2019 바벨TOP : 3월

일시 : 2019-03-30

장소 : 메가존(갈라빌딩 B1)

주제 : 언어/NLP와 관련된 최신 소식 & 연구 & 노하우 등을 공유하는 격월 미니 컨퍼런스

발표내용

1. 전창욱 / 자연어 처리 책 쓰기
2. 황명규 / elasticsearch 소개
3. 김무성 / 챗봇 학습용 유저 시뮬레이터 만들기
4. 김형락 / BERT 한국어 적용 및 실험 튜토리얼
5. 박민영 / 목표지향 챗봇을 위한 딥러닝 기반 대화상태 추적기(DST) 소개
6. 최석규, 정찬우 / Scholarly Big Data: NLP for Patent and Paper

주관:

- 바벨피쉬 페이스북 - https://web.facebook.com/groups/babelPish
- 싸이버스 페이스북 - https://web.facebook.com/thepsybus
- 메가존 - https://www.megazone.com/

행사 소개 : 바벨피쉬라는 페이스북 그룹에서 주최함, 자연어 처리 스터디 정도. 적어도 두달에 한번씩은 모이는 것을 목표로함, 돌아가면서 발표하는 형식의 스터디. 이번이 2회차, 1회차엔 언어심리학 등등을 했음.

자연어 처리 책 쓰기 - 전창욱

<발표>

제목 : 딥알못에서 자연어처리 저자까지

1. 월화수목금금금… 일의반복 하다 회사 안에서 여러가지 스터디를 개설함
   C,C++,Java,,,OS까지
2. 더이상 할게 없어서 2016 구글 hackfair에 감 MetaMong 텐서플로우, 딥러닝… 알게됨
3. 가벼운 마음으로 딥러닝 스터디에 참가함 -> 무거운 마음가짐+ 무거운 몸상태가 됨
4. 거울아 거울아 :
   a. 음성인식 구글API, 텐서플로CNN 으로 거울아거울아 만듬
   b. 음성을 인식해서 거울에 디스플레이가 뜬다.
   c. 사진촬영, 음악재생 등등
   d. 이를 가지고 다양한 공모전 참가, 수상
5. 창업! but 짤림 (스티븐 잡스)
6. 나를 돌아보기
   a. DeepNLP 부랩장
   b. 왜 자연어처리를 공부하고 있나…
   c. -> 자연어 처리 제대로 공부해보자
   d. -> 자연어 처리 책을 써보자
7. 자연어 처리를 공부하면서 우리가 가장 필요하다고 느낀 것으로 하자
8. 주말에 아들은 옆에서 공부하고 나는 책쓰고…ㅋㅋ
9. 기획서, 계획서.. 쓸게많다
10. 자연어 처리 공모전 -> 문장 자동완성 키패드 어플 만듬 -> 금상
11. http://www.yes24.com/Product/Goods/69334316
12. https://github.com/NLP-kr/tensorflow-ml-nlp
13. 지금 바로 시작하세요!

<느낌>

1. 대단하다
2. 책한번써볼까
3. 공모전 상받고싶다

elasticsearch 소개 - 황명규

<발표>

웹 개발자 출신, elasticsearch : 검색 엔진

1. elasticsearch 소개
    a. 현재 검색엔진중 최상위
    b. 오픈소스도 활발하다
2. simple things should be simple, complex things should be possible
3. 데이터 수집
    a. 라이브러리
        i. BEATS : 경량 데이터 수집기
        ii. LOGSTASH : 다양한 기능, 데이터 집계, 변환, 저장
    b. Agent
        i. Kafka
4. 출력
    a. machine learning
        i. 라이센스 수익의 대부분의 원인
5. 관리
    a. data freezing
        i. 날짜별로 인덱스 백업
        ii. 지정 기간 이후 인덱스는 snapshot 기능 활용하여 백업 후 삭제

<느낌>

1. 뭔소린지 잘 모르겠다…
2. 발표시간은 짧았다.
3. 오픈소스 참여해보고싶다

챗봇 학습용 유저 시뮬레이터 만들기 – 김무성

<발표>

제목: 사람같은 사람 만들기

1. 사람을 흉내내는 무언가를 만드는 이야기
2. 강화학습 기반 챗봇 만들기
3. 대화의 종류

a. 채팅
b. 목적 지향 대화
c. Q/A
d. 하나에 특화하는 경우가 많음

4. 챗봇의 종류
    a. 사람의 감정을 받아줄? 것이냐
    b. 대화의 질?을 높일것이냐
5. 구현방식
    a. 음.. 많음
6. 신경망 비슷하게 만들어 볼까?
7. 데이터를 어떻게 만들 것인가
    a. 사람끼리 대화해서 만들기 <- 비쌈, 모으는데도 한계
    b. 실제 머신과 상호작용하도록 <- 데이터의 현실적인 부분은 부족함 (ㅋㅋㅋ, 욕)
    c. M2M : user simulator self-play 기계끼리 시뮬레이션
8. 도메인마다 다른 어휘, 패턴, 시나리오

<느낌>

1. 음~ 잘 모르겠다.
2. 매우 실전적인? 내용이라 잘 와닿지 않는다..

