**Weekly report**

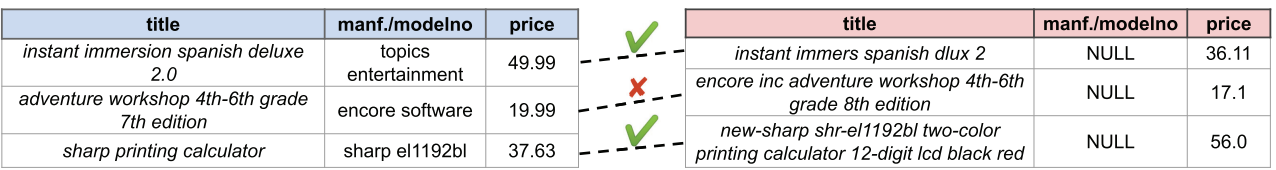
날짜 : **2023-03-02**

연구원 : **최윤석**

* **이전 수행계획**
  + **관계형 테이블 임베딩 :**
    - Value overlap + Semantic overlap + Embedding(EmbDI) 을 사용하여 테이블 융합이 가능한지 확인하는 나만의 아키텍처 연구
    - Creating Embeddings of Heterogeneous Relational Datasets for Data Integration Tasks을 인용한 논문 중 연구와 관련된 것 읽어보기 (Embedding + Semantic Overlap(외부 knowledge base)을 섞는 방식 관련해서)   
      -> SANTOS: Relationship-based Semantic Table Union Search (외부 Knowledge Base를 사용하여 Table Unionability를 알아내는 알고리즘 제시)  
      -> Deep Entity Matching: Challenges and Opportunities (Pretrained된 language model과 딥러닝을 활용해 Entity Matching을 진행)
    - 테이블 간 조인 / 유니온 하기위한 컬럼을 찾는 문제에 관해 생각 (Elmo / Bert / Word2Vec) 등도 생각 / 내가 손으로 하는 것을 자동화하는 방법으로 생각해보기
    - SemProp(Attribute overlap + Value overlap + Embedding) 문제점을 찾고 문제점을 타개할 해결방안 찾기, 큰 부분은 모방해도 되지만 독창성 있는 아이디어로 나의 연구를 할 것
  + **GAN 논문 세미나 :**

**CycleGAN | Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks** 논문 읽고 세미나 준비

* **수행결과**
* **관계형 테이블 임베딩** : (임베딩 / 융합 가능성 파트 분류해서 작성)
  + Deep Entity Matching: Challenges and Opportunities 읽고 정리

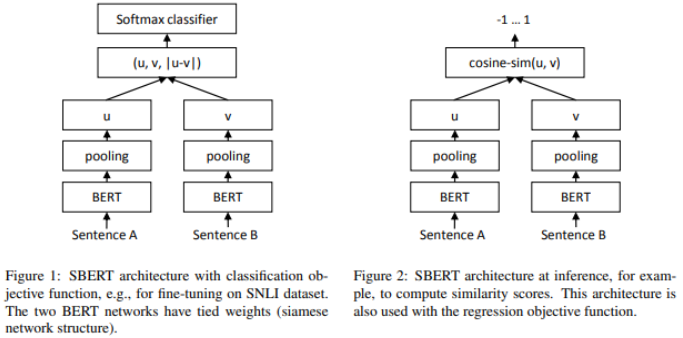


* + - 위 그림은 두 테이블간 Entity Matching(EM)을 한 결과, 매칭되는 레코드 쌍을 출력함
    - 두 레코드가 매칭되는지 판단하는 기준은 두 레코드의 similarity를 기준으로 진행함
    - Similarity는 pre-trained language models(LMs) (BERT, RoBERTa, DistillBERT)을 활용해 알아냄
    - 논문에서는 LMs를 사용하는 EM 시스템 DITTO를 제시함

텍스트, 시계, 장치, 스크린샷이(가) 표시된 사진

자동 생성된 설명

* + - DITTO의 구조는 위의 그림과 같음
    - Blocker의 역할은 매칭이 되지 않을 확률이 높은 레코드쌍들을 미리 제거해 EM시간을 단축시키는 것임
    - Blocker의 Blocking을 위해 먼저 Sentence-BERT를 라벨링된 데이터(레코드쌍과 매칭가능여부)를 활용해 추가로 학습시킴.



* + - 위의 그림은 Sentence-BERT의 구조에 대한 그림
    - 왼쪽의 구조에 입력으로 두 레코드를 넣으면, 각 레코드에 포함된 단어들의 임베딩 벡터를 출력하고 각 임베딩 벡터들의 값을 평균내 그것을 문장의 임베딩 벡터로 함.
    - 각 문장의 임베딩 벡터들과 벡터들의 차의 절댓값을 하나의 벡터로 묶어 소프트맥스 함수를 통과시키면 매칭가능 / 매칭불가능을 출력하고, 정답라벨과 비교해 학습을 진행함
    - 학습된 모델을 가지고 Table A와 Table B에 있는 모든 레코드들에 대해 레코드 쌍을 만들고, 레코드 쌍을 오른쪽 구조에 넣어서 벡터화 시킴.
    - 벡터들의 코사인 유사도를 통해 Table A의 한 레코드와 가장 유사한 k개의 Table B의 레코드를 추림. 이 Blocking을 통해 EM을 진행할 레코드쌍의 개수를 대폭 줄일 수 있음.
    - serialize는 레코드를 다음과 같이 바꿔줌





* + - 이후 serialize된 두 개의 레코드를 위와 같이 붙임
    - Inject DK는 domain knowledge를 직접 추가하고 특정 타입에 대한 포맷을 지정해 모델의 성능을 높임



