Machine Reading for QA:
Knowledge Enrichment from Korean Wikipedia

함영균 박사과정 / 최기선 교수

Context of the talk

- Exobrain project (2013-2022)
  - Industry-Academia-Research Institute Consortium
  - Korean language-based Question Answering
  - 1st phase (2013.5 – 2017.2)
    - wiseQA and wiseKB
      - E.g., IBM Watson
- wiseKB
  - Ontology-based structuring of knowledge from unstructured text
  - DBPedia and Multilingualism
    - To enhance the Korean knowledge structuring
  - http://xb.saltlux.com/

CONTENTS

1. QA and NLQ400
2. Type 1
   - Questions whose answers “can” be found in Korean DBpedia and Wikipedia in the form of “complete sentence(s)”
   - And, **OUR APPROACH** to construct KB from Wikipedia
3. Type 2
   - Questions whose answers “can” be found in Korean DBpedia and Wikipedia in the form of “a word” or “a series of words”
   - And, **OUR APPROACH** to construct KB from Wikipedia
4. AND, our other efforts for Multilinguality KB
   - Ontology Alignment between Korean properties and English DBO properties
5. What’s next?

1. QA, AND NLQ400

Introduction:
Question Answering?

- Question Answering (QA)
  - The oldest task in natural language processing (1961)
  - Related work
    - IBM Watson won the Jeopardy in 2011
    - Apple Siri, Samsung S-voice, LG Quick Voice, …

Question

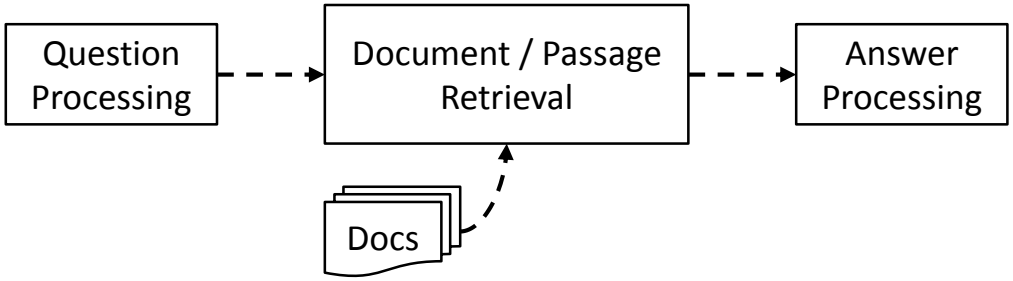
Who was the person that dropped out of Reed College, founded Apple Inc., and fought against pancreatic cancer?

Answer

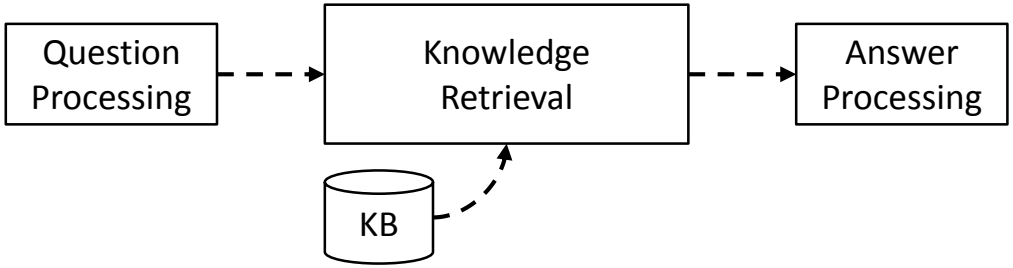
Steve Jobs

Introduction: Paradigms for QA

- IR-based approaches
  - TREC, Naver, Daum



- Knowledge-based approaches
  - Wolfram Alpha;



Introduction: KB for QA?

- Knowledge is traditionally written in natural language:
  - Only 15-20% of information in structured data
    (R Blumberg et al., 2003)

- So, in REAL QA,

  How about Wikipedia or DBpedia?

Introduction to NLQ400

- “ NLQ400 ”
  - a set of 384 pairs of Korean natural language questions and answers (NLQ&A) collected from various quiz shows.
  - questions were selected from many different fields of study.
  - e.g., Art, Biography, History, Geography, Science, Technology, etc.

