**Weekly report**

날짜 : **2023-02-23**

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) 관련 논문 이해하고 정리
  + **GAN 논문 세미나 :**

**StyleGan | A Style-Based Generator Architecture for Generative Adversarial Networks** 논문 읽고 세미나 준비

* **수행결과**
* **관계형 테이블 임베딩** : (임베딩 / 융합 가능성 파트 분류해서 작성)
  + SemProp 논문 읽고 정리
    - SemProp은 syntactic과 semantic similarities를 기반으로 관련있는 두 데이터셋 사이에 링크를 만들어 Data Discovery를 쉽게 하기위해 고안됨.

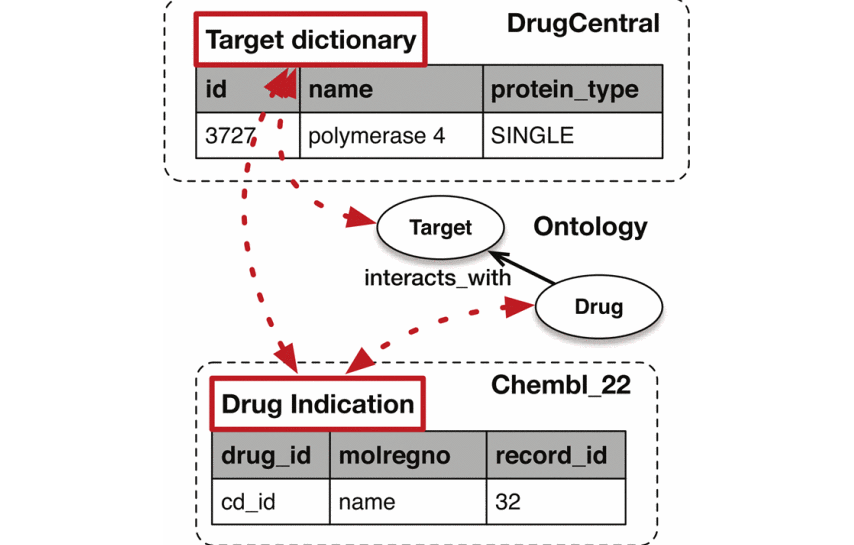


Figure . 두 테이블간 링크를 Reference data(Ontology)를 참고하여 생성하는 예시

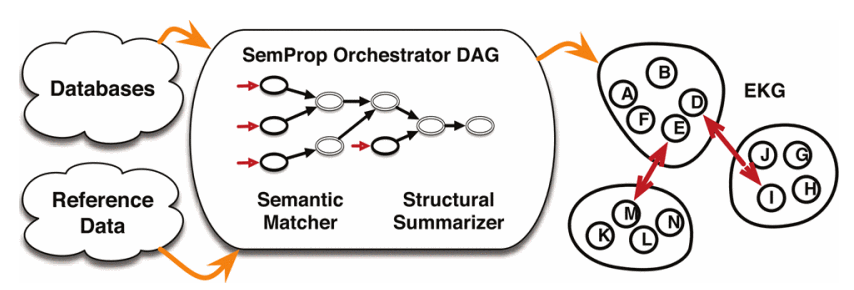


Figure . SemProp Overview

* + - Databases 에 해당하는 데이터들은 Source Elements(SE)라고 부르며,  같이 나타냄. N = all relations A = all attributes. Drug Central의 source elements는 {Target dictionary, id, name, protein\_type}임
    - Reference Data는 각 도메인에서 기준으로 간주되는 것에 대한 정보임. Ontology(클래스들의 정보와 각 클래스들의 관계정보, Drug, Protein, Target등이 클래스에 해당함)나 두 attributes가 같은 것을 나타내는 mapping table이 이에 해당됨.
    - Semantic Matcher에서 Semantic similarities는 word embedding을 통해 알아냄
      * 다만 기존의 word embedding을 그대로 사용하지 않고 coherent group을 이용한 word embedding을 사용함.
        + coherent group은 기존의 word embedding이 가진 두 가지 문제를 해결하기 위해 사용함
        + 1. word embedding은 단일 단어를 학습하는데 SE는 대부분 여러 단어로 되어있음.
        + 2. 찾고자 하는 단어가 word embedding dictionary에 없음
        + 핵심 아이디어는 모든 단어 쌍 간의 평균 유사성(임베딩에서)이 높으면 단어 그룹이 다른 단어 그룹과 유사하다는 것임.
    - Coherent group인지 판단은 coherency factor가 일정 수준 이상인지로 판단함.
      * 텍스트이(가) 표시된 사진

