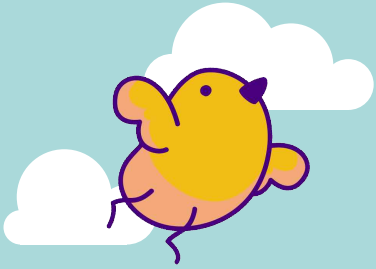
데이터 깎는 소상공인으로 살아남기

2 Mar. @LangCon 2024

조원익 (warnikchow)

목차



연사 소개

연사 소개

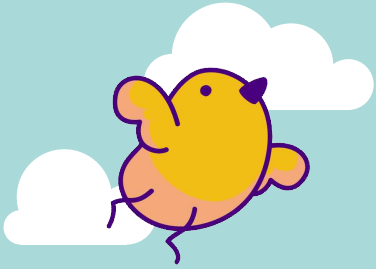
- 조원익
  - B.S. in EE/Mathematics (SNU, ’10~’14)
  - Ph.D. in EE (SNU EE/INMC, ’14~’22)
  - Staff Engineer (Samsung SAIT, ’22~)

- Interested in:
  - Computational linguistics
  - Computational social science

warnikchow

shallow and heuristic computational linguist

- Previous LangCon presentations:
  - 화행, 그 이론과 구축의 간극 (2020)
  - kosp2e - 공개 가능한 한국어 음성 번역 코퍼스 구축기 (2021)

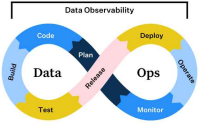


‘데이터 담당’이란?

데이터 담당

- 어떤 AI/ML조직에서 '데이터 담당'이라는 것의 의미
  - 도메인 데이터 분석?
  - 학습용 데이터 제작?
  - 대용량 데이터 관리?

- 데이터가 중요해지면서 등장한 다양한 표현들
  - Data managing
  - DataOps



https://www.montecarlodata.com/blog-what-is-dataops/

  - 기존의 Data scientist, Data engineer와 무엇이 다른가..?

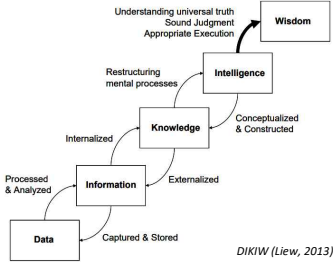
데이터에 관한 생각들

- **Data as ingredient of modeling**

- 데이터는 무조건 많아야 한다
- 데이터는 무조건 질이 좋아야 한다
- 이 데이터로 뭘 해야 한다
- 이 데이터를 보는 것이 재미있다
- **<u>이 데이터 자체가 어떤 데이터이다</u>**

- **Data as domain knowledge**

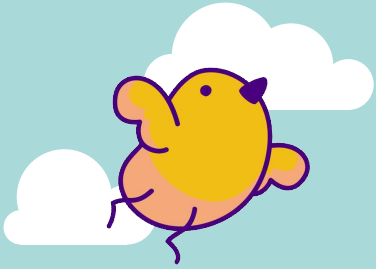
DIKW (Rowley and Jennifer, 2007)



DIKIW (Liew, 2013)

데이터 담당의 역할

- 데이터 담당에겐 기본적으로 어느 정도 기획자/모델러의 역할이 있다
- 좋은 데이터 담당이 되려면?
  - **데이터의 특성 파악**: 데이터 자체를 재미있어해야 함
  - **데이터 가공 목적**: 기획과 연결 - 이것으로 무엇을 할 것인가?
  - **데이터의 질**: 실제 좋은 데이터와 그렇지 않은 데이터를 파악
  - **데이터의 양**: 도구를 무엇을 쓸 것인가에 따라 다름
  - 기타:
    - 이 데이터 써도 괜찮나? Privacy, Harmful contents, License...
    - Machine? Human? -generated (or mixed)? 학습 측면에서 얻는 점이 있을지
    - NLP 데이터에서는
      - Text style (spoken, written, web...)
      - Language specificity (multilingual, code-switching, translated...)

