀한 이유는 몇년만에 복귀했을 때인가 ?
- 민주적 좌파의 거두를 체포한 민주적 좌파의 거두는 ?
- 최초의 기독교 박해가 최초의 기독교 박해가 된 것은 언제인가 ?
- 항공기가 납치된 후 항공기를 납치한 인물은 누구인가 ?
- 우리 역사임당의 어머니 이름은 ?

Experimental Results – QA Generation

- **VAE Generator**
  - **Successful generation**

Real Question and Answer

- Q: 임종석이 조사를 받은 뒤 인계된 곳은 어딘가 ?
- A: 서울지방경찰청 공안분실
- Q: 영생교에서 보는 사람의 실체는 원래 무엇이라고 여기는가 ?
- A: 신

Generated Question and Answer

- Q: 임종석이 농민 폭력행위같은 2월을 지도한 해는 ?
- A: 1989년
- Q: 영교는 사후 또는 무엇의 교리를 갖는가 ?
- A: 극락

Experimental Results – QA Generation

- **VAE Generator**
  - **Failed generation: mainly by repeated words**

Real Question and Answer

- Q: 노르트슐라이페의 랩타임 기록 갱신에서 실력을 인정받은 드라이버가 받는 칭호는 ?
- A: 링마이스터
- Q: 시리아가 국제 기구에서 심한 비난을 받고 바샤르 알아사드가 권력을 쥐고부터 10년 간 개선시키는 데 실패한 것은 ?
- A: 시리아의 인권

Generated Question and Answer

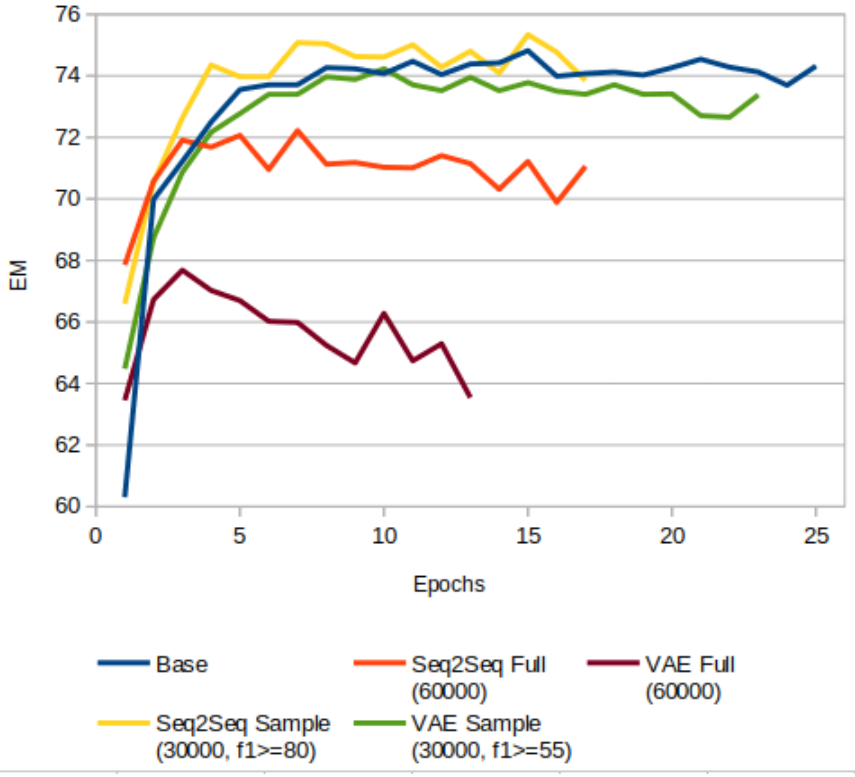
- Q: 포르알라이페의 차량 중 가장 빠른 기록이 있는 차량의 전체 순위는 ?
- A: 6분 47초
- Q: 인권 상태를 무엇이라 불렀는가 ?
- A: 시리아

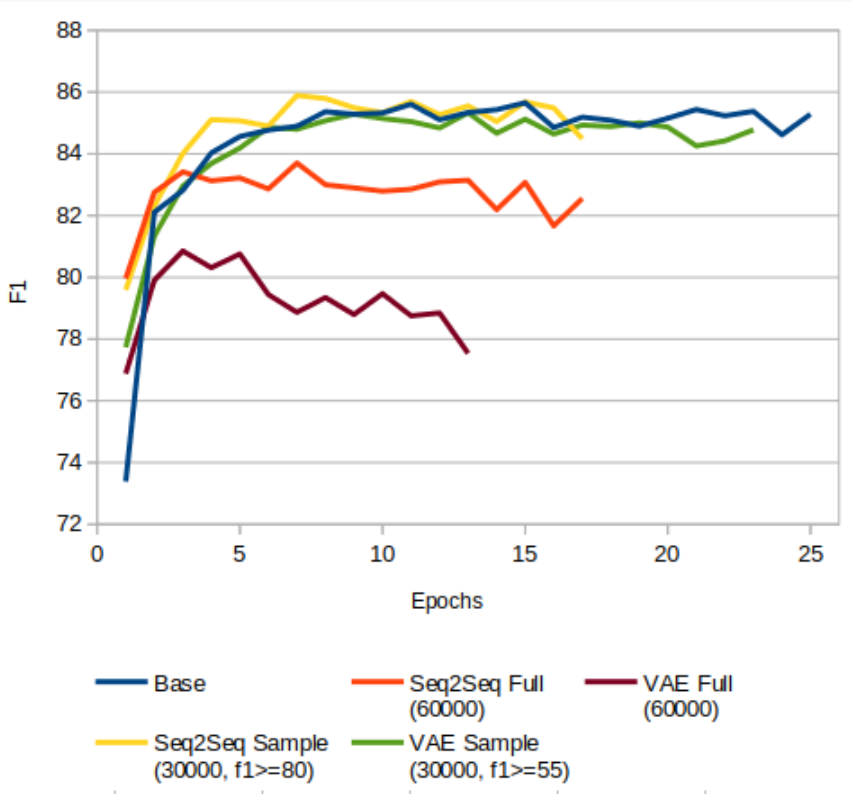
Experimental Results – QA Generation Comparison