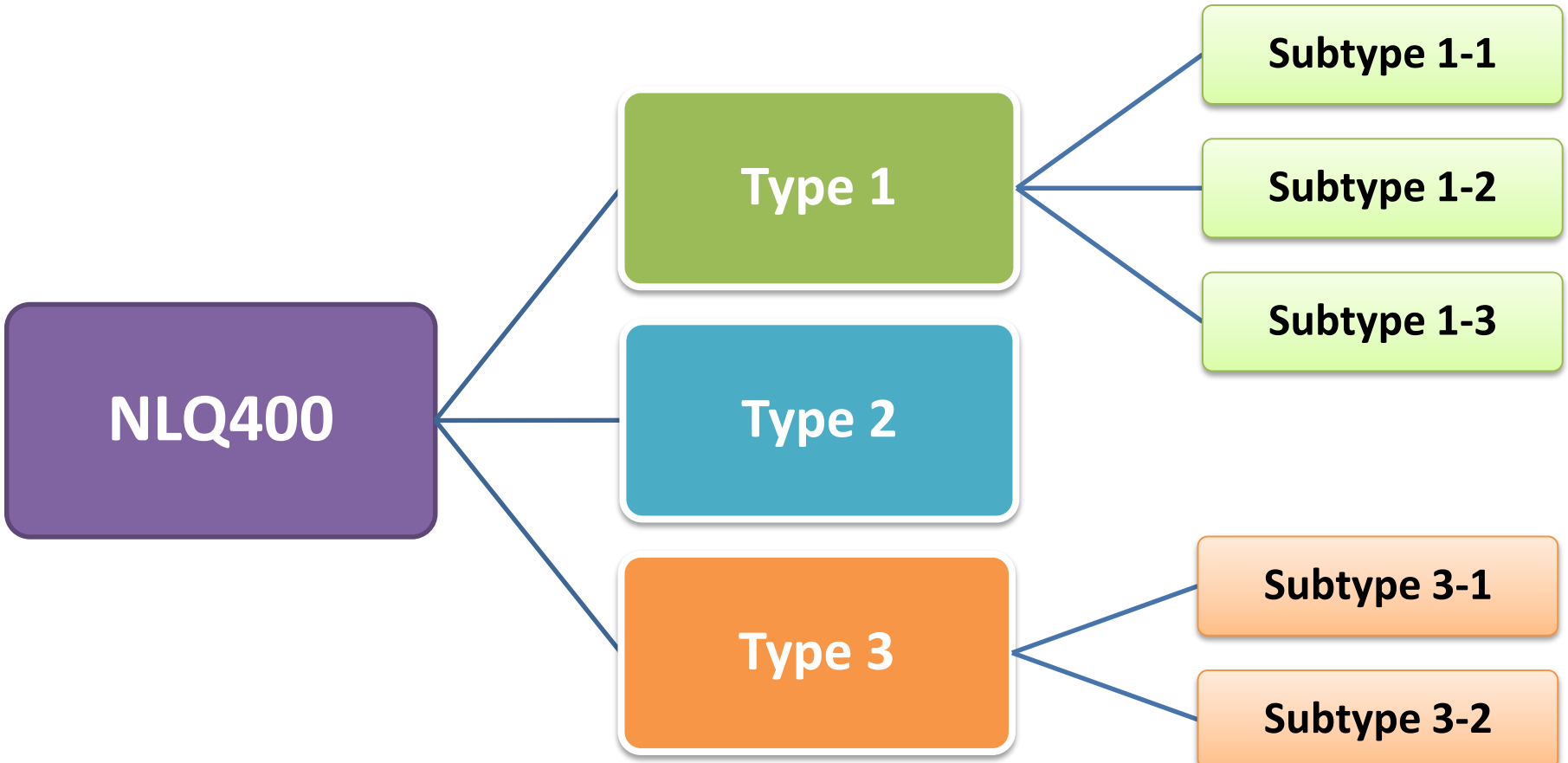
  - all answers are in one of the following formats.
    - multiple-choice
    - short-response
    - write-out

Purpose of NLQ400

- NLQ400 Set is to be used as the criteria for evaluating :
  - the state-of-the-art of knowledge bases (KB) such as Korean DBpedia and Wikipedia in particular.
  - the accuracy and performance of Question-Answering (QA) systems especially designed to process the Korean natural language.

NLQ400 Type Categorization

- These 384 Korean NLQ&A pairs in NLQ400 Set can be categorized into 3 main-types and 5 sub-types, depending on whether or not the answer to the question can be found in Korean DBpedia and Wikipedia.



Type 1 : **343 out of 384 (89.32%)**

= Questions whose answers “can” be found in Korean DBpedia and Wikipedia in the form of “complete sentence(s)”.

Type 2 : **24 out of 384 (6.25%)**

= Questions whose answers “can” be found in Korean DBpedia and Wikipedia in the form of “a word” or “a series of words”.

Type 3 : **17 out of 384 (4.42%)**

= Questions whose answers “cannot” be found in Korean DBpedia and Wikipedia.

2. TYPE 1:

QUESTIONS WHOSE ANSWERS “CAN” BE FOUND IN KOREAN DBPEDIA AND WIKIPEDIA IN THE FORM OF “COMPLETE SENTENCE(S)”.

Subtype 1-1 : **95 out of 384 (24.73%)**

= Questions whose answers "can" be found in Korean DBpedia and Wikipedia "directly" from "one single sentence".

e.g.,

Ko.NLQ#81 :

대한민국의 제16대 대통령은 누구인가?

Who was the 16th President of the Republic of Korea?

Answer in Ko.DBpedia & Wikipedia :

노무현(盧武鉉, 1946년 9월 1일 ~ 2009년 5월 23일)은 대한민국의 제16대 대통령이다.

Roh Moo-hyun (盧武鉉, 1 September 1946 - 23 May 2009) was the 16th President of the Republic of Korea.

Subtype 1-2 : **223 out of 384 (58.07%)**

= Questions whose answers "can" be found in Korean DBpedia and Wikipedia "directly" from "two or more sentences".

e.g.,

Ko.NLQ#230 :

Who was the person that dropped out of Reed College, founded Apple Inc., and fought against pancreatic cancer?

리드 대학교를 중퇴하고, 애플을 창업하였으며, 췌장암으로 투병한 사람은 누구인가?

