암묵적 관계 발견을 통한 QA용 지식베이스 증강

Discovering Implicit Relationships to Augment Web-scale Knowledge Base for QA

맹성현, 김진호
IR & NLP Lab
KAIST

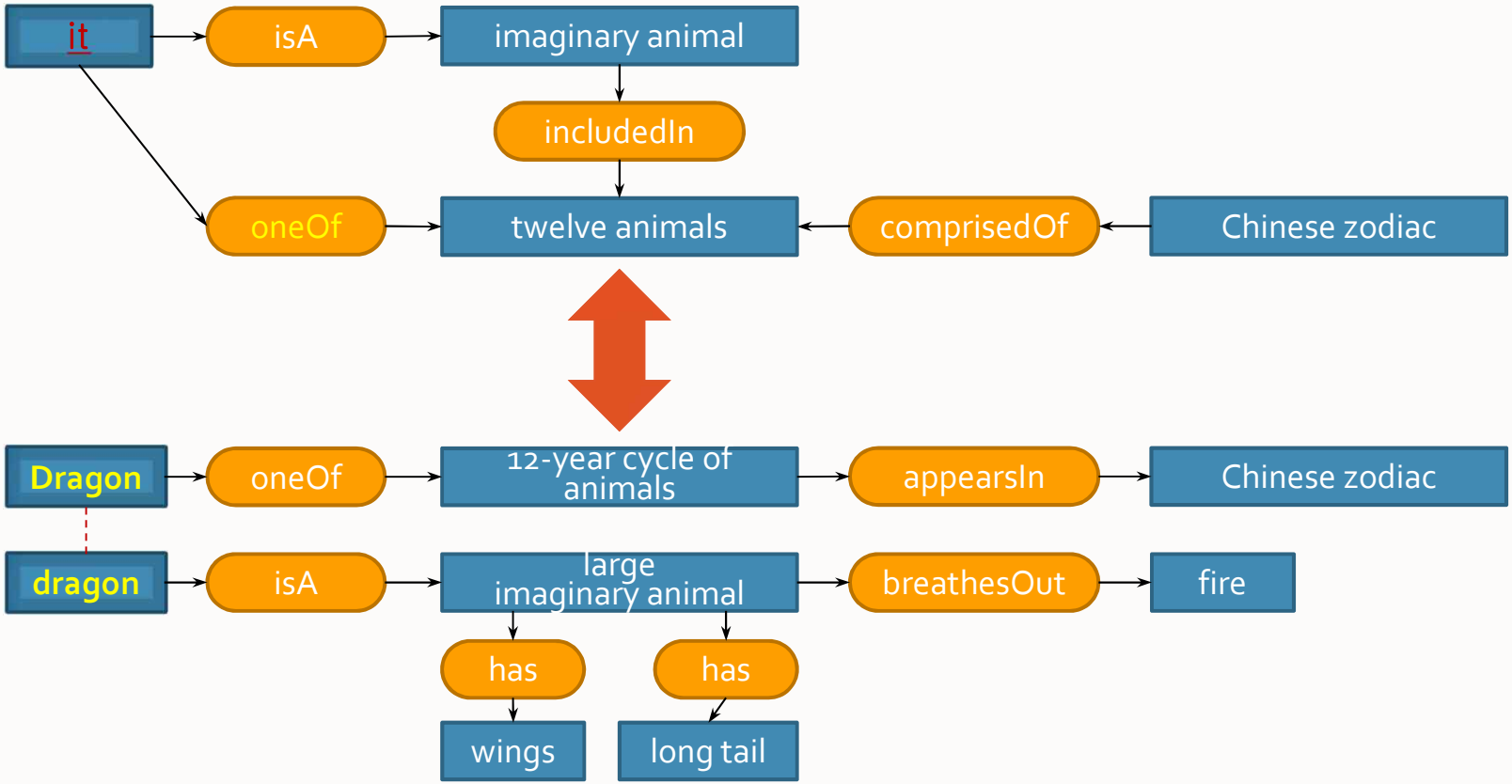
2015.08.21 Fri.

