관계 추출 태스크 가이드라인

1. 관계 추출 Task 알아보기

관계 추출 Task는 하나의 문장에서 나타나는 Entity 쌍 사이의 의미적 관계를 분류하는 Task 입니다. 문장과 Subject와 Object에 대한 설명이 주어집니다. 이 문장에서 나타나는 Entity 쌍은 관계의 주체가 되는 Subject Entity와 대상이 되는 Object Entity로 이루어집니다.

대학 데이터셋에 초점을 두어 wikipidia api를 통해 생명과학, 물리학, 문학, 수학, 러시아어, 스페인어, 중국어 등에 **대학에서 배울 수 있는 학문을 토대로** 데이터 셋을 구축했습니다.

자연어처리를 통해 대학 학문 관계 그래프를 만드는 이유는 대학에서 배우는 여러 학문에 대한 구분이 명확하지 않다고 판단했기 때문입니다. 기존 분류에 대한 의견 및 이론 정립은 자주 나왔음에도 불구하고 많은 학생들이 여전히 명확한 구분을 하지 못하고 있습니다. 대량의 정보량을 토대로 자연어 학습을 하게 된다면 '실제로 쓰이는 바'를 토대로 지식그래프를 구축할 수 있게 됩니다.

대학에서 배울 수 있는 학문들의 관계를 추출하고자 하는 연구자 분들에게 도움이 될 것이고 이외 학문에 대한 지식 그래프 및 유관 분야 소개 등으로의 확장 또한 가능할 것입니다.

각 Entity는 Type을 가지고 있습니다. Type에 대한 자세한 설명은 아래와 같습니다.

Entity Type	description	example
STU	학문, 이론	물리학, 정수론
LAN	언어, 언어학의 기초분야	아랍어, 굴절어
PER	사람, 인물	보일, 이건희
ORG	기관, 단체	UN
DAT	날짜, 일자	1937년, 르네상스 시대
LOC	위치, 장소	대한민국, 동아시아
РОН	사물, 명칭, 기타 고유명사	문예가, 공유결합

Entity 쌍 사이에 나타날 수 있는 관계는 총 14개의 클래스 중 하나로 분류합니다. 관계 클래 스는 구체적으로

- 1) 학문 또는 이론(STU) 중심 관계 8개
 - 학문 중심 관계는 Subject Entity가 학문일 때 발생할 수 있는 관계로서 "학문:하위 학문", "학문:상위 학문", "학문:별칭", "학문:기여자", "학문:시대", "학문:연구 집단" 등이 이에 해당합니다.

- 2) 언어 또는 관련 이론(LAN) 중심 관계 6개
 - 언어 중심 관계는 Subject Entity가 러시아어, 프랑스어, 일본어 등의 언어일 때 갖는 관계로서 "언어:하위_언어", "언어:상위_언어", "언어:파생물", "언어:사용_지역", "언어:별칭", "언어:사용 집단" 등이 이에 해당합니다.
- 3) "관계 없음" 1개로 구성됩니다.
 - "관계_없음"은 주어진 문장에서 Entity 쌍이 아무 관계가 없음을 의미합니다. (2-2 관계 없음에서 자세히 설명)
- 전체 15개의 관계 클래스에 대한 자세한 설명 및 예시는 링크를 꼭 참고해주세요!

관계 추출 태스크에 대한 이해를 돕기 위해 다음의 예시를 살펴보겠습니다.

Example #1:

보일이 원래 발견자가 아님에도 보일은 1662년에 제시한 <mark>보일의 법칙</mark>으로 가장 잘 알려져있다.

Entity	name
Subject	보일의 법칙 (STU)
Object	보일 (PER)

위의 문장에서 Entity 쌍을 이루는 Subject Entity는 "보일의 법칙" 이며, Object Entity는 "보일" 입니다. 또한 위 문장에서 "보일의 법칙"와 "보일" 사이의 의미적 관계는 "학문:기여자" 클래스라고 분류할 수 있습니다.