Answer in Ko.DBpedia & Wikipedia :

Steven Paul "Steve" Jobs (February 24, 1955 – October 5, 2011) was an American pioneer of the personal computer revolution of the 1970s. Jobs briefly attended Reed College in 1972 before dropping out, …. He co-founded Apple in 1976 …. In mid-2004, he announced to his employees that he had a cancerous tumor in his pancreas.

스티븐 폴 "스티브" 잡스(1955년 2월 24일 ~ 2011년 10월 5일)는 미국의 기업인이었다. 1972년 리드 대학교에 다니다 1학기만 수강한 후 중퇴하였으며, …. 1976년 스티브 워즈니악, 로널드 웨인과 함께 애플을 공동 창업하고, …. 한편 그는 2004년 무렵부터 췌장암으로 투병생활을 이어왔다.

Subtype 1-3 : **25 out of 384 (6.51%)**

= Questions whose answers "can" be found in Korean DBpedia and Wikipedia "indirectly" where "some processes of reasoning are required around several sentences".

e.g.,

Ko.NLQ#12 :

Who was the last princess of Joseon?

조선의 마지막 황녀는 누구인가?

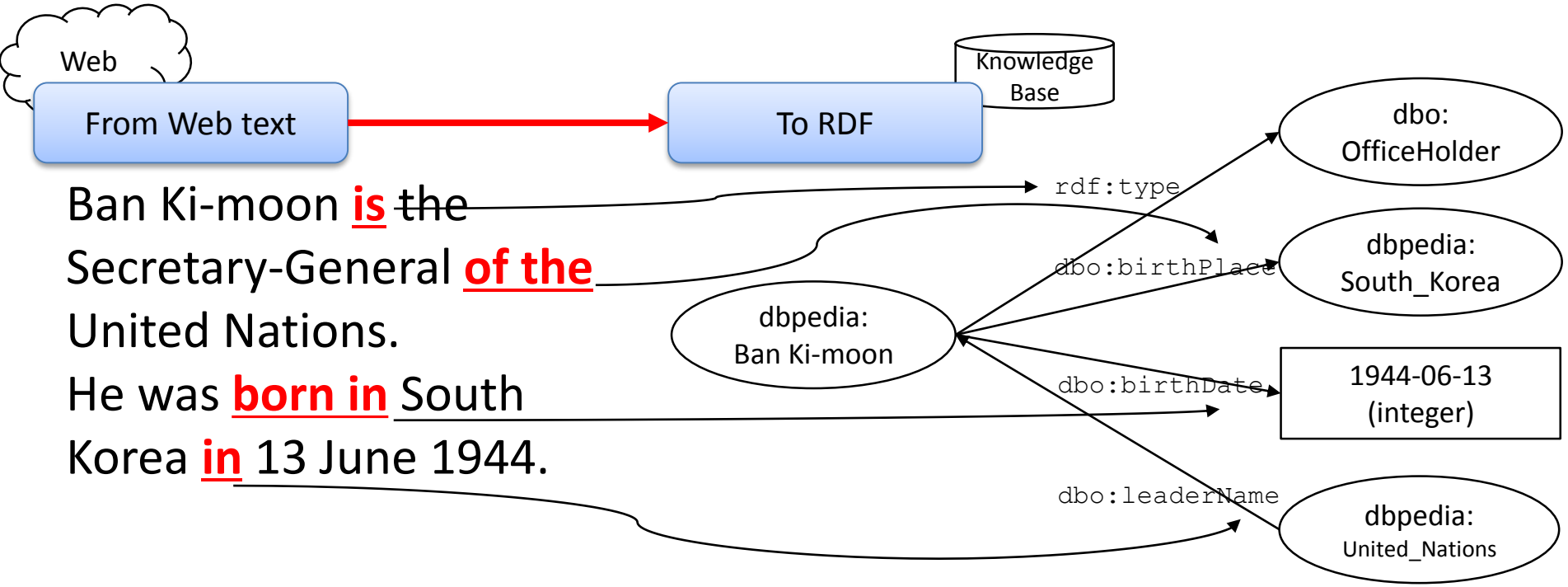
Answer in Ko.DBpedia & Wikipedia :

Princess Deokhye (德惠翁主, 25 May 1912 - 21 April 1989) was the daughter of Emperor Gojong and the noblewoman Yang. Gojong was the 26th and the last king of Joseon. Gojong used to have 9 sons and 4 daughters, but only 3 sons and 1 daughter survived, so that Princess Deokhye was actually the only daughter of Gojong.

덕혜옹주(德惠翁主, 1912년 5월 25일 ~ 1989년 4월 21일)는 조선의 제26대 왕이자 대한제국의 초대 황제였던 고종과 귀인 양씨 사이에서 태어난 딸이다. 고종은 조선의 제26대이자 마지막 임금이다. 고종은 총 9남 4녀의 자녀가 있었지만 3남 1녀만이 성년이 될 때까지 생존하여 덕혜옹주가 사실상 유일한 딸이 되었다.

Learning by Reading for TYPE 1

What is “Learning by Reading”?



- related work:
  - NELL, Watson
  - BOA (Bootstrapping LOD)

Introduction:
Information Retrieval vs Knowledge Retrieval

	IR	KR
Model	Statistical & probabilistic model	Semantic model, inference model
Query	Natural language	Knowledge structure (SQL, SPARQL,,,)
Organization	Table, Index	Knowledge unit, Knowledge Structure (schema, ontology,,,)
Representation	Natural language, Markup language	Graph, Predicate logic, production rule, semantic network, ontology,,,
Storage	Document collections	Knowledge base
Retrieved Result	Sections or documents	A set of knowledge unit

Knowledge?
→ Structured information?