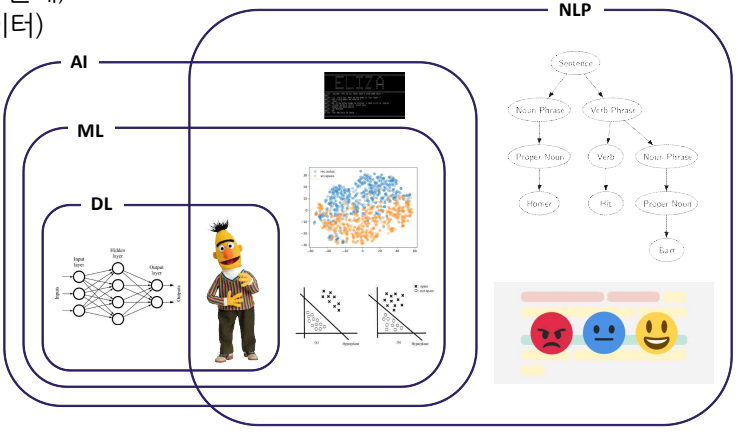


한국어 NLP 데이터 동향 변화

LangCon 2024

- NLP as...
  - Task (해결해야 할 문제)
  - Corpus (언어 데이터)
  - Algorithm (기술)

https://runestone.academy/ns/books/published/pythonds/Trees/ParseTree.html



NLP 데이터 동향 변화

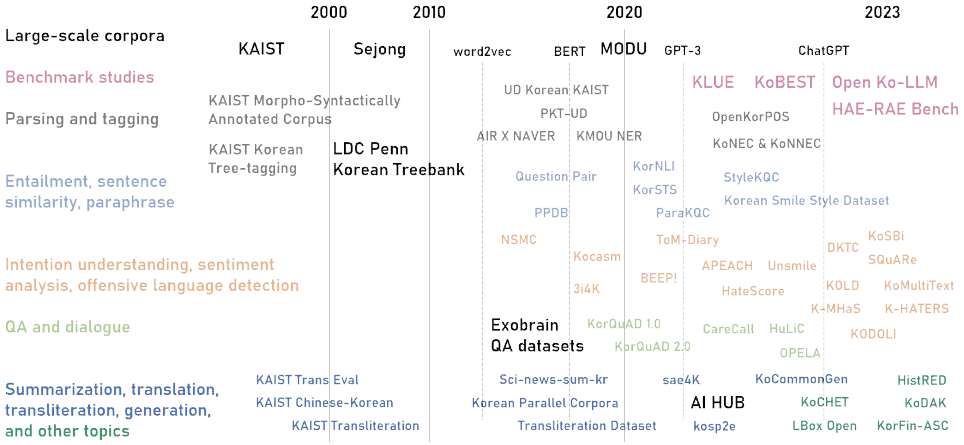
- Conventional한 data-driven NLP
  - Task가 정해짐
    - 어떤 model을 활용할 것인가? (pretrained model이 있나?)
      - 어떤 데이터로 훈련시킬 것인가? (fine-tune vs. from-scratch?)

- Model 발전에 따른 dataset 양상 변화
  - From-scratch → Pretrained LM → Prompt-based LLM
    - Model의 기본 능력치 상승
    - Task 이해를 위한 학습 데이터 필요성 감소

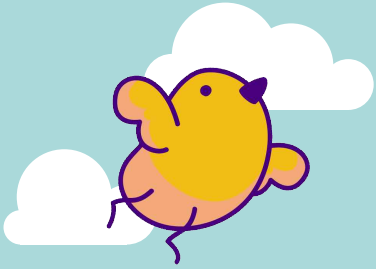
- 문제 해결과의 상관성: 더 이상 challenge되지 않는 task들은 무엇인지?

한국어 NLP 데이터 동향 변화

Language Conference 2019



Revisiting Korean Corpus Studies through Technological Advances (Cho et al., 2023)



LLM 시대의 NLP 데이터 구축

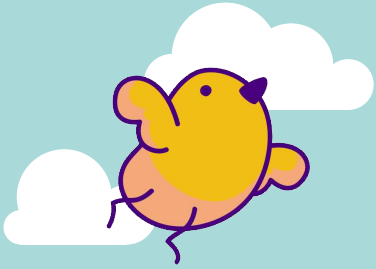
LangCon 2024

LLM 시대의 데이터 공개 양상

- Large-scale for pretraining/instruction tuning,
  Small scale for analysis or evaluation
  - Pretraining을 위해서도 다양한 domain/형식의 데이터 공개
  - Small scale로는 조금 더 focused goal을 위함
- LLM생성-휴먼검수/자동검수
  - 기존에 사람이 생성/사람이 검수하던 체제에서 기계 생성/사람 검수로 변화
- 번역데이터(대량)
  - 기존에 사람이 번역/사람이 검수하던 체제에서 기계 번역/사람 검수로 변화
- Model-dependent, multi-task, domain-specific etc.