2. Annotation 가이드라인

2-1) 관계의 방향성

Subject에 대한 Object의 관계와 Object에 대한 Subject의 관계는 다를 수 있습니다. 즉, 경우에 따라서 관계의 방향성이 존재합니다. 다음의 예시를 통해서 자세히 알아보겠습니다.

Example #2:

보일은 1662년에 제시한 보일의 법칙으로 가장 잘 알려져있다.

Entity	name
Subject	보일 (PER)
Object	보일의 법칙 (STU)

Example #1과 다르게, 이번에는 "보일"이 Subject, "보일의 법칙"이 Object Entity입니다. 이 경우에 관계는 "학문:기여자"가 아닙니다. 보일은 인물이기 때문에 데이터 셋에서 사전

에 정의한 관계에 부합하지 않습니다. 따라서 "관계 없음"으로 분류하는 것이 맞습니다.

즉, 관계에는 방향성이 존재할 수 있고 동일한 Entity 조합이여도 Subject와 Object가 무 엇이냐에 따라서 관계는 완전히 달라질 수 있습니다.

2-2) 관계 없음

문장에서 Entity 사이에 의미적 관계를 유추할 수 없거나, 있더라도 정의된 15개의 관계 클래스에 존재하지 않을 수 있습니다. 이러한 경우, 주어진 샘플을 "관계_없음" 클래스로 분류합니다.

Example #3:

갈릴레오 갈릴레이는 실험과 관찰을 통해 과학적 지식을 발견하였고, 이를 수학적 모형으로 서술하여 이후 **물리학 발전**에 큰 영향을 주었다.

Entity	name
Subject	물리학 (STU)
Object	갈릴레오 갈릴레이 (PER)

위의 문장에서 "물리학 발전"과 "갈릴레오 갈릴레이" 사이에는 어떠한 관계도 유추할 수 없습니다. 이 경우에 Entity 쌍 사이의 관계는 "관계_없음"이 됩니다. 혹여나 관계가 있더라도 태스크에서 정의된 14개의 관계 클래스("관계_없음" 제외)에 존재하지 않는다면 이 역시 "관계 없음"으로 분류합니다.

2-3) STU Entity의 기준

STU Entity는 STUDY의 약어 표현으로 학문 또는 이론에 해당하는 것을 지칭했습니다. 가령 생물학, 물리학 등의 학문을 표현하는 Entity와 양자이론, 상대성이론 등의 이론으로 판단 가능한 범위를 모두 포함한 Entity를 STU Entity로 설정했습니다. e.g) OO학,OO론, OO법칙

Example #5:

<mark>공유 결합</mark>이 형성되는 결합은 <mark>발열반응(發熱反應)</mark>인데, 이때 방출되는 에너지의 양이 그 결합의 결합 에너지이다.

Entity	name
Subject	발열반응 (POH)
Object	공유결합 (POH)

위 문장에서 "발열반응"과 "공유결합" 사이의 관계로 "관계_없음"라고 설정해야합니다. 자칫 화학에서 많이 사용하는 표현이기 때문에 STU Entity라고 생각하기 쉽지만 학문 또는 이론 을 말하는 것이 아닌 해당 Entity 안에 포함되는 작은 개념이기 때문에 POH로 분류됩니다. 이와 같은 경우 "관계 없음"으로 분류하는 것이 맞습니다.

2-4) LAN Entity의 기준

LAN Entity는 프랑스어, 한국어, 등 언어만 뜻하는 것이 아닙니다. 언어학에서 다루고 있는 언어학의 분류 형태론에는 굴절어, 교착어, 고립어가 있듯이 **언어학의 기초분야에 해당하는** 개념들까지 포함한 것을 LAN Entity로 설정했습니다.

Example #4:

이렇게 아랍어는 형태론적으로는 굴절어에 속한다.