Introduction:
Knowledge Extraction?

- Knowledge is traditionally written in natural language:
  - Only 15-20% of information in structured data
    (R Blumberg et al., 2003)

- Knowledge extraction (KE)?
  - KE system Automatically produce machine-readable information from Web contents (structured/semi-structured/unstructured)
- KE from text?
  - ~= Constructing (or, bootstrapping or Populating,,,) KB from text

- 2 major approaches in KE from text
  - 1) OpenIE
  - 2) Ontology population (producing A-Box)

Introduction:
Knowledge Extraction from text – Open IE

Text

김부식은 삼국사기를 쓴 작가이다

Open IE approach

이다(김부식, 작가)
쓰다(김부식, 삼국사기)

- Lexicalized predicate (relation)
  - No supervision
  - No pre-defined schema
  - Scalability
- Related work
  - openIE, NELL, Watson,,,

Introduction:
Knowledge Extraction from text – Ontology population

Text

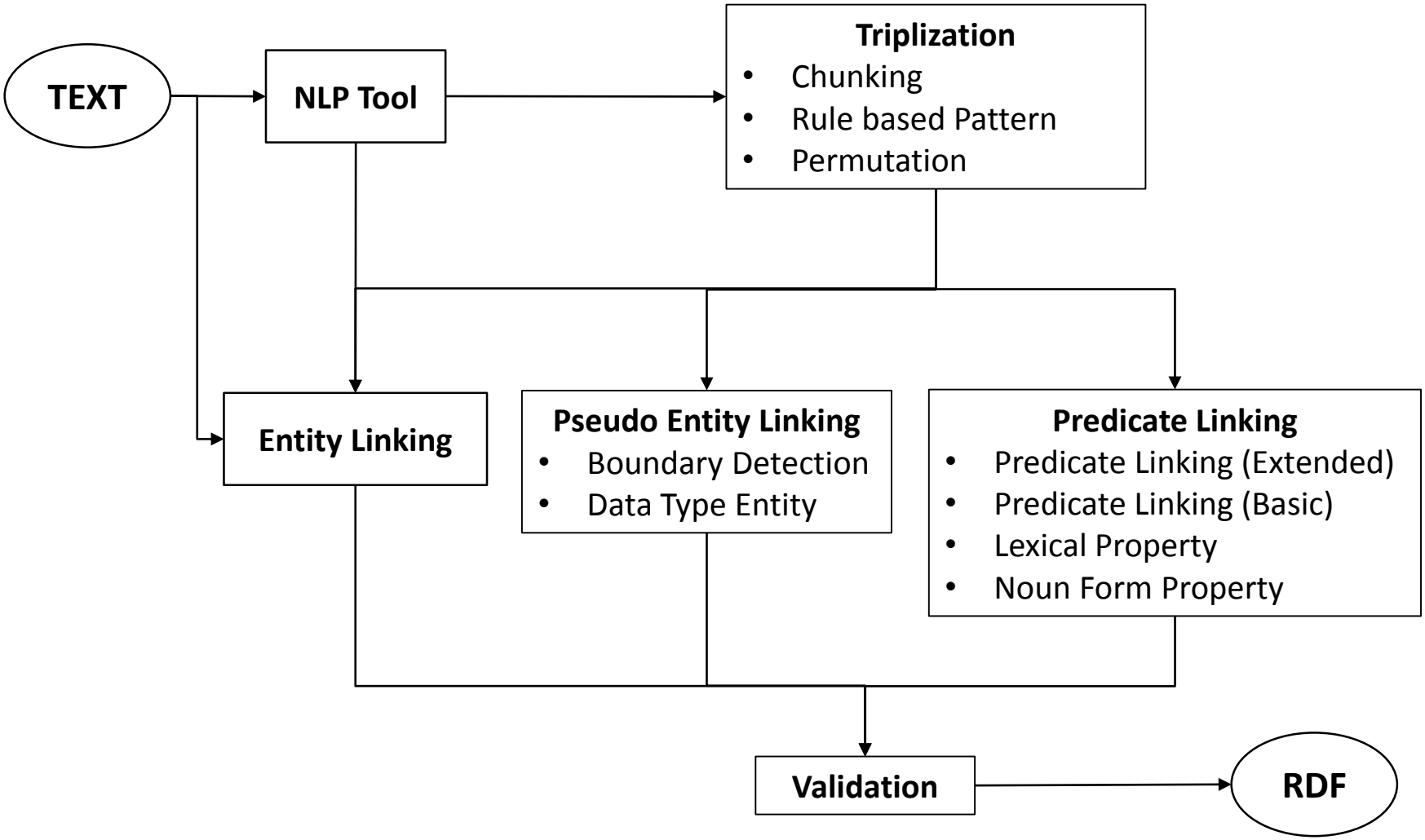
김부식은 삼국사기를 쓴 작가이다

Ontology Population

<res:김부식> <rdf:type> <class:작가>
<res:김부식> <property:write> <res:삼국사기>

- Ontological predicate (relation)
  - Supervision
  - Pre-defined schema (property)
  - Well-defined meaning
- Related work
  - BOA, Ontology Lexicalization, and other ontology population researches,,,