LLM 시대의 개별 언어 데이터 공개 방향

- 결국 English 방향을 따라가겠지만
  이를 최대한 leverage하고 언어를 덜 타는 쪽에 집중하기
  - 언어를 덜 타는 쪽 - Semantics가 조금 바뀌어도 원래 task의 goal에 크게 영향 주지 않는 경우
    - Question answering
    - Multi-turn dialogue function calling
  - 언어를 타는 쪽 - 소상공인으로 할 수 있는 것?
    - 미묘한 표현 변화에 민감한 영역
      - SA, OLD, NLI, similarity 등
    - Cultural/Societal한 측면을 반영하는 연구
      - e.g., KoBBQ, KoSBi



마치는 글

데이터 깎는 노인

- 이미 있다

터를 분석하고 있는 노인이 있었다. 마침 우리 서비스도 스마트하게 만들어볼까 하여 우리 서비스의 로그 데이터가 담긴 USB 메모리를 들이밀며 분석해달라고 부탁했다. 값을 굉장히 비싸게 부르는 것 같았다.
"좀 싸게 해 줄 수 없습니까?" 했더니,
"이 정도 데이터 가지고 에누리 하겠소? 비싸거든 다른 데 가서 하시우."
대단히 무뚝뚝한 노인이었다. 값을 흥정하지도 못하고 잘 분석해 달라고만 부탁했다.

그는 잠자코 열심히 분석하고 있었다. 처음에는 빨리 분석하는 것 같더니, 몇 시간이 지나도록 이리 돌려 보고 저리 돌려 보고 궁뜨기 시작하더니, 마냥 늑장이다. 내가 보기에는 그만하면 다 됐을 것 같은데, 자꾸만 코드를 수정하고 그래프도 더 띄워보고 있었다. 나에게 우리 서비스의 내용과 현황에 대해 이것저것 캐묻기도 했다. 심지어 R Studio 까지 띄우고, 여기저기에 전화를 걸어 정체를 알 수 없는 사람들과 이야기를 하기도 했다. 인제 다 됐으니 그냥 지금까지 한 결과만 달라고 해도 통 못 들은 척 대꾸가 없다. 차 시간이 빠듯해 왔다. 갑갑하고 지루하고 조조할 지경이었다.

"더 분석하지 않아도 좋으니 그만 정리해서 주십시오." 라고 했더니, 화를 버럭 내며,
"끓을 만큼 끓어야 밥이 되지, 생쌀이 재촉한다고 밥이 되나." 한다. 나도 기가 막혀서,
"해 달라는 사람이 좋다는데 무얼 더 한다는 말이오? 노인장, 외고집이시구먼. 차 시간이 없다니까요."
노인은 퉁명스럽게,
"다른 데 가서 하시우. 난 결과 안 주고 돈도 안 받겠소."
하고 USB 메모리를 MacBook에서 확 뽑는다. 지금까지 기다리고 있다가 그냥 갈 수도 없고, 차 시간은 어차피 틀린 것 같고 해서, 될 대로 되라고 체념할 수밖에 없었다.

https://inforience.net/2017/04/23/oldman/

마치며

- 데이터 작업의 본질은 데이터의 이해에 있다 (tool에 넣을 재료도 잘 다듬어야 하고, 실제로 무언가 가치를 창출하기를 원한다면 이해해야 한다)
- Domain 전문가, 데이터 담당, 모델러의 브릿징은 무슨 모델을 쓰든 (쓰지 않아도) 중요하다
- 데이터 담당은 data in-out/modeling 모두 고려해야 하는 기술 '기획자'이다
- 한국어 특화로 무언가를 해야 한다면, model, domain을 모두 잘 보고 최대한 있는 것을 leverage하자 - 그리고 한국어 특화가 되는 부분에 집중하자