BERT 한국어 적용 및 실험 튜토리얼 - 김형락

(영어 분석과 한국어 분석은 다르기 때문에..)

<발표>

1. NLP Deep Learning History
    a. RNN 구조
    b. LSTM 구조
    c. Seq2Seq모델 구조
    d. 어텐션 Seq2Seq 모델 구조
    e. transformer
        i. 선행 지식이 있을 때 더 잘 할수 있다. 의 아이디어
        ii. 셀프 어텐션 : 문장에서 특정 단어가 문장에서 얼마나 중요한지, 가중치 설정
        각 단어의 연관성 파악

iii. -> BERT가 나옴

2. BERT란?
    a. Transformer 모델 사용
    b. Bidirectional
    c. Sota(state-of the art)

<느낌>

1. 하핫; 모르겠당

목표지향 챗봇을 위한 딥러닝 기반 대화상태 추적기(DST) 소개 - 박민영

<발표>

1. Goal oriented Dialogue System?
    a. 사용자가 특정 task를 수행하는 데 도움을 주는 시스템
2. Modular system VS End-to-End System
3. Dialogue state tracker(DST)
    a. How about 6 pm? :
    b. I am busy at 6, book it for 7 pm instead
    c. 대화의 state 분석, 그 상황에 대한 후보와 확률 반환
4. Delexicalised Model VS NBT vs GLAD
    a. GLAD가 제일 좋은 성능을 기록함
5. Neural Belief tracker(NBT)
    a. 대화속에서 slot-value 쌍을 탐지하기 위해 설계된 모델
6. DNN < CNN
7. CNN 대신 BERT로 해보자 (사이드 프로젝트 : NBT with BERT)
8. word vector 를 concat 한 상태에서 BERT 적용
    a. 별로였다 ㅋㅋㅋ

<느낌>

1. 이런 사이드프로젝트 좋은거같다.

Scholarly Big Data: NLP for Patent and Paper - 최석규, 정찬우

<발표>

제목:NLP for scholarly Big Data

가천대 랩소속

1. Scholarly Big Data
    a. 새로운 연구가 쏟아져 나옴에 따라 생기는 엄청난 양의 학술지에 포하모디어 있는 정보 작가, 제목, 그림, 표 등등
    b. 연구분야 예측, 연구자 매칭, 내용 기반 분석, 논문 인용, 논문 요약, 논문 표절 등에 쓰임
    c. google 학술검색, 등등
    d. text, bibliography
2. patent landscaping
    a. 기술 개발에 있어 특허 침해를 방지하고, 트랜드를 추적하기 위해 특허를 분류하고 시각화하는 업무
    b. 한계점
        i. 특허 분류를 위한 전문인력 필요
        ii. 비효율적 프로세스를 반복적으로 실행
        iii. -> 시간적*금전전 비용 발생
        iv. -> 자동화 하려는 노력 -> Automated Patent Landscaping (2018, google)
    c. Contribution
        i. 자동화 특허 랜드스케이핑 프로세스 제안
    d. 특허 검색~ 빅쿼리 쿼리~ Pandas GBQ~ 했음
    e. 특허분류 -> GCN / Abstract Text -> Transformer
    f. BERT가 성능이 잘 안되더라.. -> 특허 특성상 키워드가 중요해서? 그런듯함, 대신 Transformer 사용
    g. 특허 분류 코드를 그래프 관점으로 보고 GCN을 썼다.
3. Experiment result
    a. 놀라운 성능이 나옴

b. 구글 모델보다 뛰어났음 (올ㅋ)
c. 분류 코드와 인용정보를 함께 사용하는 것보다 분류코드만 사용하는 것이 성능이 좋음 (자동차 논문에 수학, 물리라든지 관련적은 것들이 인용돼서 그런듯함)

제목: Paper recommendation

1. 연구에 대한 발표인듯
2. BERT가 좋다
3. 데이터셋이 별로 없어서 힘들었다

<느낌>

1. 집간다~
2. NLP가 이런곳에도 쓰이는구나