Autoknow:
Self-driving knowledge collection for products of thousands of types

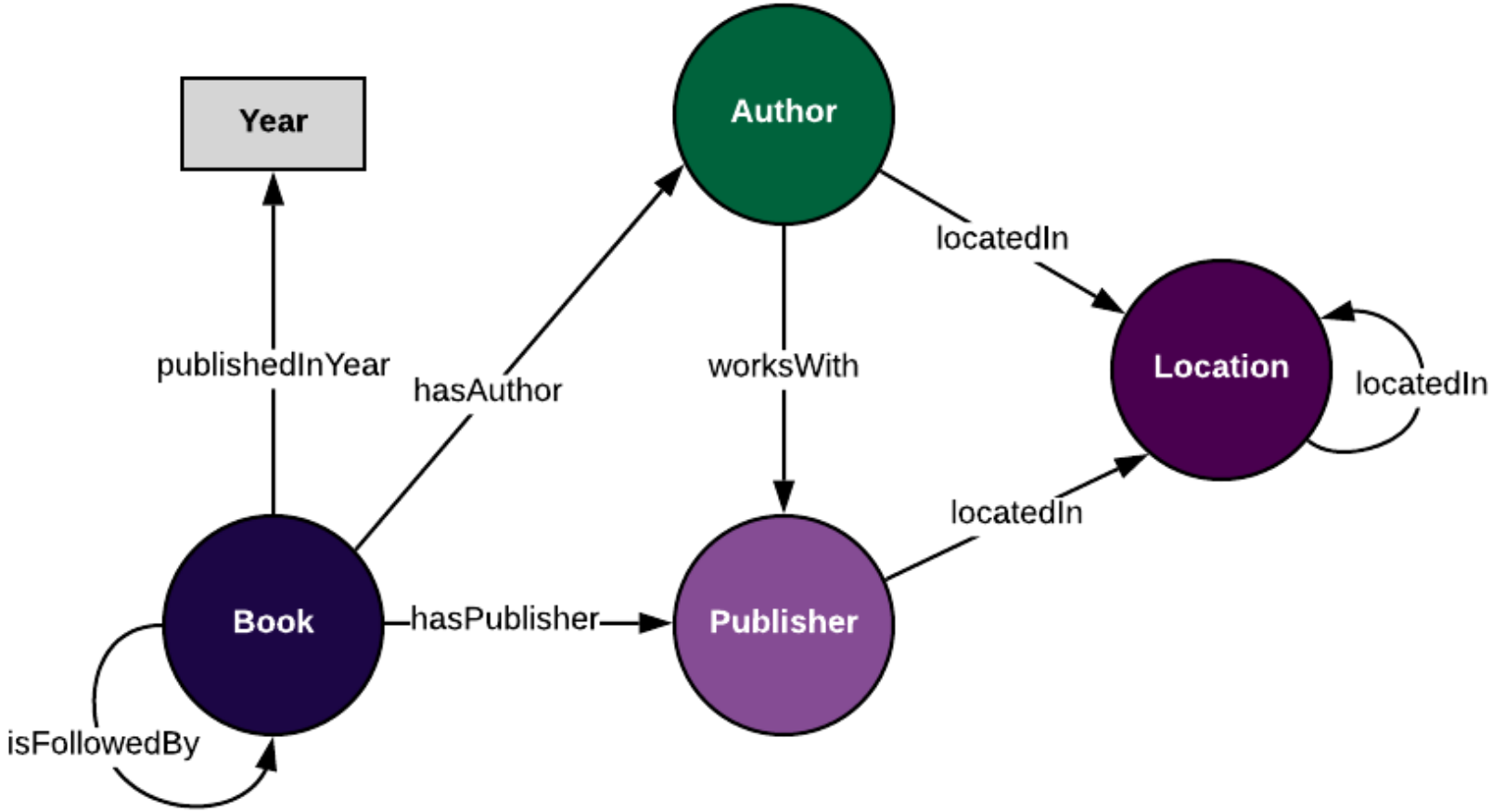
KDD 20

- **연구 목적**

"고품질의 포괄적인 Knowledge Graph 생성 자동화"