CONTENTS

1
INTRODUCTION

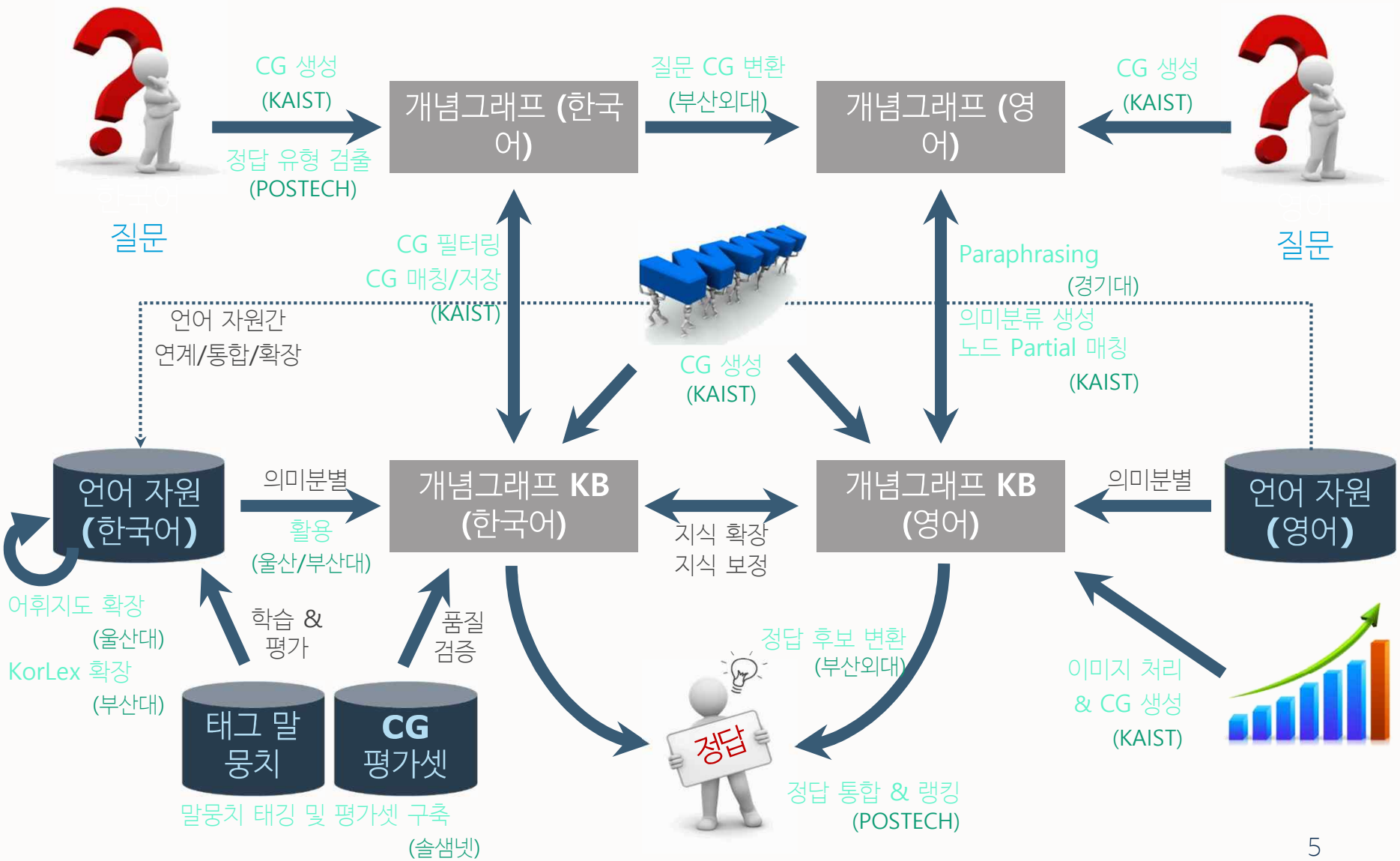
개념그래프(CG) 기반 QA

질문	띠를 나타내는 12마리의 동물 중에서 상상의 동물은?
질문	It is imaginary animal among the twelve animals in the Chinese zodiac.



Dragon (zodiac) from Wikipedia	The Dragon is one of the 12-year cycle of animals which appear in the Chinese zodiac
dragon from Longman dictionary	a large imaginary animal that has wings and a long tail and can breathe out fire

CGQA Flow


CG 생성
(KAIST)
정답 유형 검출
(POSTECH)
개념그래프 (한국어)
질문 CG 변환
(부산외대)
개념그래프 (영어)
CG 생성
(KAIST)
한국어
질문
영어
질문
CG 필터링
CG 매칭/저장
(KAIST)
언어 자원간
연계/통합/확장
CG 생성
(KAIST)
Paraphrasing
(경기대)
의미분류 생성
노드 Partial 매칭
(KAIST)
언어 자원
(한국어)
의미분별
활용
(울산/부산대)
개념그래프 KB
(한국어)
지식 확장
지식 보정
개념그래프 KB
(영어)
의미분별
언어 자원
(영어)
어휘지도 확장
(울산대)
KorLex 확장
(부산대)
학습 &
평가
품질
검증
정답 후보 변환
(부산외대)
이미지 처리
& CG 생성
(KAIST)
태그 말
뭉치
CG
평가셋
정답
말뭉치 태깅 및 평가셋 구축
(솔샘넷)
정답 통합 & 랭킹
(POSTECH)

1.1 Motivation

Knowledge Base

Open Information Extraction

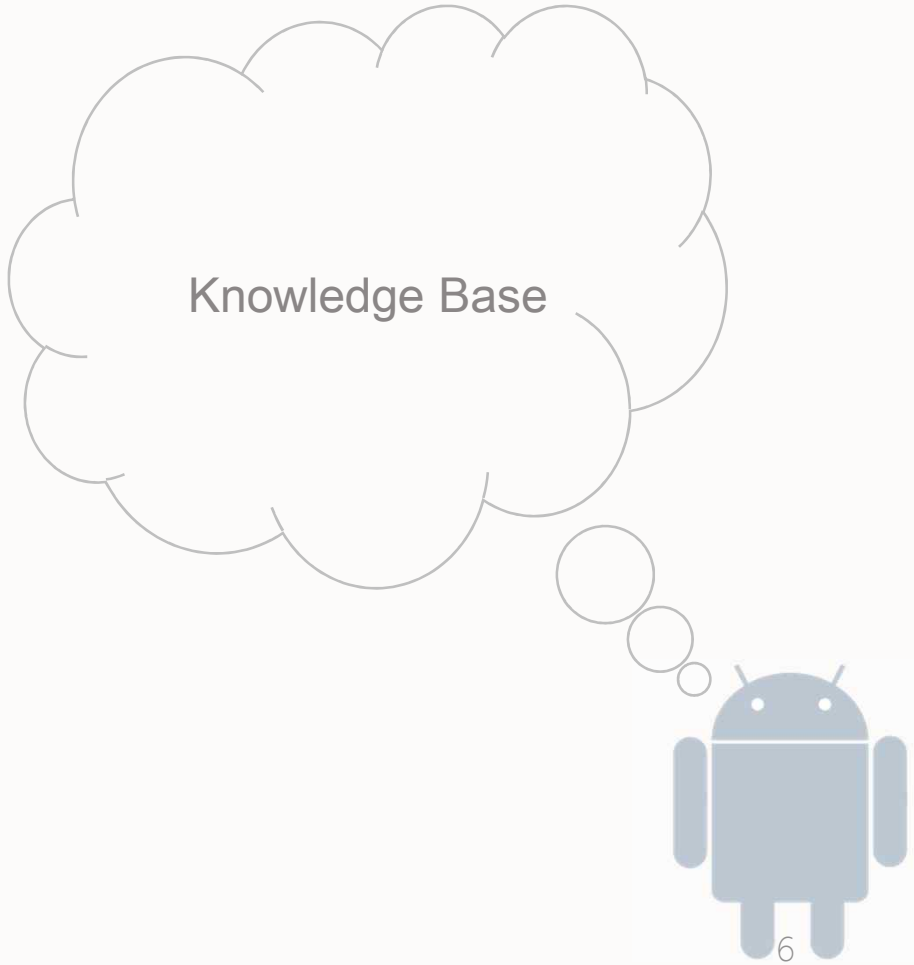
Read the Web
Research Project at Carnegie Mellon University