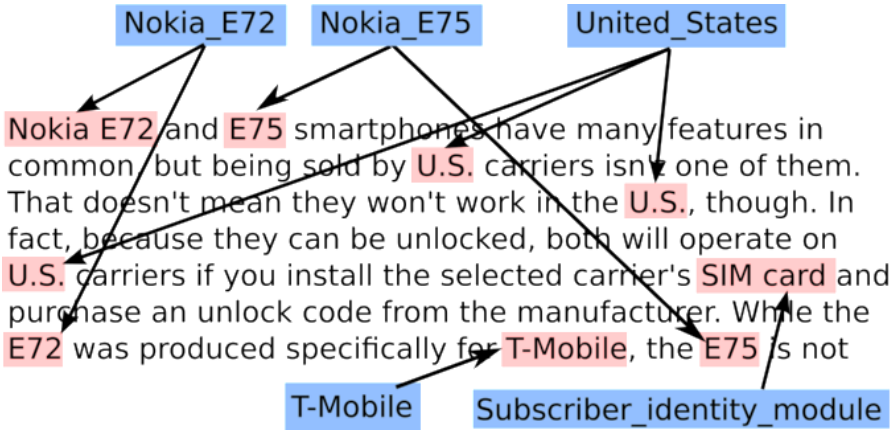
L2K Architecture



Learning by Reading: Entity Linking

Entity Linking

- The task of determining the identity of entities mentioned in text.
  1. Named Entity Detection(NED): Detecting entity 'mentions'.
  2. Entity Disambiguation: Linking each mention with an appropriate entity.



Youngsik Kim, Key-sun Choi, 2015,
Entity Linking Korean Text: An Unsupervised Learning Approach using Semantic Relations, CoNLL

Motivation #1: Handling NED

- Current approaches (such as AGDISTIS, 2014) use existing NED systems, such as FOX.
  - FOX reports a best-case performance of over 95% f-score.
    - An ensemble learning approach utilizing multiple classifiers.
    - Using English datasets. What about Korean?

- NED for certain languages like Korean is hard!
  - Lack of capitalization and strict whitespace separation.

- Domain mismatches

→ **We need to handle NED as well!**

a surface form

그의 목적은 아라비아 반도 여행이다.

His goal is to tour the Arabian Peninsula.

not a surface form

보급품이 반도 남지 않았다.

Not even half of the supplies remain.

Youngsik Kim, Key-sun Choi, 2015,
Entity Linking Korean Text: An Unsupervised Learning Approach using Semantic Relations, CoNLL

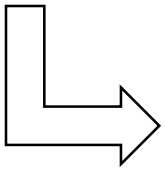
Motivation #2: Problems of Past Research

- Youngsik Kim, Younggyun Hahm, Dosam Hwang, Key-Sun Choi, A Non-morphological Entity Boundary Detection Approach for Korean Text, NLIWoD, 2014.

1. This research used a specific Wikipedia dump to construct the entity dictionary and train the system, meaning that the domain was fixed to the Korean Wikipedia.

<대한민국> [[수나라|수]], [[당나라|당]]과 싸우며
<1695년_잉글랜드_총선> 어떠한 [[정당|당]]이나 [[정부|정부]]도
…

Wikipedia links



Entity Detection dictionary

- 수 → 수나라
- 당 → 정당, 당나라
- 정부 → 정부
- …

2. The training data used by this system:
   - Manually annotated golden standard data.
   - Link data within the Wikipedia dump.

- Training data is hard to obtain because it requires manual intervention.

Youngsik Kim, Key-sun Choi, 2015,
Entity Linking Korean Text: An Unsupervised Learning Approach using Semantic Relations, CoNLL

OUR APPROACH (especially, for Multilingual)

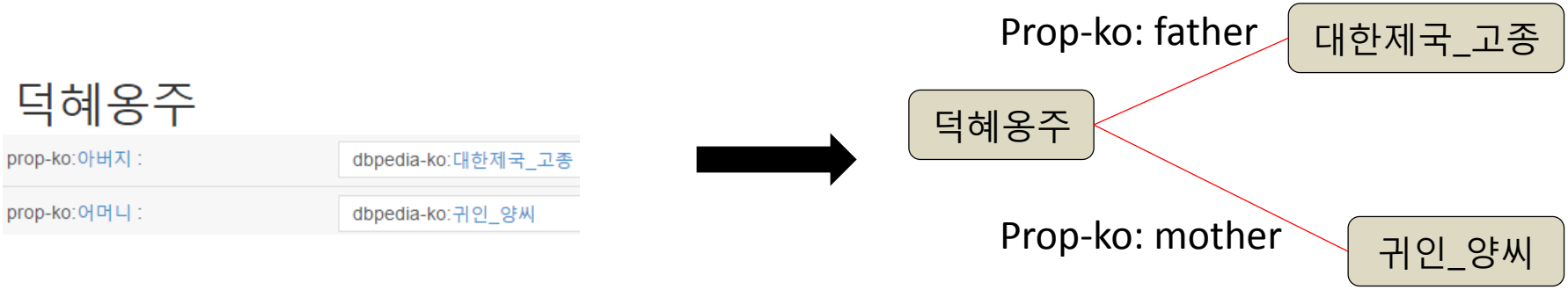
- Develop an entity linking system that:

1. Can be trained to detect entities within any given domain.
2. Only requires easy-to-obtain resources for training.
   - No resources that require human intervention.
3. Performs at least as well as the state-of-art system for Korean.
4. Can be easily configured for other resource-scarce languages.