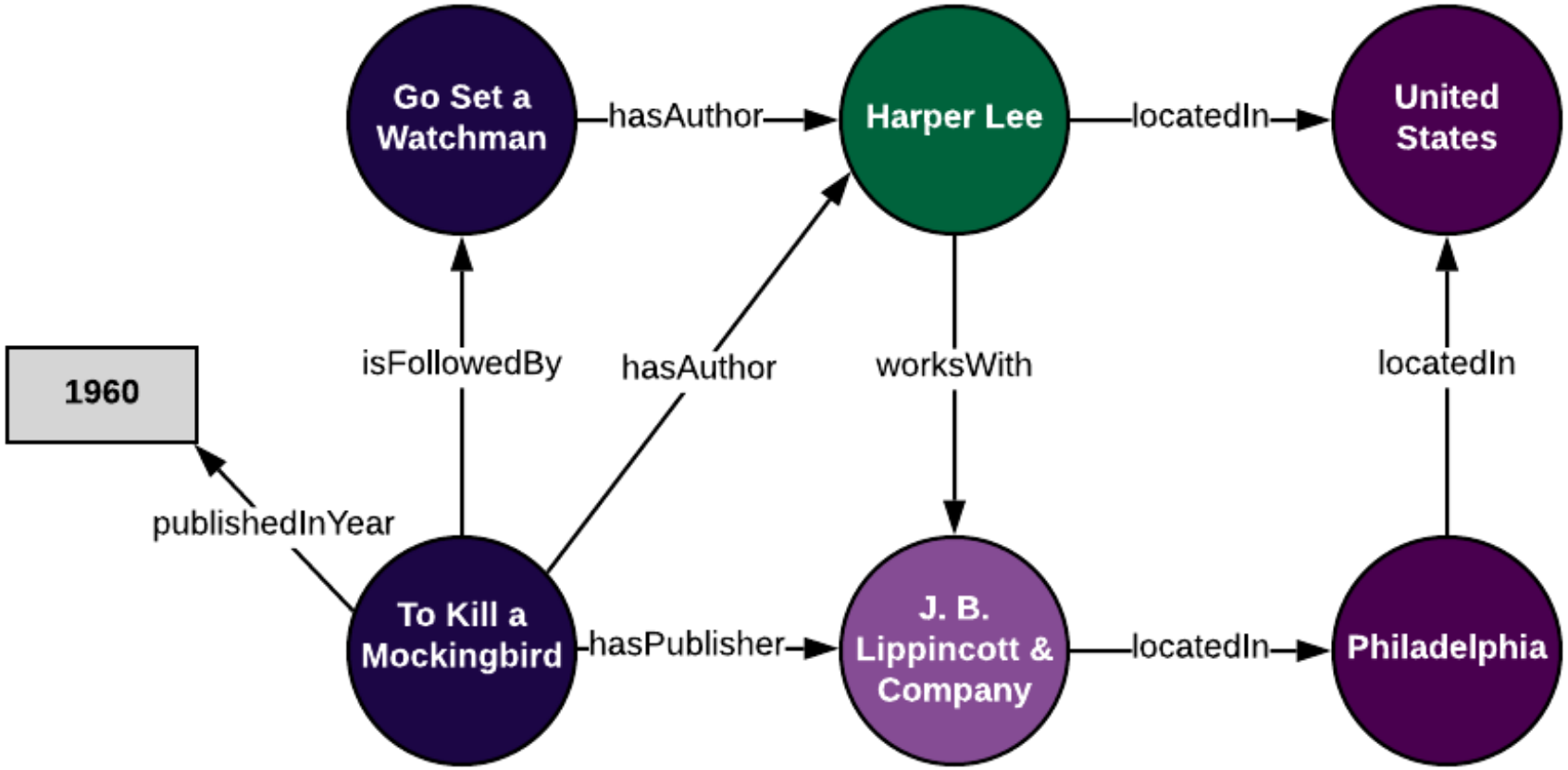
기존 지식 그래프는 보통 스포츠, 음악, 영화와 같이 데이터가 잘 정리/구축 되어있는 도메인으로 지식 그래프 구축을 하고, 수동으로 전문가에 의해 생성됨.