Entity	name
Subject	아랍어 (LAN)
Object	굴절어 (LAN)

위 문장에서 "아랍어"와 "굴절어" 사이의 관계로 "언어:상위_언어"라고 설정해야합니다. "굴절어"가 사용하는 언어가 아닌 형태론적 분류 개념이기에 언어가 아니라고 생각할 수 있으나. LAN Entity는 언어학의 기초분야에 해당하는 개념들까지 포함 하는 것을 기준으로 하기때문에 LAN Entity가 맞고 이와 같은 경우 "언어:상위_언어"에 해당합니다.

2-5) 관계의 시점

"언어:사용_지역", "언어:사용_집단" 등의 관계들은 시간에 따라서 달라질 수 있습니다. 이런 관계의 경우 **문장의 시점을 기준**으로 관계가 존재하는지 판단합니다. **현재 시점이 아닌 문장의 시점을** 기준으로 한다는 점에 주의합니다.

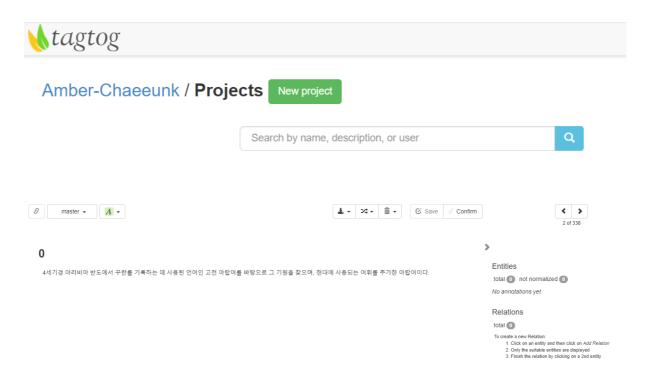
Example #5:

미국령 괌, 팔라우, 북마리아나 제도 그리고 미국에서 독립한 마셜 제도와 **미크로네시아 연방** 모두 한때 **에스파냐어** 사용자가 있었다.

Entity	name
Subject	에스파냐어 (LAN)
Object	미크로네시아 연방 (LOC)

위 문장에서 "미크로네시아 연방"에 한때 "에스파냐어" 사용자가 있었다라는 의미가 담겨있습니다.현재 미크로네시아 연방에서 에스파냐어를 계속 사용하고 있을지, 아니면 모두 사라 졌을지는 외부 지식 없이는 모르는 일입니다. 즉, 이를 주어진 문장에서 유추하는 것은 불가 능합니다. 적어도 위 문장에서 말하는 시점에서는 미크로네시아 연방에서 에스파냐어를 사용하고 있었으므로, 위의 예시는 "언어:사용 지역"으로 분류합니다.

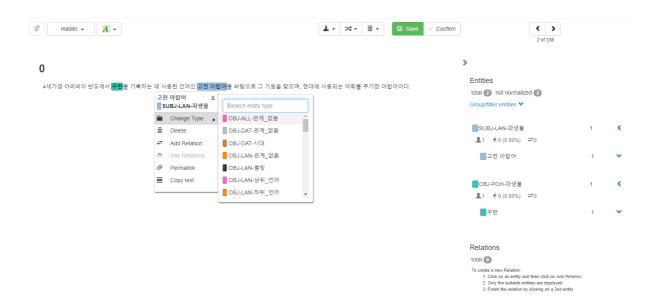
3. Annotation 환경



작업자 화면 예시:

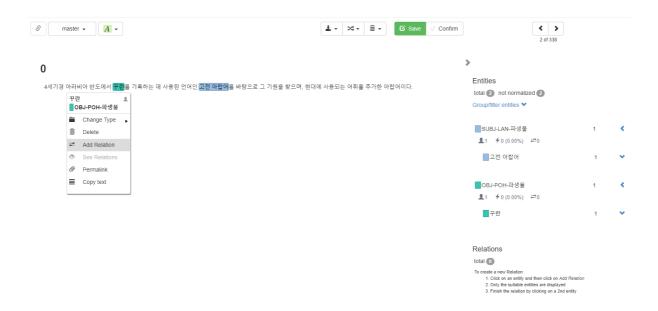
작업자의 화면에는 위와 같이 1) 좌측에 하나의 문장, 우측에 현재 선택된 Subject/Object Entity, Relations가 표시됩니다.

