

내 스팀덱으로 로토무 만들기 Building Large Language Model at Scale



신정규
래블업 주식회사
@inureyes



BACKERJD.AI

lablup

안녕하세요!

• Lablup Inc. : Make AI Accessible

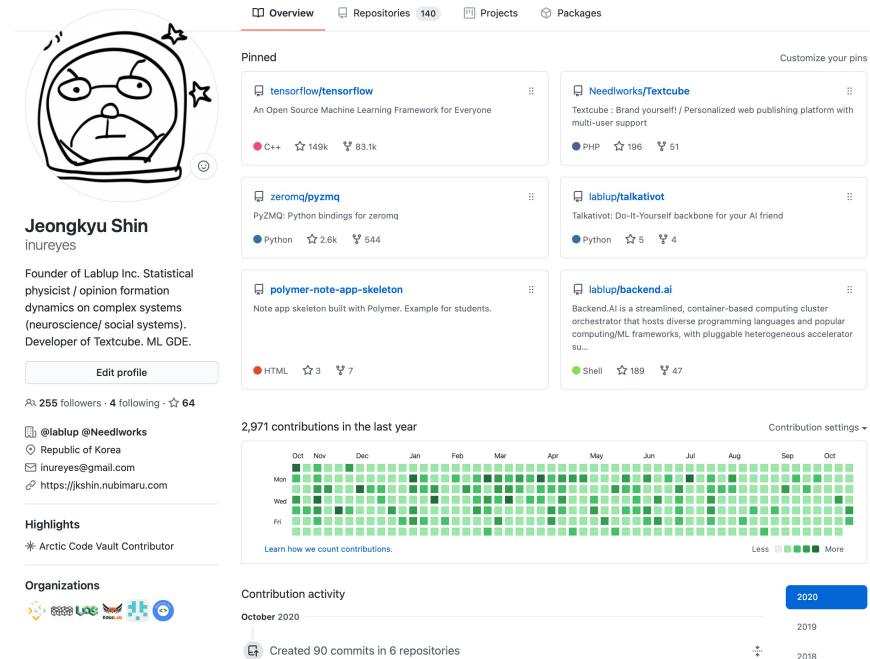
- 오픈소스 머신러닝 클러스터 플랫폼: Backend.AI 개발
- <https://www.backend.ai>

• Google Developer Expert

- ML / DL GDE
- Google Cloud Champion Innovator
- Google for Startup Accelerator Mentor

• 물리학 / 뇌과학

- 통계물리학 박사 (복잡계 시스템 및 계산뇌과학 분야)
- (전) 한양대학교 ERICA 겸임교수 (소프트웨어학부)



Building AI Chat bot using Python 3 & TensorFlow

Jeongkyu Shin
Lablup Inc.

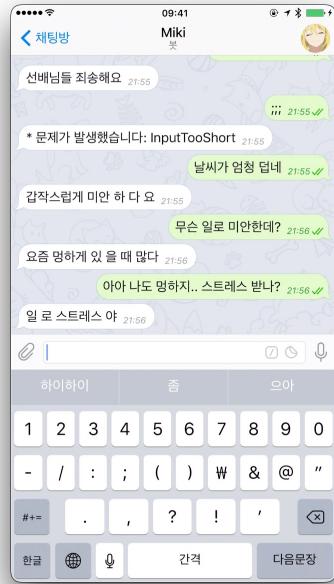


Illustration © Bandai Namco Games

PYCON KOREA 2017

Let Android dream electric sheep:
Making emotion model for chat-bot with
Python3, NLTK and TensorFlow

Jeongkyu Shin
Lablup Inc.

LAC

KossLab

한양대학교 HANYANG UNIVERSITY Back to the Basic

The screenshot shows a desktop environment with several open windows. One window is titled "PyCon Demo" and displays a browser page with the URL "127.0.0.1:3000". Another window shows command-line output for Python code, including error messages about TensorFlow compilation. A third window shows more command-line logs. To the right of the windows, there is a cartoon illustration of a blue cat-like character and a girl with blue hair and a white dress. A text box in the top right corner says "지금 마을 극복 하는 것이 중요해".