DBpedia

yago
select knowledge

Freebase

WIKIDATA



1.1 Motivation

Augmenting Knowledge Base

Augmenting Knowledge Base

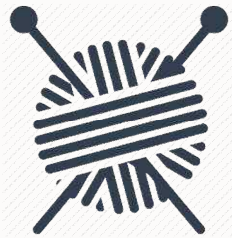
새로운 Entity와
Relation을 추가

Knowledge Base

'I have only
0.0013% of knowledge'
(possible relation links)

기존 Entity와
Relation을 수정

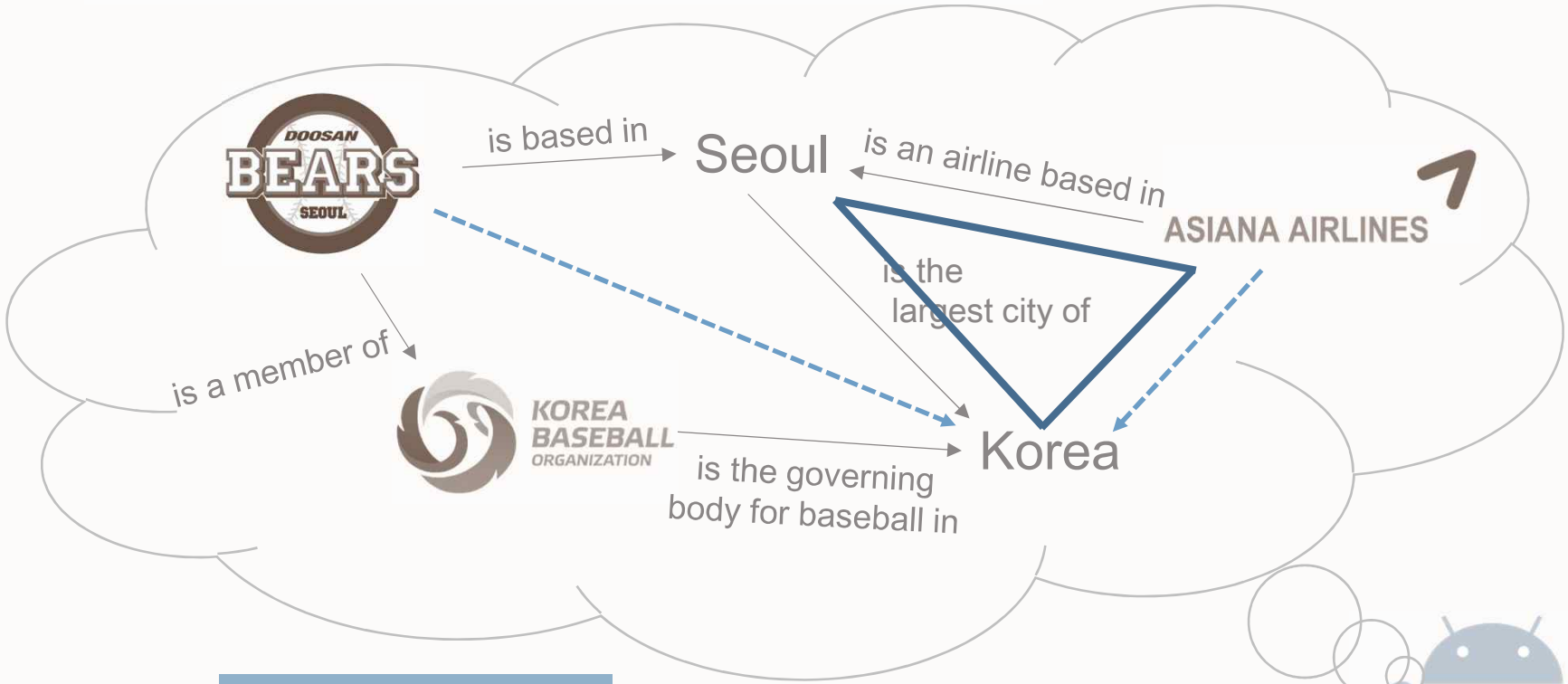
기존 Entity 사이의
새로운 Relation을 추가



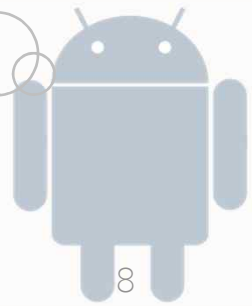
1.1 Motivation

Open Information Extraction and Implicit Relationship

Open Information Extraction (OIE)