- **Ontology 란**
  : 도메인에 존재하는 사물들의 유형과, 이를 설명하는 데 사용되는 속성을 정의하는 Semantic data model.



Ex) 책에 대해 생성한 Ontology

- **Knowledge Graph 란**
  : Ontology를 프레임워크로 사용하여, 실제 데이터를 추가해 만든 그래프.



Ex) 책에 대해 생성한 Ontology를 기반으로 '앵무새 죽이기' 책에 대해 생성한 Knowledge Graph

(출처 : https://enterprise-knowledge.com/whats-the-difference-between-an-ontology-and-a-knowledge-graph/)

- **기존 Knowledge Graph 생성**

  주로 음악, 영화, 스포츠 등 데이터가 잘 정리되어 있는 도메인에서 성공적으로 이루어짐.

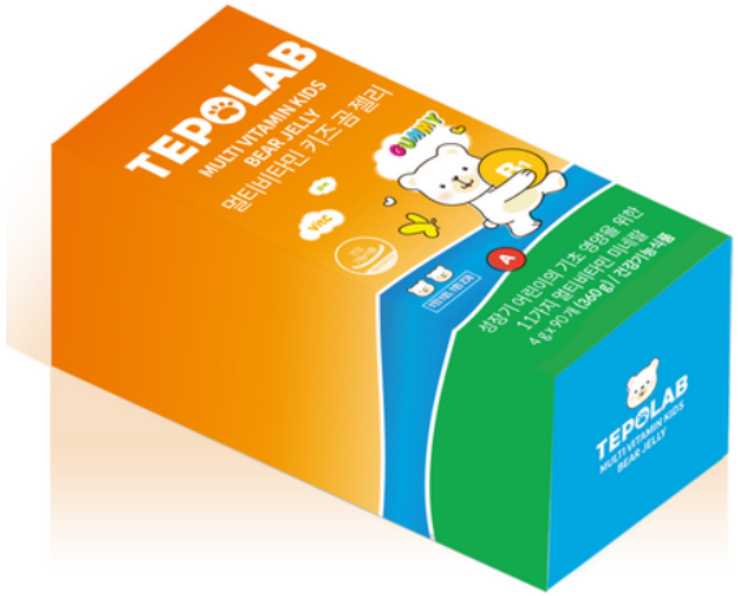
  - 원래 데이터가 풍부하고, 구조가 잘 잡혀있었음.
  - 이미 잘 정리된 데이터 소스 존재
  - 해당 도메인의 내용 복잡도가 처리할 만한 수치
  - 수작업으로 Ontology 를 구축해도 몇 주면 가능

  → 잘 정리되어 있지 않다면?

- **문제 정의**

1. C1-Structure-sparsity
2. C2-Domain-complixity
3. C3-Product-type-variety

1. C1-Structure-sparsity



테포랩 젤리영양제 성장기 종합비타민 어린이 영양제 "90구미" 키즈 곰젤리 멀티비타민 글루콘산아연

★★★★★ 108개 상품평

33,500원

최대 335원 적립

무료배송
모레(금) 8/25 도착 예정

판매자: 일성 랩 | 판매자 상품 보러가기 | 다른 판매자 보기(2)

판매자 평가 94% (16)

배송사: CJ 대한통운

개당 캡슐/정 × 수량
90정 × 1박스

제품명	테포랩 어린이 멀티비타민 키즈 곰젤리
제조업소의 명칭과 소재지	용기 및 제품 별도표시

2. C2-Domain-complixity

- 하위 유형

  ex) swimsuit > athletic swimwear
- 동의어

  ex) swimsuit = bathsuit
- 중복 유형

  ex) fashion swimwear – twopiece swimwear
- 시간에 따라 변화하는 속성

  ex) TV의 wifi 기능

3. C3-Product-type-variety

- 가지각색의 다양한 제품 유형들

- 비슷한 유형도 서로 다른 속성, 제목, 설명 등의 텍스트 패턴을 가짐.

→ 모든 제품 유형에 대한 학습 데이터 수집은 cost가 많이 들고,
수많은 모델을 유지 관리하는 것은 비효율적이다.