        자동 생성된 설명  
        coherency factor 식
      * SE에 속한 원소 하나가 예를 들어, Drug Description이라 하면, 이것은 Drug의 vector와 Description의 vector 두 개의 벡터로 임베딩해 나타낼 수 있다.
      * 는 각각 임베딩된 단어 벡터 집합 X에 속한 하나의 벡터
      * 모든 단어 쌍의 코사인 유사도를 평균내서 구함
      * 같은 coherent group에 속한 단어들은 유사한 의미를 가진다고 볼 수 있음
    - Semantic Matcher로 SEMA를 사용함.
      * 두 elemenet가 coherent group이면 SEMA(+), 아니면 SEMA(-)
    - Syntatic similarities는 SynM을 사용해 파악함. SynM은 인스턴스와 스키마(Attribute and Table name)의 유사성을 jaccard similarity로 판단해 두 테이블이 Syntatic적으로 유사한지 판단함.
    - 최종적으로 여러 테이블간의 관계를 나타내는 Enterprise Knowledge Graph(EKG)를 생성함

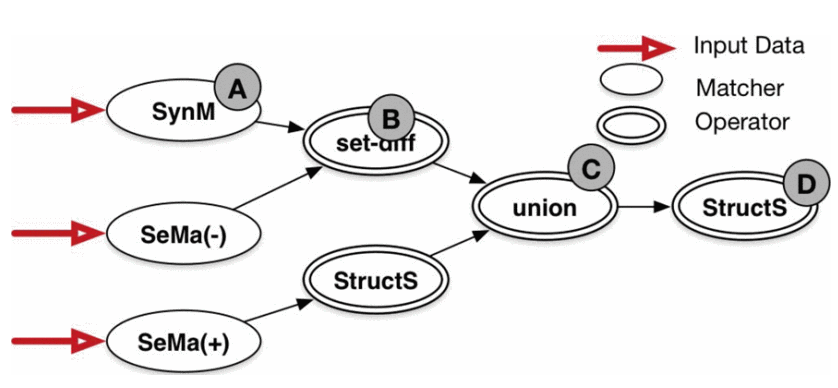


Figure . SemProp Orchestrator DAG의 자세한 그래프

* + - DAG는 Directed Acyclic Graph의 줄임말로 방향이 있는 비순환 그래프를 의미함.
    - 먼저 SynM을 통해 스키마 기반 유사한 것들의 링크를 만듦
    - 그 중 SEMA(-)인 즉 의미적으로 유사하지 않은 것들은 차집합 계산을 통해 제거함
    - SEMA(+) 즉 의미적으로 유사한 것들 사이에 링크를 구함
    - StructS는 Structural Summarizer로 잘못 생성된 링크나 불필요하게 많이 생성된 링크를 제거해줌
    - SEMA(+)에서 생성된 링크중 불필요한 것을 제거하고 차집합 했던 링크들과 합집합 연산을 해줌.
    - 마지막으로 StructS에서 잘못 생성된 링크나 불필요한 링크를 제거해줌.
* **GAN 논문 세미나**
  + **StyleGan | A Style-Based Generator Architecture for Generative Adversarial Networks** 논문 읽고 세미나 준비완료
* **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 (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**
  + **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