OIE Knowledge Base



1.2 Problem Statement

Open Knowledge Acquisition

Infer all possible implicit relationships between two entities
for which no relationship exist in OIE knowledge base

Input	A set of Triples	
	An Entity Pair	<Asiana airlines, Korea>
Output	Implicit Relationships of input	<Asiana airlines, major_airline, Korea>

Infer Implicit Relationship	주어진 두 개체 관계를 설명할 수 있는 적절한 관계명 유추
All Possible Relationships	두 개체 사이에 가능한 모든 관계명을 유추
No Textual Context	지식베이스 Triple 집합 만을 활용

1.3 Related Work

Knowledge Acquisition

2010
SHERLOCK
Schoenmackers & Etzioni
University of Washington

Open IE의 관계명을 활용하여
Inference Rule을 자동으로 추출

Output
{A, Cause By, C}+{C, Short For, B} → {A, Cause By, B}

2011
Saeger & Torisawa
NICT

Class Dependent Pattern과
Partial Pattern을 활용하여
Target Relation의 추가적인 Instance를 추출

Output
Cause(Hypertension, intracranial bleeding)

2014
ProbKB
Yang et al.
University of Florida

SHERLOCK Rule의 조합을 통해
새로운 Inference Rule 추출

Output
{A, Born in, B}+{A, Born in, C}→ {B, Located in, C}

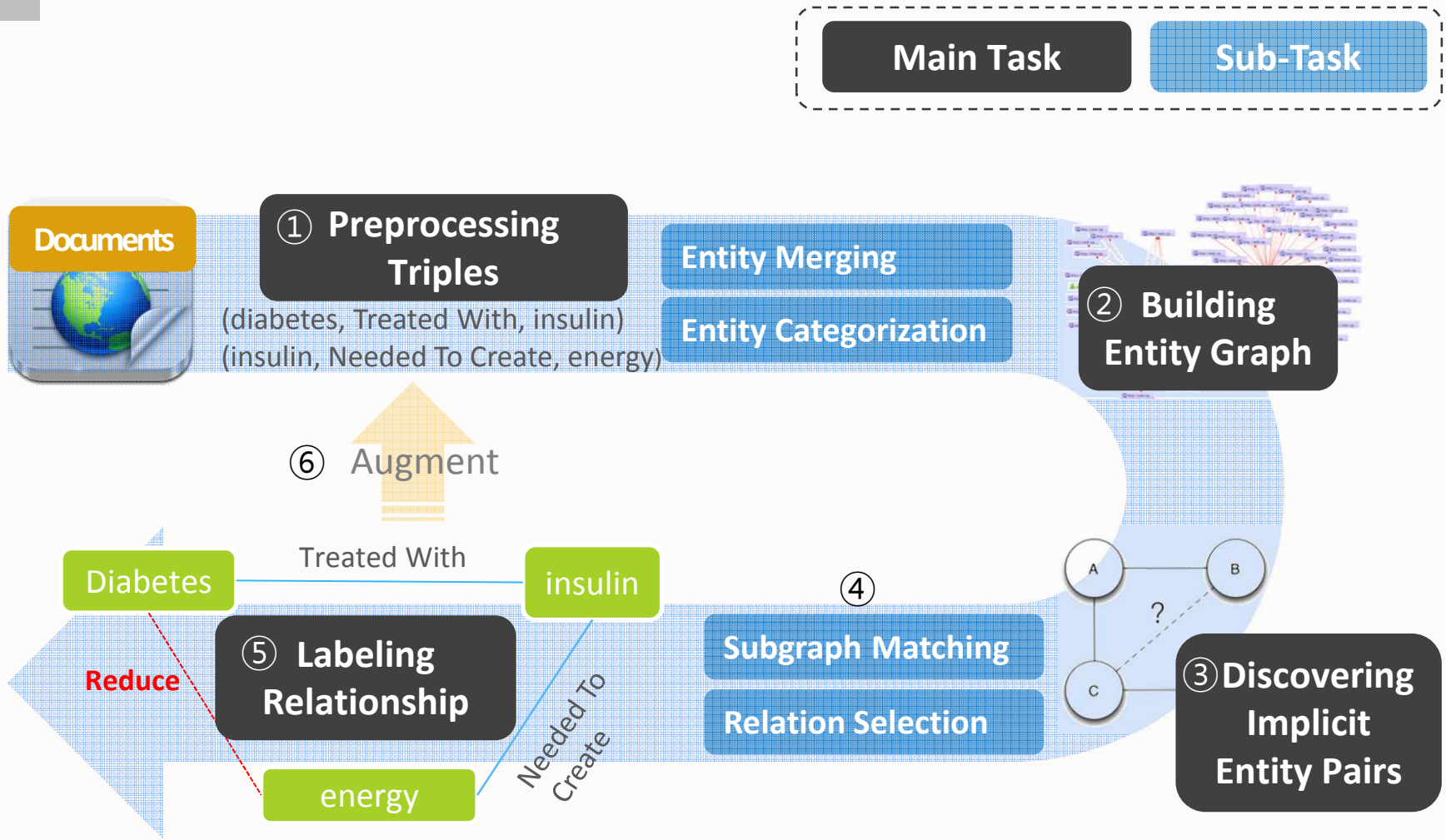
2014
Knowledge Vault: Link Prediction
Dong et al.
Google

Random Walk를 활용하여
Target Relation별 새로운 Instance를 추출

Output
Profession(Charles Dickens,?} → Writer

2
THE METHOD

2.1 Overview


Main Task
Sub-Task
Documents
① Preprocessing Triples
(diabetes, Treated With, insulin)
(insulin, Needed To Create, energy)
Entity Merging
Entity Categorization
② Building Entity Graph
⑥ Augment
Diabetes
Treated With
insulin
Reduce
⑤ Labeling Relationship
energy
Needed To Create
④
Subgraph Matching
Relation Selection
A
B
C
?
③ Discovering Implicit Entity Pairs