Youngsik Kim, Key-sun Choi, 2015,
Entity Linking Korean Text: An Unsupervised Learning Approach using Semantic Relations, CoNLL

Semantic Relations

- Semantic relations: 2 entity URIs which have a relationship in the KB.

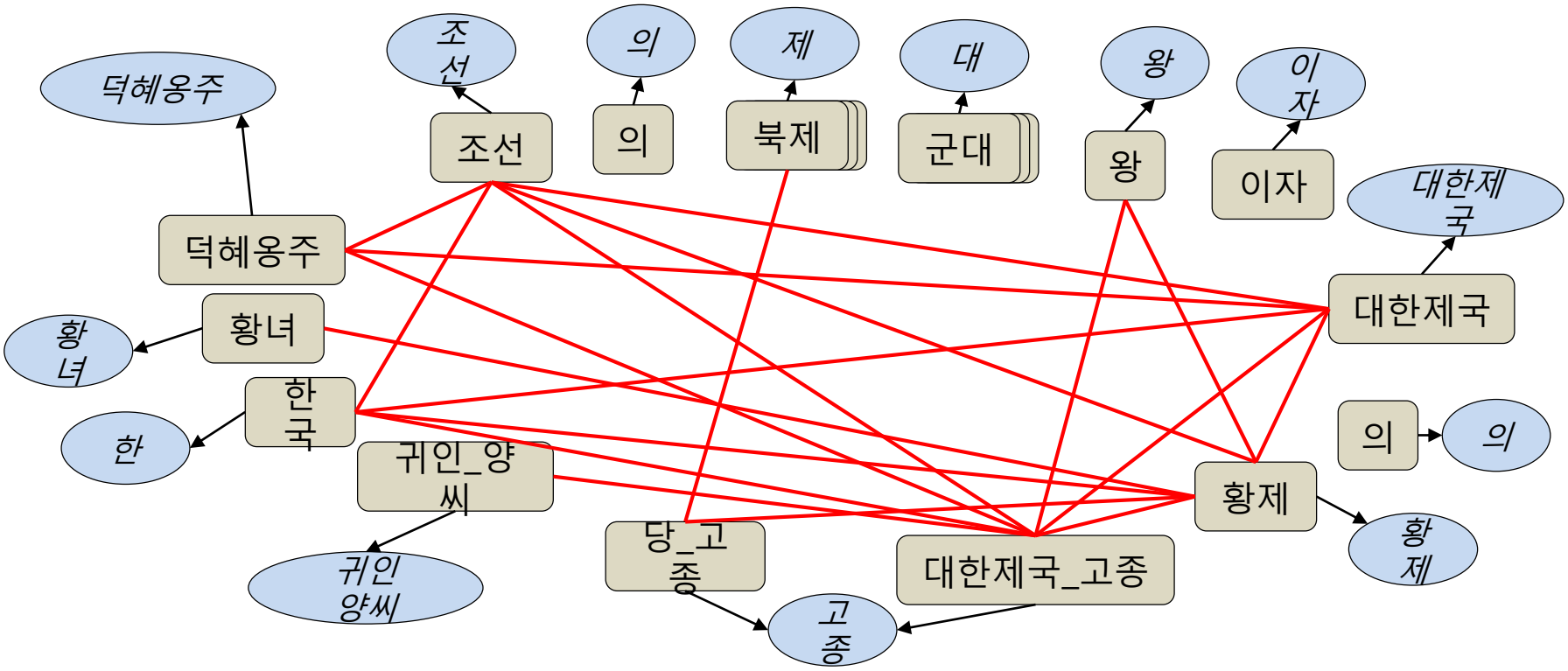


- Given any text & a list of entity candidates, we can build a graph of semantic relations.

Youngsik Kim, Key-sun Choi, 2015,
Entity Linking Korean Text: An Unsupervised Learning Approach using Semantic Relations, CoNLL

Graph of Semantic Relations

덕혜옹주는 조선의 제26대 왕이자 대한제국의 초대 황제 고종과 귀인 양씨의 한 황녀이다.



Youngsik Kim, Key-sun Choi, 2015,
Entity Linking Korean Text: An Unsupervised Learning Approach using Semantic Relations, CoNLL

- Observation: High semantic relation degree → Actual entity?