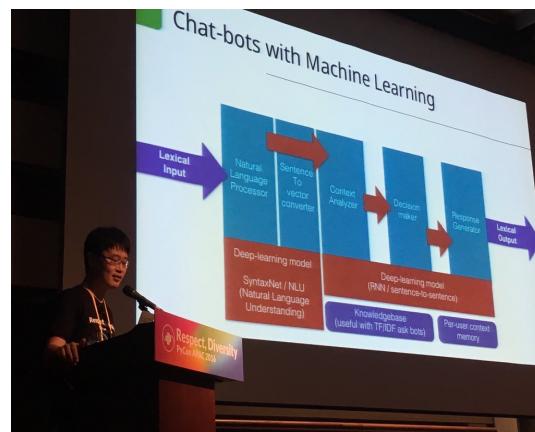
이 주제의 세 번째 시리즈





가지 않은 길

- 2015년에 래블업을 시작하고
- 2016년에 AI용 백엔드 솔루션과 함께 언어 모델을 사이드로 만들며
- 2017년까지 언어 모델에 관한 다양한 일들도 하고 새로운 시도도 했으나
- 2017년 하반기에 언어 모델을 접고 Backend.AI에 전념
 - 제 취미로만 남았죠
- 왜 취미로만 남았나?
- 2017년 9월
 - GDE Summit, 폴란드 크라쿠프 항공 박물관



5



구글 어시스턴트 팀 리드 및 열 명 남짓한 전세계 ML GDE들과
비공개 미팅 1시간동안 생각 후 접음.



집중해야 하는 일은 언어모델이 아니었음.

그 날 보고 들은 것

- 2017년 하반기 구글의 목표: 더이상 언어 모델 성능을 좋게 만드는 것이 아니었음
- 이미 끝난 문제
 - (내 삽질은 전부 그거였는데)
- 그럼 무슨 문제를 풀고 있었나?
 - 이번주 토요일에 랭콘 오시면 알려 드립니다.
- 그 후 일어난 일들

언어 모델: 2017~2018년

• 2017년

- 통계적 방법으로 7년간 만들어진 구글 번역 서비스의 성능을
- 4주 동안 인공 신경망을 번역에 도입하는 태스크포스팀의 실험 결과가 능가
- 두 달 후 기존 팀 해체 및 모든 번역 엔진 교체
- 1년 후 모바일에서 오프라인 번역을 인공신경망 기반으로 제공

• 2018년

- 번역기 개발 중, 언어쌍에 상관없이 공통된 인공 신경망 구조가 항상 생긴다는 것을 발견
- 언어 템플릿 신경망+추가적 훈련 = 번역기를 빠르게 만들 수 있음
- 언중이 만 명 미만인 언어의 번역기도 만들 수 있었음
 - ✓ 수백만 문장 쌍 -> 수 천 문장으로 줄어듦
- 이 과정의 부산물
 - ✓ **Transformer**, Universal Sentence Encoder, **BERT**, **Duplex**

언어 모델: 2019~2020년

• 2019년

- Transformer가 굉장히 일반적인 논리 구조를 만들 수 있음을 발견함
- "언어"가 무엇인가?에 대한 논의
 - ✓ 언어는 인간에게는 소통을 위한 도구이지만, **수학적으로는 연관된 정보를 논리에 따라 나열하는 방법**
 - ✓ "언어"를 잘하게 된다는 것의 의미가 무엇인가?
- XLNet, T5의 등장

• 2020년

- 논리 구조의 집중 포인트 차이
 - ✓ 정보를 투사하는 것이 중요한가? 정보를 최종적으로 표현하는 것이 중요한가? / BERT vs GPT
- GPT-3의 등장
- 수학적 접근: Transformer는 GNN의 특수 표현형?
 - ✓ GNN (Graph Neural Network, 2018)은 **대상의 관계를** 표현하는 그래프를 훈련하는 신경망
 - ✓ 2021년에 증명

거대 언어 모델: 2021~2022년

- **Zero-shot 번역 훈련**

- 아예 문장 쌍 데이터 없이 번역이 가능할까?
 - 24 언어 번역 모델을 zero-shot으로 개발 (Google, 2022)

- **OPT-175B (Meta, 2022)**

- 사전 훈련하여 공개한 모델 중 가장 큰 사이즈 (1750억 파라미터)
 - 모델 동작 시 Nvidia V100 16장 GPU 요구 (512GB) / 실제 동작시 사용 메모리는 약 350GB (A100 5장)
 - 모델 자체보다, 모델을 만들면서 고생한 모든 내용을 기록으로 남겨서 공개한 내용이 심금을 울림

- **Galactica (Meta, 2022)**

- 논문 작성 모델 (2022년 11월): 이런 일도 무난하게 할 수 있다!
 - 종종 오류를 내는 것으로 비판 받아 사흘만에 공개 종료
 - 전략의 실패...

- '거대' '언어' '모델' 이 그래서 뭐냐?
- 얼마나 거대하고
- 언어를 어떻게 처리하고
- 그걸 무슨 수로 서비스로 만드는가?
- 난 그걸 어떻게 해 볼 수 있을까?





더 이상의
자세한 설명은
생략한다.
23Langcon에서
이야기합니다.