3-1) Entity 선택하기

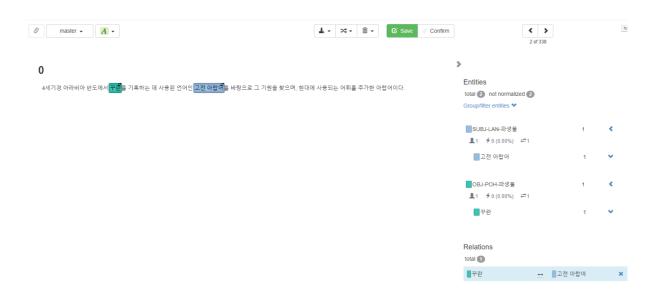


가이드라인에 기반하여 주어진 문장에서 Subject/Object Entity를 선택한 후 그에 맞는 Type을 결정합니다. Type 마지막 부분은 relation의 관계를 표현합니다.(OBJ-DAT-시대) subject/object를 동일한 관계로 선택합니다 만약, 관계가 존재하지 않는다면 "관계_없음"을 선택합니다.

3-2) 관계 추가하기

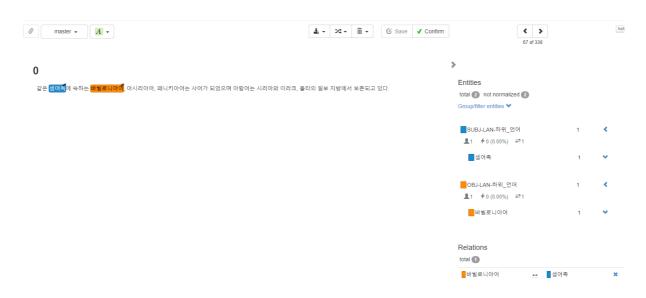


위 사진과 같이 Add Relation 탭을 선택하여 설정한 Subject Entity와 Object Entity사이 관계를 이어줍니다.



위 사진과 같이 우측 하단에 관계가 추가된 것이 보이면 완료된 것입니다.

3-3) 한 문장에서는 하나의 관계



위 사진에 있는 문장은 한 문장에 여러 관계가 있는 문장입니다. (셈어족, 바빌로니아어) 쌍뿐만 아니라 (셈어족, 아시리아어) 쌍도 관계가 존재합니다. 그렇다고 하더라도 한 문장에서는 한 쌍의 Entity, 즉 하나의 관계만을 선택하면 되겠습니다.

3-4) Hate / Bias / Privacy (중요!)

만약 데이터에 혐오 표현, 편향적 표현, 개인정보가 포함되어 있다면 "Hate / Bias / Privacy" 체크박스를 눌러주세요. 구체적으로 다음의 경우가 이에 해당합니다.

- 타인 및 집단에 대한 혐오를 표출하거나 기분 나쁘게 할 수 있는 표현을 내포하는 경우 (e.g., 비하, 모욕, 비꼼)Ex) "다음 카페 여성시대 회원들은 해당 기사를 두고 "ㅉㅉㅋㅋ", "어쩌라고", "남잔 관심 없어용~" 등 반응을 보였다." → "단체:하위_단체" + "Hate / Bias / Privacy" 체크Ex) "한 네티즌은 신재은 씨의 비현실적인 몸매를 보고 "에이 이건 좀 보정이네"라는 반응을 보였다." → "관계_없음" + "Hate / Bias / Privacy" 체크
 - 성별, 성 정체성, 성적 지향, 인종, 장애, 출신지, 국적, 가족 형태, 종교 등 개인/집단의 특성에 근거하여 적대감을 표출하거나 비하하는 글
 - 성희롱 및 성적 모욕, 기타 욕설을 포함한 모욕
 - 。 사실 적시를 통한 모욕
 - 낙인 표현 (일베, 메갈, 대깨 등)
 - 반어, 냉소, 비인륜적 표현 등을 활용한 기분 나쁜 언사