Entity Candidate	#
(고종, 대한제국_고종)	7
(조선, 조선)	5
(대한제국, 대한제국)	5
(한, 한국)	4
(황제, 황제)	4
(덕혜옹주, 덕혜옹주)	3
(고종, 당_고종)	2
(왕, 왕)	2
(제, 북제)	1
(귀인 양씨, 귀인_양씨)	1
(황녀, 황녀)	1
(의, 의)	0
(이자, 이자)	0

(sample threshold ∂ = 3)

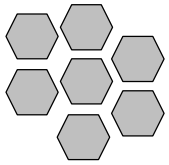
Semantic relation degree ≥ ∂.
Most are actual entities. (Except (한, 한국))

Are actual entities.
Semantic relation degree ≤ ∂.
If we only use 'Semantic Relation Degree' to detect entities, we miss these!
(Or pick incorrect (제, 복제), (고종, 당_고종) as well)

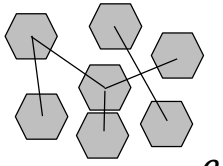
Youngsik Kim, Key-sun Choi, 2015,
Entity Linking Korean Text: An Unsupervised Learning Approach using Semantic Relations, CoNLL

Training Process

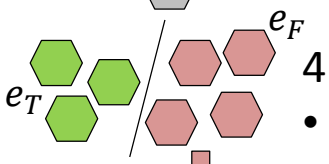
1. Part-of-speech tagging
   - Perform morphological analysis on the given text.



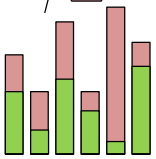
2. Entity candidate detection
   - Parse a list of all possible entity candidates from the given text.



3. Semantic relation search
   - Build a graph of all semantic relations.



4. Entity candidate classification
   - Classify each candidate based on the amount of semantic relations.

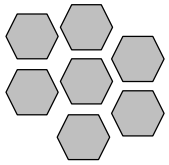


5. Feature training
   - Train features based on the entity candidate classification.

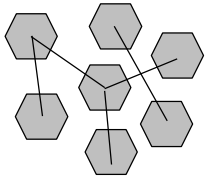
Youngsik, Key-sun Choi, 2015,
Entity Linking Korean Text: An Unsupervised Learning Approach using Semantic Relations, CoNLL

Entity Linking Process

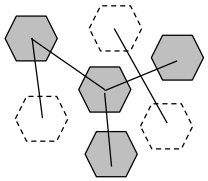
1. Part-of-speech tagging
   - Perform morphological analysis on the given text.



2. Entity candidate detection
   - Parse a list of all possible entity candidates from the given text.



3. Semantic relation search
   - Build a graph of all semantic relations between.

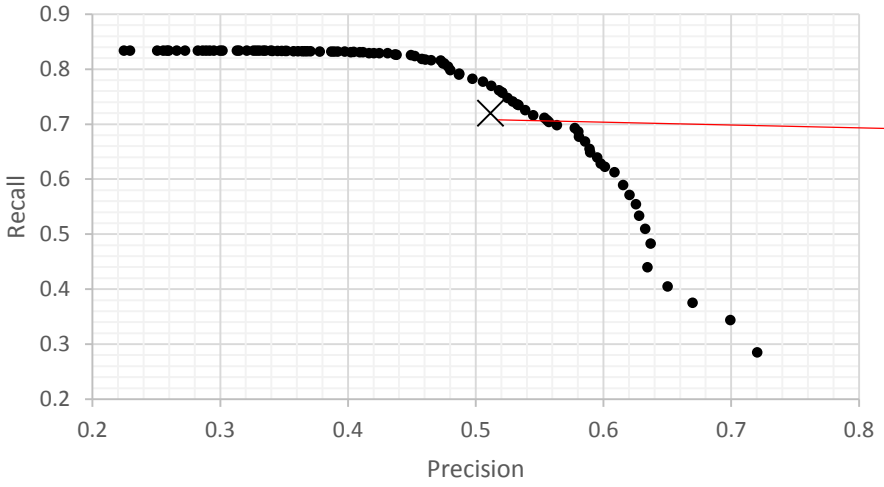


4. NED & disambiguation
   - Score all entity candidates, and select only candidates with a score higher than a given threshold $0 \leq \gamma \leq 1$.
   - For overlapping candidates, only select the candidate with the highest score.

Youngsik, Key-sun Choi, 2015,
Entity Linking Korean Text: An Unsupervised Learning Approach using Semantic Relations, CoNLL

System Evaluation

- Training data: 4,000 Korean Wikipedia documents.
- Test data: 60 Korean Wikipedia abstract documents manually annotated with entity data.
  - 3 annotators, 2-person agreement

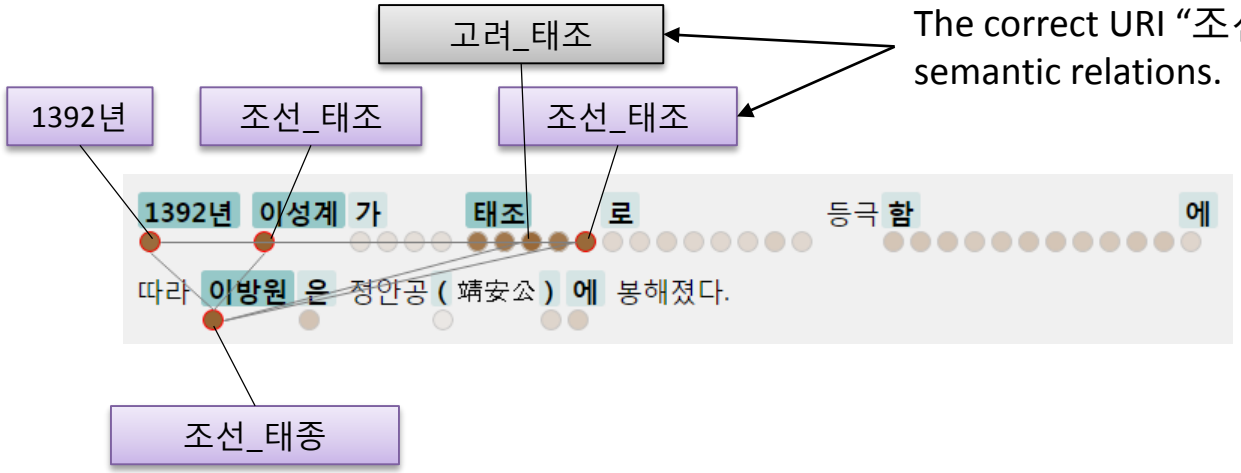


- F-score for our system: 0.630
- Better than the f-score of (Kim, 2014) (0.598)