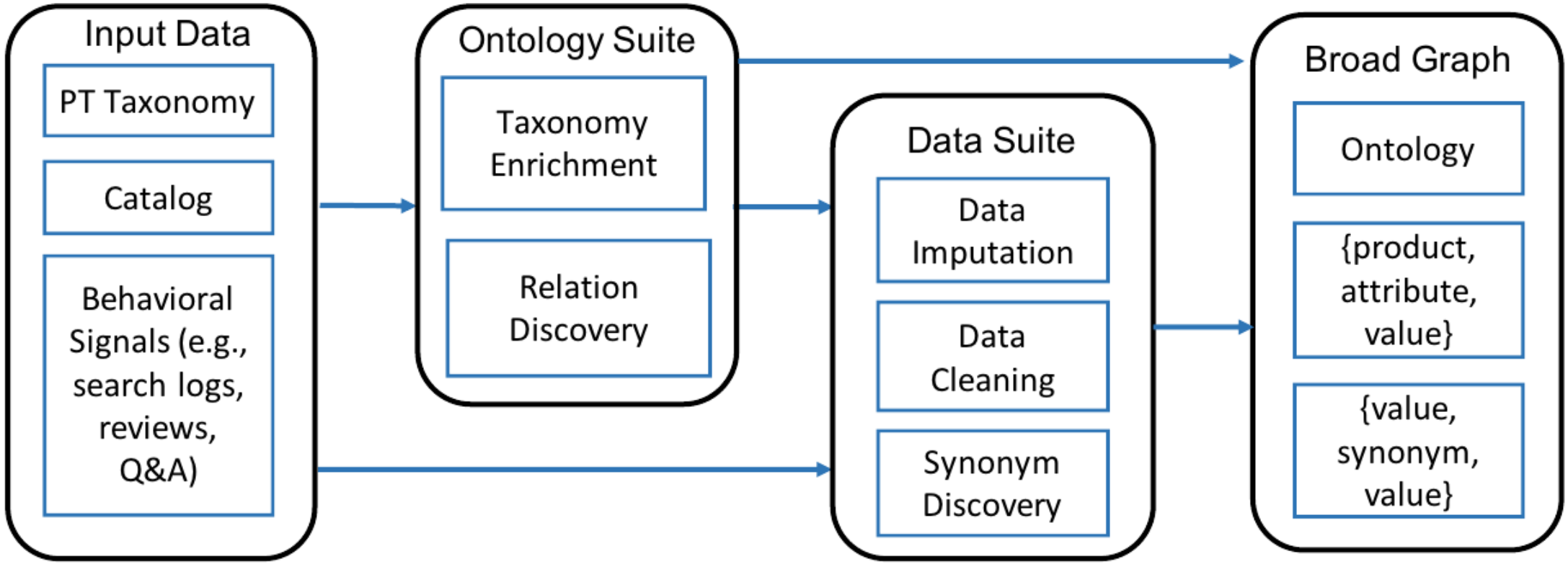
- **Self-driving**

  : 자율주행과 같이 적은 사람의 개입으로 환경을 이해하는 방식을 차용

- **특징**

  1. Automatic
  2. Multi-scalable
  3. Intergrative

- AutoKnow Architecture



- **Input Data**

1) Product Catalog

- product taxonomy
- product attribute
- products
- product attribute value

2) Customer behavior logs

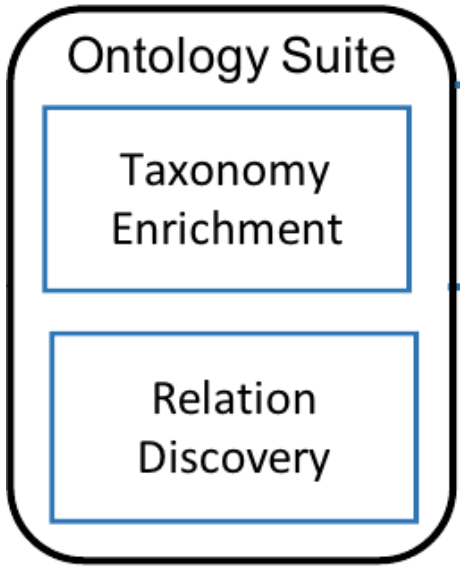
- query
- purchase log
- customer reviews
- Q&A

- **Broad Graph**

: 논문에서 Broad Graph 라고 하는 지식그래프 사용.