- 특정 집단 및 특성에 대한 편견을 내포하는 경우 (e.g., 성차별, 지역차별)Ex) 사진 속 조이는 실크 소재의 탑 드레스를 입어 섹시하면서도 고혹적인 느낌을 선보였다. → "관계 없음" + "Hate / Bias / Privacy" 체크Ex) 신천지는 모지리들만 가입하나봐. → "관계 값음" + "Hate / Bias / Privacy" 체크
 - 특정 성별을 갖거나, 성 정체성 및 지향성, 성 관련 사상 등을 가진 사람들이 특정한 방식으로 사고 및 행동할 것이라는 편견
 - 그 외에도, 특정 인종, 출신국, 종교, 정치적 견해 등에 해당하는 사람들이 특정한 방식으로 사고 및 행동할 것이라는 편견
- 개인정보를 포함하는 경우 (e.g., 이름, 연락처, 인적사항)Ex) "구입문의는 홈페이지 (또 는 홍길동 010-1234-1234)에 문의하면 된다." → "관계_없음" + "Hate / Bias / Privacy" 체크Ex) "아이디는 '_mariahwasa'다." → "관계_없음" + "Hate / Bias / Privacy" 체크
 - 개인에 대해 그 사람을 특정할 수 있는 인적사항(이름, 고유식별정보, 상세 주소, 날짜정보, 각종 자격증 및 면허 번호, 전화번호나 계좌번호와 같은 숫자 정보, 이메일주소, 아이디와 같은 식별코드 등)이나 사건 등이 기록된 문서 혹은 문장
 - 단, 공인 혹은 네이버나 카카오의 인물정보 서비스에 등재된 사람의 경우, 접근이용이한 정보에 대해서는 공개된 개인정보로 간주하지 않습니다.

4. 자주 묻는 질문

Q1. "공용어", "공식 언어", "외래어", "신조어" 는 타입이 "LAN", "POH" 중 어느 것에 해당하나요 ?

- 언어학의 기초분야에 해당하는 개념들까지 포함한 것을 LAN Entity로 설정했습니다.
- "외래어", "신조어"는 언어학의 기초분야에 해당되는 용어이기 때문에 "LAN"으로 표기 합니다.
- "공용어", "공식 언어"는 언어학으로 보기 힘들기 때문에 "POH"로 표기합니다.

Q2. 상위, 하위 개념이 겹치는데 하나로 통일 할 수 있나요?

- Subj, Obj 의 위치를 기준으로 의미 해석이 달라질 수 있는 요소가 존재합니다.
- 해석이 달라지는 경우 관계_없음으로 나타내지 않기 위해 각각 다른 관계로 설정하였습니다.
- 상위, 하위 개념은 방향성이 존재하는 명확한 관계입니다. 두 관계를 하나의 관계로 합 친다면 그 정보를 잃어버리게 되기에. 웬만하면 두 관계 모두 유지하는 것이 좋습니다.

Q3. 두 개의 학문에서 "영향이 있다", "근원이 되었다", "필수적 과목이다", "영역이 추가되다"는 학문 : (하위, 상위) 개념에 속할 수 있나요?

- 아니요! **학문:영향 관계**에 해당합니다. 하위, 상위 개념은 트리 구조 상 상하 개념이 존 재하는지 확인해주시면 됩니다.
- ex) python 은 java의 영향을 받아 만들어진 언어이다.
 - 。 해당 문장은 "학문:영향" 관계입니다.