2.2 Preprocessing

Entity and Relational phrase normalization

Goal

- 동일한 의미를 지닌 Entity와 Relation을 하나로 통합
- 이후 모듈에서의 매칭 효율성 향상

Entity

Original Unique Entities — 100%

Entity Merging — 94.4%
- 'a', 'the' 제거
- Frequency가 높은 쪽으로 결정

Entity Filtering (For Efficiency) — 30.1%
- 1회 등장 Entity 제거

Relation

Original Unique Relations — 100%

Relational Phrase Normalization — 64%
- 29 stopwords 제거
- POS 'MD','DT' 제거
- Lemmatization

2.2 Preprocessing
Entity Categorization

28.3% **Goal** Entity의 Class를 부여하여 매칭 정확도 향상

10.7% **WordNet Based**

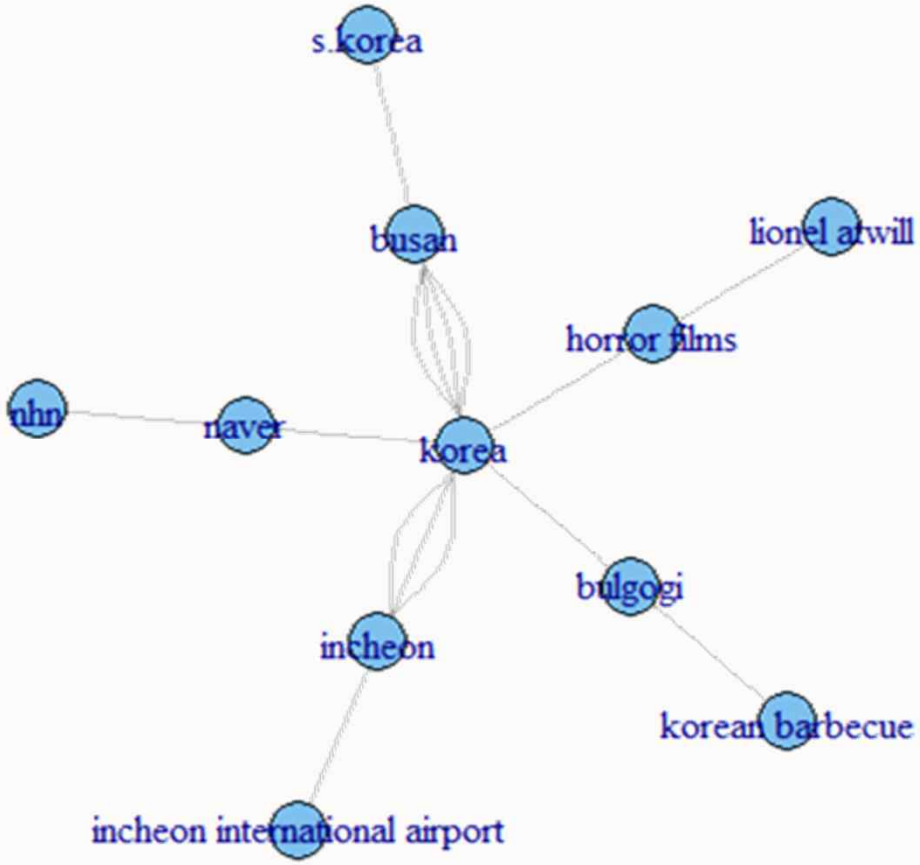
- 첫 번째 Synset의 첫 번째 Direct Hypernym을 Class로 부여

89.3% **Triple Based**

- 패턴 추출 : {A, [be 동사 + noun], B}
  → A의 클래스는 noun
  예. {Asiana Airlines, is an airline based in, Seoul}
  → 'Asiana Airlines'의 클래스는 'airline'

- Meronym : {A, [part of | member of], B}
  → A의 클래스는 B
  예. {Astronomy, is a part of, physics}
  → 'Astronomy'의 클래스는 'physics'

2.3 Constructing Entity-Relation Graph



Directed Graph G = (V, E)

$V(G)$ = a set of Entities in the triple set T

$E(G)$ = a set of relational phrases in T

Direction of Edge(e_1, e_2) = ($e_1 \rightarrow e_2$) in t

Edge Weight(e_1, e_2) = PMI(e_1, e_2)

$$PMI(e_1,e_2) = log\left(\frac{p(e_1,e_2)}{p(e_1)*p(e_2)}\right)$$

$$p(e_1,e_2) = \frac{|triples \quad with \quad e_1,e_2 \ in \ T|}{|T|}$$

2.4 Finding Implicit Entity Pairs

Goal

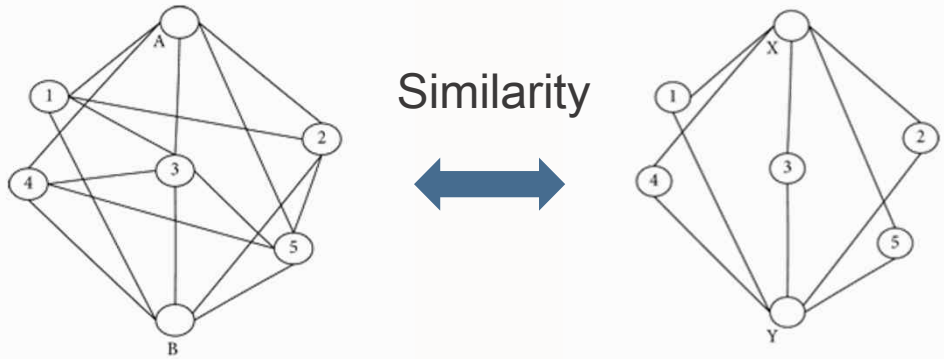
Implicit Relationship을 찾아낼 Entity Pair의 추출 및 랭킹

Algorithm

The “Link Prediction on Social Network” Algorithm (Dong et al. 2013)

Intuition

두 node A, B와 Common Neighbors로만 이루어진 SubGraph가
원래의 그래프와 비슷할 수록, 두 노드 A, B는 연관될 가능성이 높다.