G = (Topic Type Entity, Attribute, Edge)

- Ontology Suite



1. Taxonomy Enrichment
   : 기존 taxonomy에 존재하지 않는 새로운 제품 유형 식별
   → 새로 발견한 유형과 기존 유형 간의 관계 결정
   → Taxonomy 보강

2. Relation Discovery
   : 특정 속성이 제품 유형에 적용되는지 여부 판단
   → 적용 가능할 시,
   구매 결정에 있어서 해당 속성의 중요도 점수 결정

- Taxonomy Enrichment → address C2

1) Type extraction

제품 제목, 고객 검색 쿼리에서 새로운 제품 유형 인식해야 함.

→ Open-word tagging 모델 채택 + Type extraction을 BIOE sequential labeling 문제로 공식화

Table 2: Example of input(text)/output(BIOE tag) sequences for the type and flavor of an ice cream product.

Input	Ben	&	Jerry's	black	cherry	cheesecake	ice	cream
Output	O	O	O	B-flavor	I-flavor	E-flavor	B-type	E-type

→ product type = "ice cream" / product flavor = "black cherry cheesecake"

- **Taxonomy Enrichment → address C2**

1) Type extraction

제품 제목, 고객 검색 쿼리에서 새로운 제품 유형 인식해야 함.

→ Open-word tagging 모델 채택 + Type extraction을 BIOE sequential labeling 문제로 공식화

→ 학습된 extraction 모델을 입력 데이터에 적용

→ 새로운 제품 유형 추출

- **Taxonomy Enrichment → address C2**

2) Type Attachment

추출한 유형을 기존 Taxonomy 로 정리

→ 기존 유형과 새로운 유형 사이 Hypernym 관계가 존재하는지 판단하는 이진 분류 문제

→ 고객 행동 데이터를 기반으로 그래프를 구성하여, GNN 모델을 통해 type representation을 얻고, 분류를 위해 semantic feature과 결합

→ 이진 분류기에 공급 및 모델 학습

- Relation Discovery

목표

- 다양한 속성 중 구매 결정에 영향을 미치는 속성 식별
- 많은 유형에 적용 가능한 속성이어야 함

  ex. '맛' 속성 = 과자 적용 가능, 샴푸 적용 불가능

- 언급 수가 많은 속성이 중요도도 높고, 많은 유형에 적용 가능하다.

- **Relation Discovery**

제안하는 방식

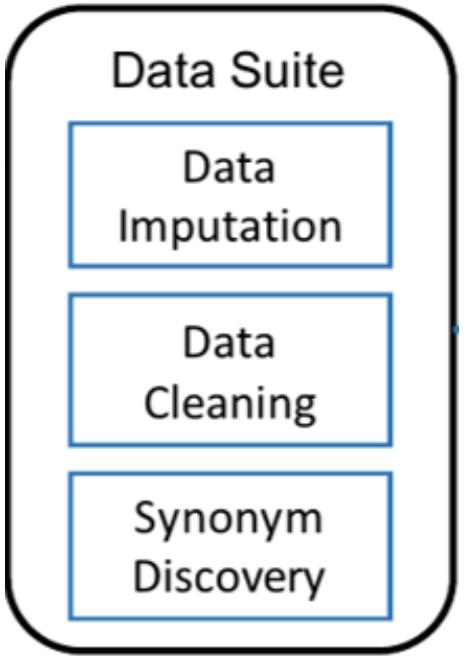
- 적용 가능성을 결정하기 위한 분류 모델과 속성 중요도를 결정하기 위한 회귀 모델 학습

- 두 모델 모두 Random Forest 사용

- 고객 행동 반영하는 2가지 Feature 사용

  1. 특정 제품 유형에 대한 속성의 적용 범위와,
     제품 프로필에서 속성에 대한 언급 빈도로 파악한 **판매자의 행동**
  2. 검색 쿼리, 리뷰, Q&A 등에서 속성을 언급하는 빈도로 파악한 **구매자 행동**

- **Data Suite**



1. Data Imputation
2. Data Cleaning
3. Synonym Discovery

- **Data Imputation**

  - 기존 방식

    : 추출한 속성 값을 임베딩하여,

    BiLSTM+CRF(Conditional Random Field)를 사용한 sequence labeling

$$(y_1, y_2, \ldots y_L) = \mathrm{CRF}(\mathrm{BiLSTM}(e_{x_1}, e_{x_2}, \ldots, e_{x_L})),$$

  - → 생성한 sequence tag를 통해 제품의 속성 식별 가능.