Q4. 파생 언어는 "언어:하위_언어" 에 해당하나요?

- 네! 파생언어는 "언어:하위 언어"에 해당됩니다.
- ex) 현대 아랍권 세계에 있어서 푸스하와 암미야의 관계는, 중세 가톨릭 교회가 지배했던 시기의 **라틴어**와 **로망스어군**의 관계, 즉 후자가 전자로부터 파생되어, 여러 변종의 언어로 갈라져나가는 점에서 유사하다 볼 수 있다.

O5. "언어:사용 지역" 의 범위는 어느정도 인가요?

- 문장 내에서 사용하고 있는 언어라는 명확한 표기가 있는 경우에만 "언어 : 사용 지역" 관계입니다
- ex) 대한민국에서는 중학교 2학년때부터 러시아어를 배울 수 있으며 대학수학능력시험 에서도 제2외국어로 러시아어 I 을 선택할 수 있지만, 러시아어 II 는 수능에 없다.
 - 지역에서 사용하고 있는 것이 아니라 배울 수 있다는 의미이기 때문에, 해당 문장은 "언어: 사용_지역"이 아닙니다

Q6. 모어는 "언어:상위언어"에 해당하나요?

- 네! 모어는 "언어:상위 언어"에 해당합니다.
- ex) 일본어에는 본래 **일본어**의 고유어인 **야마토 고토바**가 있었고, 한자는 뒤에 중국에서 직접 전해지거나 혹은 한반도 서해안(특히 백제 또는 고구려)을 경유하여 전래되었다.

Q7. 언어와 학문의 기준은 무엇인가요?

• 2. Annotation 가이드라인 2-3, 2-4 표현 되어 있습니다.

Q8. 타입이 존재하지 않는 문장은 태깅하나요?

• 대학에서 배울 수 있는 학문 및 언어가 주제이기 때문에 해당 내용과 관련 없는 문장은 태깅하지 않습니다.

O9. 사전적 정의가 외부지식인가요?

- 네, Entity를 Mask 처리 했을 때, 문맥만 보고 정의합니다.
- ex) 이는 대영제국의 대외 팽창으로 인해 전 세계로 영어가 확산되고, 제2차 세계 대전 이후 미국의 경제적 문화적 영향력이 증가한 데에 기인한다.
 - 미국과 영어를 Entity로 가정 하였을 때, 문맥에서 미국이 영어를 사용한다는 라고 할 수 없습니다.
 - 。 즉, "언어: 관계_없음"입니다.

Q10. Subject와 Object가 동일해도 되나요?

• 네, 가능합니다

Q11. "분야"는 "언어: 사용집단"에 해당하나요?

- 아니요. 집단의 의미에 분야는 포함되지 않습니다.
- ex) 다른 이유는 노르만 정복 이후 라틴어나 프랑스어가 정치, 학문, 예술의 전 분야에서 사용되면서 음절이 긴 토박이말들이 사라졌다는 점이다.
 - 정치, 학문, 예술은 "집단"에 포함되지 않습니다.

O12. "예술"과 "시"는 타입이 "STU"인가요?

- 예술은 "학문", 시는 "파생물"로 정의하였습니다.
- 즉, 예술은 "STU", 시는 "POH"입니다.

Q13. "포함관계"는 "학문: (상, 하)위_관계" 인가요?

- 아니요. 포함관계는 "학문 : 요소"입니다.
- ex) 물리학은 만물을 구성하는 기본적인 대상인 원자 내부의 아원자 입자(양성자, 중성자, 전자)부터 가장 크다고 여겨지는 우주 전체까지 현상의 광범위한 범주를 포괄한다.

。 원자는 물리학의 요소로 표현할 수 있습니다.

Q14. 경영학:기업과 같은 관계는 STU:연구 집단 인가요?