* + - 위와 같이 [MODEL NO] 같은 태그를 추가해 해당 부분을 모델이 EM을 진행할 때 참고하게 도와줌
    - 또한 특정 타입에 대한 포맷을 지정해줄 수 있음. 예를 들어 휴대폰 번호는 XXX-YYYY-ZZZZ의 형식을 사용하게 하고, 퍼센티지는 5.00%처럼 소수점 아래 둘째 나타내는 형식을 사용하게 지정할 수 있음.
    - 이러한 포맷 지정은 DITTO가 수치형 데이터에 대해 잘 매칭할 수 있게 해줌.
    - Summarize는 레코드의 길이가 너무 길어서 발생하는 문제를 해결하기 위해, 레코드 내에 필요 없는 데이터를 없애 레코드의 길이를 줄이는 데 사용함.
    - Summarize로 없애는 데이터를 선정하기 위해 DITTO는 TF-IDF를 사용함.
    - TF-IDF는 TermFrequency(TF)와 Inverse Document Frequency(IDF)의 곱으로 나타내지며, TF는 한 문서에서 특정 단어가 나온 횟수를 의미하고, IDF는 DF의 역수에 로그를 취한 것임. DF는 한 단어가 몇 개의 문서에 나타났는지를 의미함.



* + - Idf는 DF에 1을 더한 값을 분모에 넣고 문서의 개수 D를 분자에 넣어 자연로그를 취해준 값임.
    - TF-IDF의 값이 큰 단어일수록 특정 문서내에서 중요한 단어임.
    - Augment는 training data를 augment해 DITTO 모델의 성능을 높이는 방법임.
    - Augment는 span(token들의 집합)-level / attribute-level / entity-level로 나뉘어짐
    - span의 순서를 바꾸거나 일부를 삭제하거나 attribute의 순서를 섞거나 일부를 삭제하거나 serialize 된 레코드 순서를 바꾸는 것으로 augment를 진행함
    - 최종적으로 augment된 데이터들과 기존 데이터들을 모델에 넣어 매칭되는 레코드쌍을 출력으로 얻음
* **GAN 논문 세미나 :**
  + **CycleGAN | Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks** 논문 읽는 중
* **수행계획**
  + **관계형 테이블 임베딩 :**
    - Value overlap + Semantic overlap + Embedding(EmbDI) 을 사용하여 테이블 융합이 가능한지 확인하는 나만의 아키텍처 연구
    - Creating Embeddings of Heterogeneous Relational Datasets for Data Integration Tasks을 인용한 논문 중 연구와 관련된 것 읽어보기 (Embedding + Semantic Overlap(외부 knowledge base)을 섞는 방식 관련해서)   
      -> SANTOS: Relationship-based Semantic Table Union Search (외부 Knowledge Base를 사용하여 Table Unionability를 알아내는 알고리즘 제시)
    - 테이블 간 조인 / 유니온 하기위한 컬럼을 찾는 문제에 관해 생각 (Elmo / Bert / Word2Vec) 등도 생각 / 내가 손으로 하는 것을 자동화하는 방법으로 생각
    - SemProp(Attribute overlap + Value overlap + Embedding) 문제점을 찾고 문제점을 타개할 해결방안 찾기, 큰 부분은 모방해도 되지만 독창성 있는 아이디어로 나의 연구를 할 것
    - DITTO의 저자가 깃허브에 올린 DITTO의 소스코드 실행 및 이해
  + **GAN 논문 세미나 :**
    - **CycleGAN | Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks** 논문 읽고 세미나 준비 (PPT 30분 이내 발표) , 각 논문의 핵심사항을 알아내는게 목적. 연구 동기, 배경, 관련연구 등은 생략하고 이해한 범위 내에서 핵심사항 중심으로 발표 진행. 성능 평가는 한 페이지(메인 그래프만)로 충분함. 그림과 키워드 위주로 발표. 예제 위주 이해 위주 친절한 자신만의 이해한 언어로 진행
* **기타사항**
  + **리포트 작성 시, 논문의 내용정리 같이 남이 한 부분에 대한 설명은 자세하게 하지 말고 그 중에서 내가 얻은 아이디어나 내 연구에 관한 부분을 자세하게 설명할 것**
  + **데이터 융합 프로세스**



* **융합 가능도 척도**
  + 임베딩 거리 기준 값을 정하고 기준값보다 임베딩 거리가 가까우면 융합 가능한 데이터셋이라 판단함
  + 융합 가능한 데이터셋을 사람이 하나씩 확인하여 유의미한 융합이 되는 데이터셋을 찾음
  + 을 융합 가능도 척도라 정의함



혼합형 데이터셋 예시

등급의 semantic

융합 가능도 딥러닝 아키텍처 : 임베딩 벡터 + 메타데이터(융합 가치를 정량화할 수 있는 무언가)   
조인할 수 있는 컬럼이 필요함(A : 주민번호, B : 주민등록 번호로 조인이 가능한 테이블이면?) 융합을 하는 건 가치가 있기 때문, 가치가 있다? -> 임베딩이 가깝게 되어있다.

조인할 수 있다, 없다 -> 사람이 판단

이걸 조인해서 가치있는 정보가 나오나? -> 모델이 판단

Id를 가지는 컬럼이 의미는 같지만 겉모양이 다를때

그래프 임베딩

정형테이블을 그래프로 바꾸는 논문읽기

그래프 정보를 임베딩한 정보와 기존 윤종찬 연구원과의 차이점 발견

나만의 새로운 패러다임 찾기

데이터셋은 영문 데이터만 사용

융합 가능성 평가는 임베딩 성능을 확보한 후에