Formula

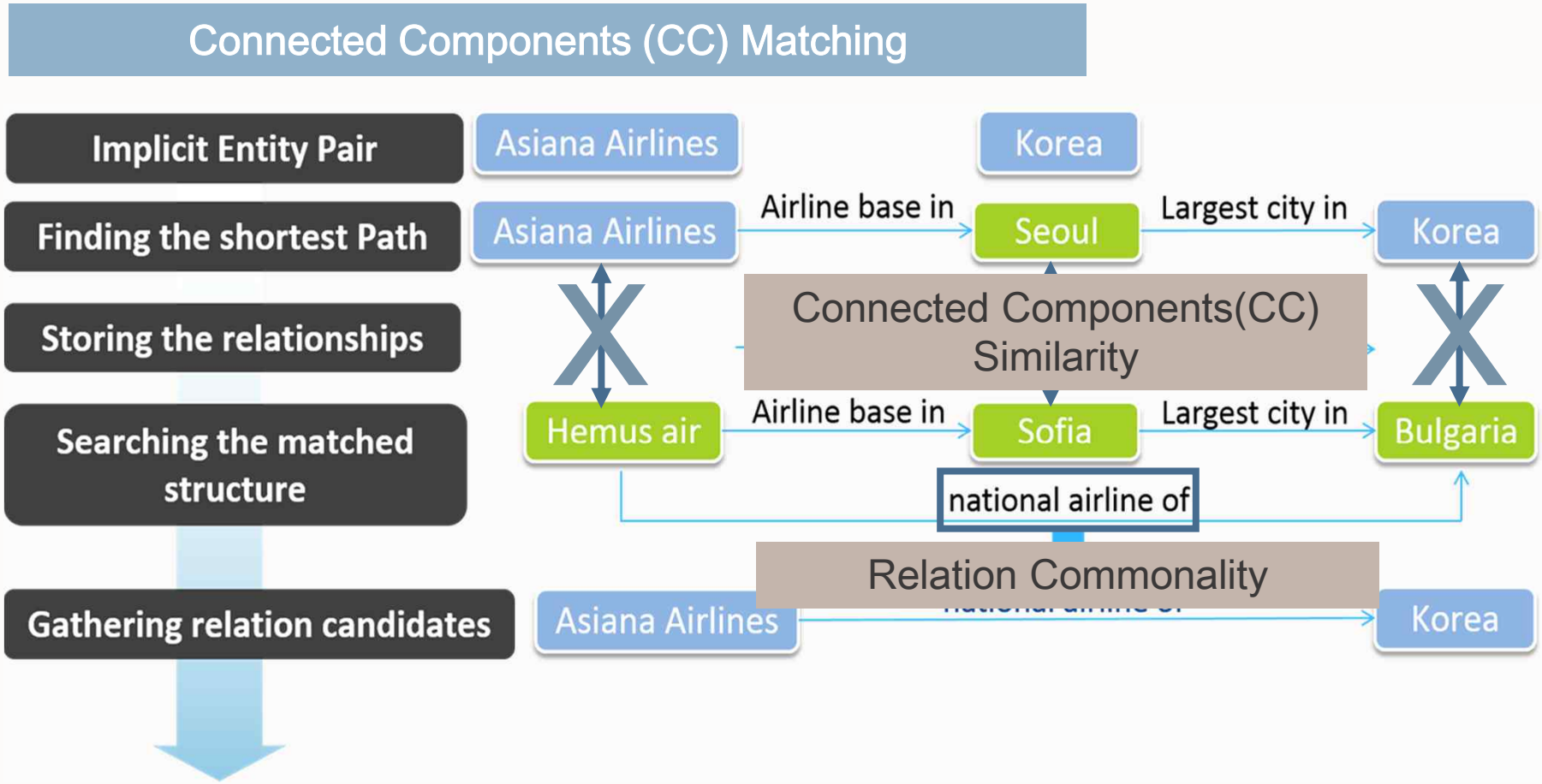
$$\text{Similarity}^{CNGF}(e_1, e_2) = \sum_{z \in \Gamma(e_1) \cap \Gamma(e_2)} \frac{|\text{degree}(z)|}{bg\ d_{cn}}$$

Output

30만 개의 Triple set → 약 787만 개의 Implicit Entity Pair 추출

예. (the us., united states : 61.88), (Europe, European union : 43.41)

2.5 Discovering Implicit Relationships



3
EVALUATION

3.1 Evaluation Dataset

Characteristics of Dataset

	Open Information Extraction	yago select knowledge
실험데이터 크기	300,000 Triples	300,000 Triples
원본데이터 크기	3,000,000 Triples (Lin et al 2012)	5,652,463 Triples (2015.06)
Unique Entity	983,410	2,824,974
Unique Relation	540,620	36
Example	{Ben Kingsley, was born in, Yorkshire}	{Bruce Murray, plays_for, FC Luzern}
Source	ClueWeb `09	Wikipedia, WordNet
Characteristic	• Entity의 Disambiguation 필요 • 다양한 Relation	• Entity의 표현에 일정 부분 Disambiguation이 이루어짐 • 한정된 Relation