- 아닙니다. 문맥상 기업이 경영학에서 다루는 요소로써 기능한다면 학문:요소이고, 그런 문맥이 아닌 대부분의 경우, 학문:관계없음입니다.
- 기업은 경영학과 밀접한 관계가 있고, 경영학을 공부하고 사용해야합니다. 그러나 경영학을 연구하는 집단으로 보기에는 다소 무리가 있습니다.
- 일차적으로 드러난 관계에 집중하시면 됩니다. 너무 넓게 그리고 포괄적으로 볼 필요는 없고, 일차적으로 문맥상 관계가 없다면 **관계없음**으로 보시면 됩니다.
- ex) sbj:경영학적 지식을 통해 성공적으로 obj:기업을 운영할 수 있게 되면 사회에서 요구하는 제품과 서비스를 제공해 고객의 삶의 질에 도움이 되며 근무자의 직무 만족까지 충족되어 기업이나 주주의 이익 뿐만이 아니라 인간의 생활상, 국가경쟁력 향상 등, 생태계의 순환처럼 사회에 큰 기여를 할 수 있다.

Q14. STU:시대는 학문이 첫 등장한 시대만을 일컫는 말인가요?

- 아닙니다. 학문이 첫 등장한 시대를 포함하여, 학문과 밀접한 관련이 있는 시대가 문장 속에 존재한다면, 둘 사이 관계 역시 **STU:시대**로 보아도 무방합니다.
- ex) <sbj:케인즈 경제학>(Keynesian economics)은 <obj:20세기> 영국의 경제학자 존메이너드 케인즈의 사상에 기초한 경제학 이론이다.
- 그러나, 특정 학문과 별 관련이 없는 시대라면 둘 사이 관계를 STU:시대로 보는 것은 무리가 있습니다.

O15. 언어:파생물에서 파생물의 기준이 뭔가요?

- 기본적으로는 그 언어로 쓰인 저작물, 작품 등과 같은 사물을 의미합니다. 아래와 같은 문장이 그 예입니다.
- ex) 4세기경 아라비아 반도에서 <obj:꾸란>를 기록하는 데 사용된 언어인 <sbj:고전 아 랍어>를 바탕으로 그 기원을 찾으며, 현대에 사용되는 어휘를 추가한 아랍어이다.
- 조금 더 포괄적으로는. 그 언어로부터 파생된 사물들을 모두 포함할 수 있습니다.
- 다만 특정 언어로부터 생겨난 새로운 언어 같은 경우, **언어:하위_언어** 관계가 더 적절한 관계입니다.

Q16. 학문:요소, 학문:영향 이 두 관계가 헷갈립니다.

- 학문:요소는 Object가 Subject에 직접 포함되는 요소일 때를 가리킵니다.
- 학문:영향은 Object가 Subject의 영향을 받은 경우를 가리킵니다.
- 학문:영향은 학문:요소에 포함하기는 어렵지만 영향을 받았거나 밀접한 관계인 경우를 의미합니다. 즉, 조금 더 넓고 포괄적인 관계라고 볼 수 있습니다.
- 따라서, Object가 직접적으로 Subject에 포함되는 요소인 경우 **학문:요소**라고 보고, Object가 Subject에 완전히 포함되지는 않지만 영향을 받았다는 것이 문장에 드러난 경우 **학문:영향**으로 봅니다.
- 학문:영향의 예) 이 점에서 <obj:경영학>은 <sbj:경제학>의 이론을 적용하며, 심리학, 통계학, 교육학, 기초공학 등과 연계하여 기업의 경영을 위한 응용 학문이다.
- 학문:요소의 예) <sbj:경영학>의 기본 연구는 기능적으로 구분하면, 생산관리, 인사조직, 마케팅, 재무, <obj:회계>와 최근에 포함된 경영정보시스템 등으로 구분할 수 있다.