- **Seq2Seq vs VAE generator**

Experimental Results – QA Model

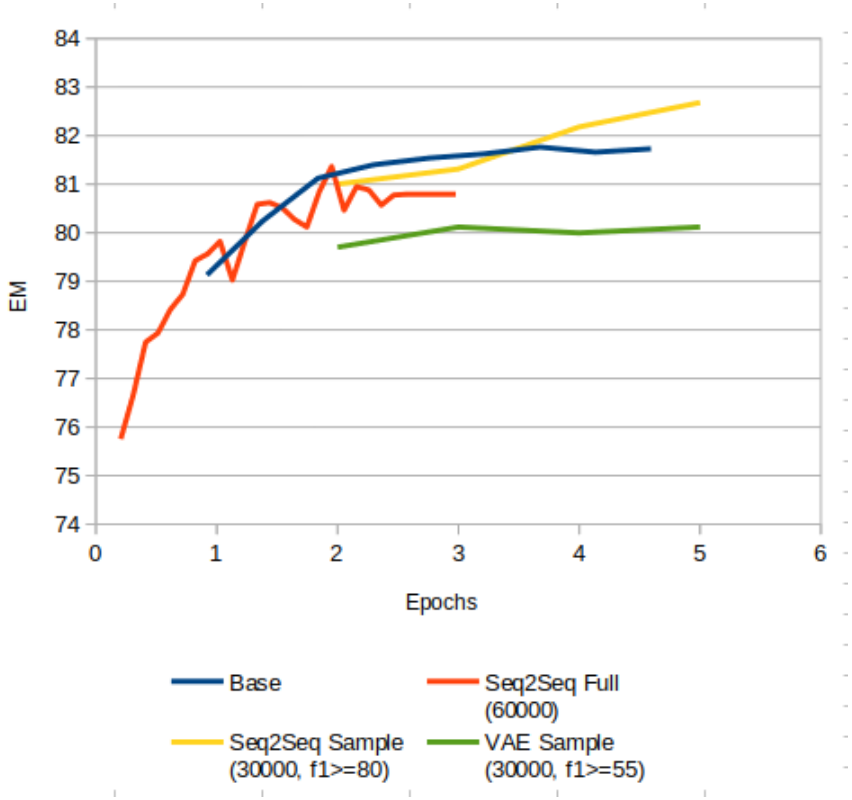
- **BiDAF**
  - **Data augmentation by Seq2Seq improves score!!**
    --> But data filtering is required

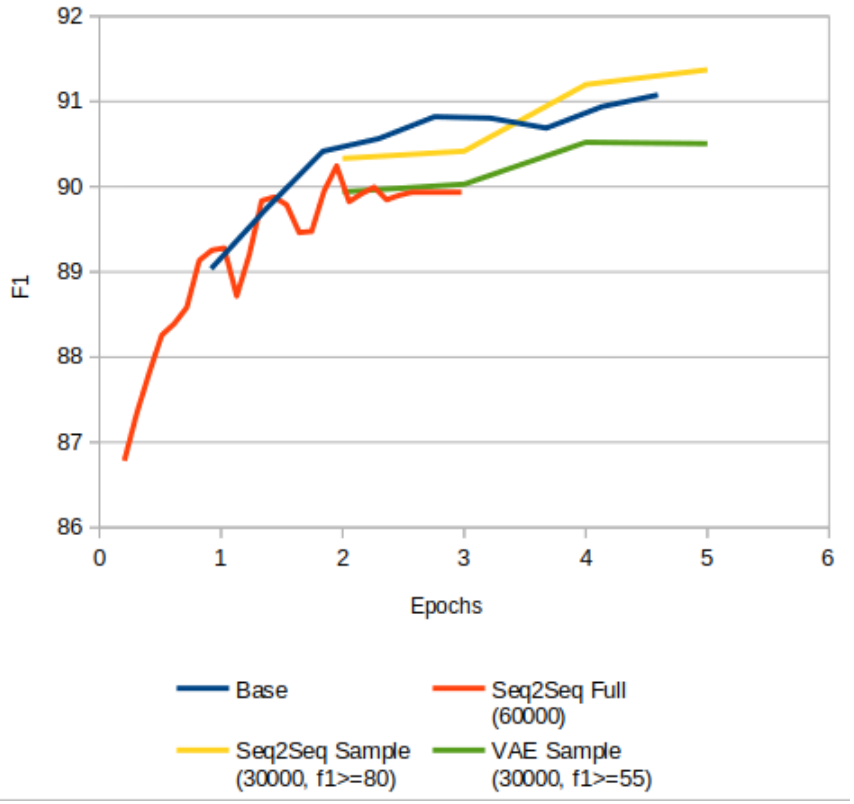




Experimental Results – QA Model

- **BERT**
  - **Seq2Seq augmentation also works**





Experimental Results – QA Model

- **Lightweight BERT**

Experimental Results – Model Pruning

- **BERT vs Lightweight BERT**

Summary

- 요약

- 느낀점