3.2 Evaluation Method

Correctness Judgment by Search Engine

Evaluation Process

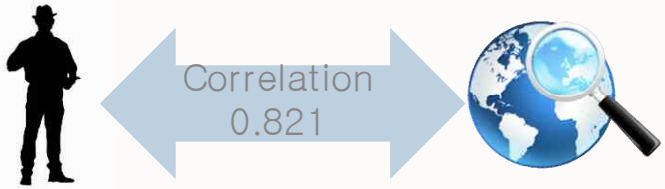
- 임의의 triple (e_1, r, e_2)를 실험 데이터셋에서 제거
- (e_1, e_2)를 query로 입력하여, implicit relation r`을 추론
- 새롭게 추출된 (e_1, r`, $e_2$)의 진위여부 평가 : (r=r`) 또는 (e_1,r`,e_2)가 옳으면 정답
- 실험 데이터셋의 모든 triple을 순차적으로 평가

Judgment by Search Engine

Query triple → phrase
e.g. 'seoul is a city of Korea'
Search 완전일치검색
Correct 결과 문서가 2개 이상일 경우

Validation Test

Google bing



156개 정답 판단
44개 오답 판단

141개 정답 판단
59개 오답 판단

- 검색엔진이 정답으로 평가한 트리플 전체는 사람도 정답으로 판별
- 검색엔진이 오답으로 평가한 트리플 중 일부는 사람이 정답으로 판별
- 사람이 오답으로 평가한 트리플 전체는 검색엔진도 오답으로 판별

3.3 Comparison with Other Methods

SHERLOCK

- 19,785 results from 7,368 patterns

Baseline #1

- Basic Transitive Inference Rule
- Rule : Entity A ⊂ Entity B ⊂ Entity C' → Entity A ⊂ Entity C
- Relation : 'in'을 포함하는 모든 Relation (포함관계)
- 앞의 Relation보다 뒤의 Relation이 더 포괄적일 때만 적용