    하지만, 대량의 제품 유형 및 속성으로 확장하기 어려움. → C3

- **Data Imputation**

  - 제안하는 방식 : taxonomy-aware sequence tagging approach

$$(y_1, y_2, \ldots y_L) = \mathrm{CRF}(\mathrm{CondSelfAtt}(\mathrm{BiLSTM}(e_{x_1}, e_{x_2}, \ldots, e_{x_L}), e_T))$$

1. e_T (= pretrained hyperbolic-space embedding)을 이용하여 노드 간의 계층적 관계 보존 + CondSelfAtt Layer 추가하여, e_T 가 attention weight에 영향 줄 수 있도록 설계
2. 공유 BiLSTM Layer 사용하여, sequence tagging과 제품 분류 동시에 학습하는 멀티 태스크 학습 사용

- **Data Cleaning**
  - Data Imputation 과 마찬가지로 많은 유형과 속성에 적용시킬 수 있게 하는 것이 목표 (C3)

  - 모델을 학습하기 위해, input Catalog로부터 학습 레이블 자동 생성

  - 여러 브랜드에서 자주 나타나는 값들로 예제를 생성하고,
    이런 positive 예제에서 negative 예제 만들기 위해 임의로 세 절차 중 하나 적용

    1. 임의의 단어로 attribute 값 치환
    2. 임의로 제목에서 n-gram 선택
    3. 다른 attribute 값 임의로 선택하여 대체

- Synonym Finding

1. 고객 행동 신호에 CF 적용 → 유사성이 높은 제품 쌍 = 동의어 후보 쌍

2. 간단한 로지스틱 회귀모델 훈련하여 후보 쌍이 정확히 동일한 의미 갖는지 판단

   - 고려 특징 : 편집 거리, 사전 학습된 MT-DN 모델 점수,
     고유 단어와 공통 단어에 대한 feature

- Experimental Result

  - Raw Data

    : Amazon 제품 Catalog 중 식료품/건강/미용/유아 4가지 도메인에서 제품 선택.

    (특징 : Sparse한 데이터, Type의 종류와 크기가 도메인 별로 차이남.)

    (도메인 별로 빈도 높은 상품들 임의로 선택)

Product Domain	Grocery	Health	Beauty	Baby
#types	3,169	1,850	990	697
med. # products/type	1,760	18,320	27,150	28,700
#attributes	1,243	1,824	1,657	1,511
med. #attrs/type	113	195	228	206

  - Resulting PG

#Triples	#Attributes	#Types	#Products
>1B	>1K	>19K	>30M

- **Quality Evaluation**

  - Metrics : 세 종류 triples 고려

    1) 제품 타입을 갖는 것 (product-1, hasType, Television)

    2) attribute 값을 갖는 것(product-2, hasBrand, Sony)

    3) Entity 간의 관계를 정의하는 것(chocolate, isSynonymOf, choc)

  - Defect rate

    : 잘못되거나 누락된 (product, attribute) 쌍 비율

- Quality Evaluation

	Attribute 1		Attribute 2		Attribute 3	
Grocery	Input	PG	Input	PG	Input	PG
Applicability	38.51%		7.53%		10.00%	
Precision	68.61%	82.59%	49.94%	77.30%	55.10%	55.10%
Recall	37.17%	83.15%	1.43%	80.96%	54.58%	54.59%
F-measure	48.22%	82.87%	2.78%	79.09%	54.84%	54.85%
Defect Rate	62.91%	21.14%	98.58%	30.72%	49.50%	49.49%

Input 데이터에 비해 평균적으로 개선된 수치

- Precision : 7.6 %
- Recall : 16.4%
- Defect Rate : 14.4%