예. {A, is a city of, B} + {B, is located in, C} → {A, is located in, C}

Baseline #2

- Rank by Cosine Similarity Value

Baseline #3

- Rank by Connected Components Similarity Value

Baseline #4

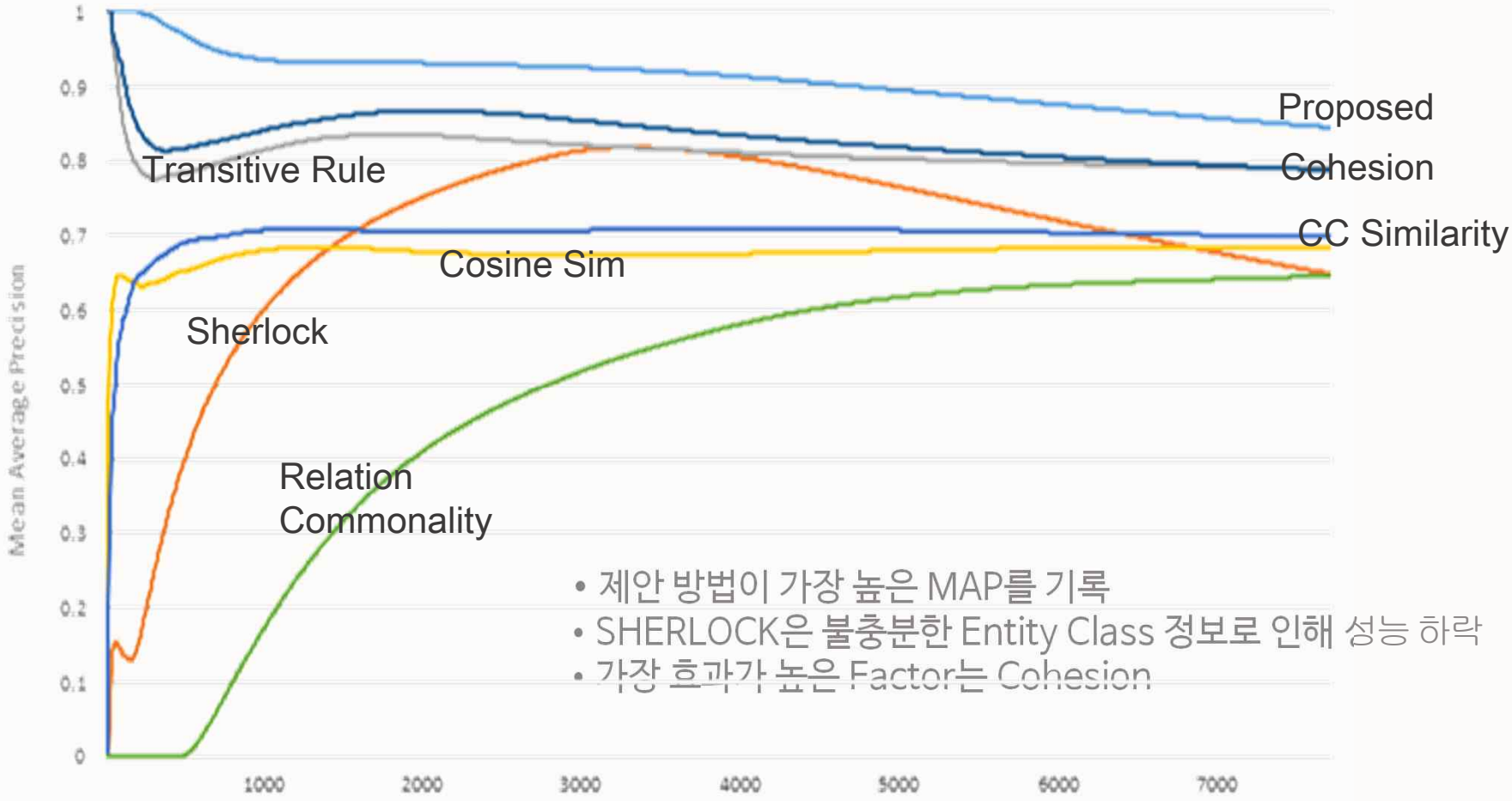
- Rank by Relation Commonality Value

Baseline #5

- Rank by Cohesion Value

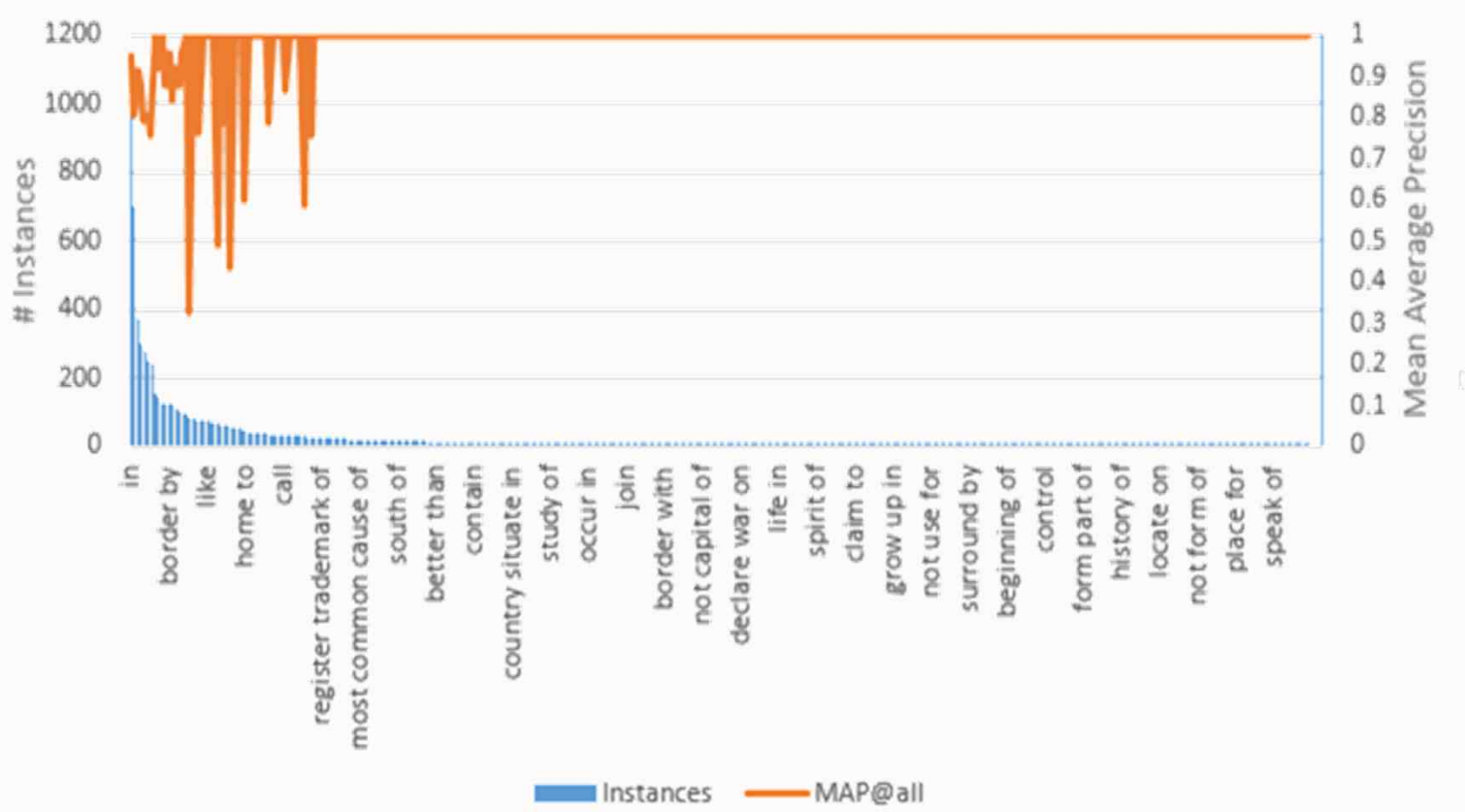
3.3 Comparison with Other Methods

Evaluation Result (OIE – Reverb Data Set)



3.5 Evaluation with Individual Relations

Evaluation on OIE ReVerb Dataset



- Instance가 1~2개인 Relation의 정답률이 매우 높음
  - Instance가 적은 Relation은 상대적으로 등장횟수가 적고 정답이 될 확률이 낮은 경우가 많음
  - 하지만 정답으로 선택될 경우에는 높은 Evidence를 가지고 있는 상태이므로 정답확률이 높아짐

4
CONCLUSION

4.2 Conclusion and Future Work

Conclusion

- 새로운 문제 정의 : Open Knowledge Acquisition
- Knowledge Acquisition을 위해 Graph Structure를 활용한 새로운 방법 제안
- 다양한 데이터셋에 적용 가능
- 다양한 연구와 결합되어 시너지 생성 가능

  Entity Resolution, Entity Linking, Relation clustering…

Future Work

- Entity Linking 및 Semantic class learning을 통한 매칭 오류 제거
- Entity Resolution과 Relation Clustering을 통한 재현율 향상
- 향상된 그래프 탐색 및 매칭 알고리즘을 통한 성능 개선


Q&A
QUESTIONS & ANSWERS SESSION