Youngsik Kim, Key-sun Choi, 2015,
Entity Linking Korean Text: An Unsupervised Learning Approach using Semantic Relations, CoNLL

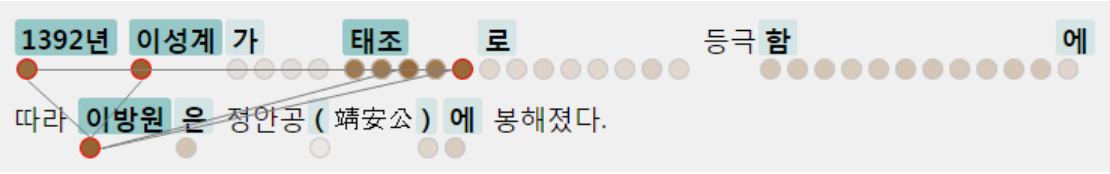
Entity Linking Example

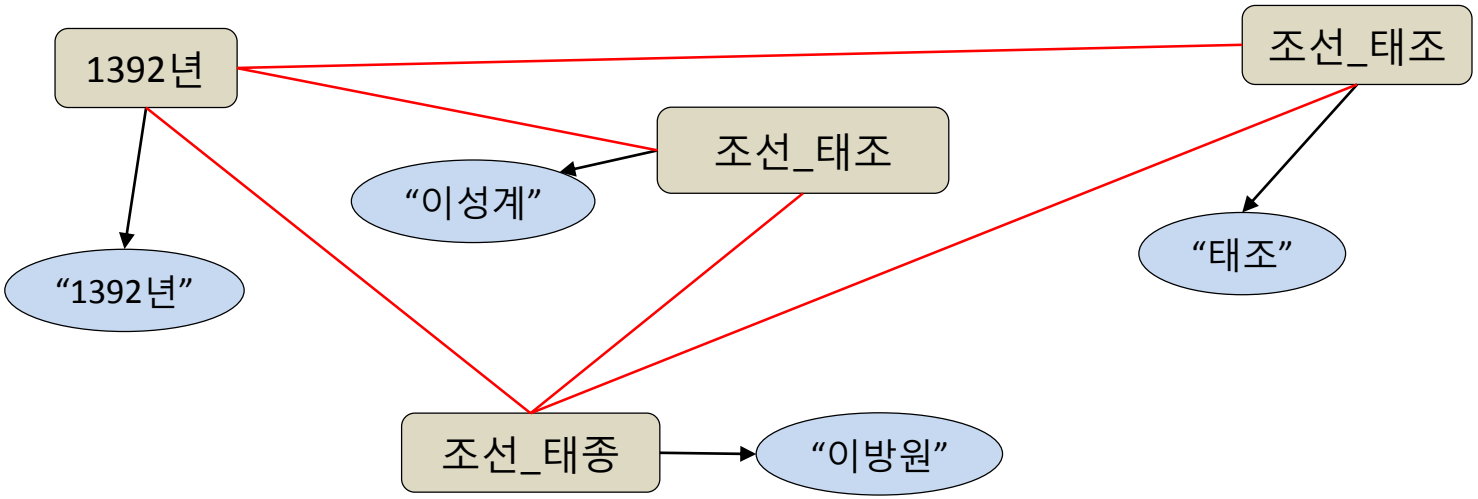
1392년 이성계가 태조로 등극함에 따라 이방원은 정안공(靖安公)에 봉해졌다.



Youngsik Kim, Key-sun Choi, 2015,
Entity Linking Korean Text: An Unsupervised Learning Approach using Semantic Relations, CoNLL

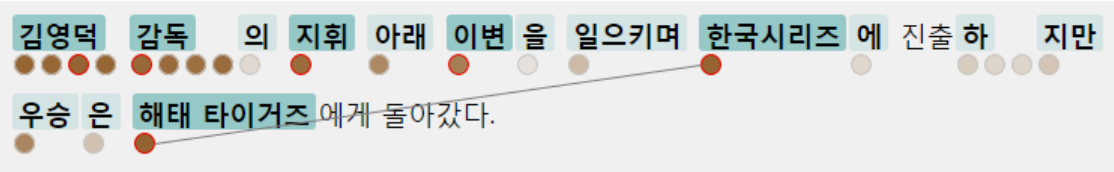
Entity Linking Example





Youngsik Kim, Key-sun Choi, 2015,
Entity Linking Korean Text: An Unsupervised Learning Approach using Semantic Relations, CoNLL

Entity Linking Example 2



- Dependence on the KB leads to errors when few relations can be found.

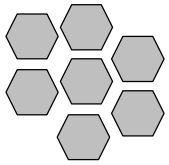
- “감독”
  - Answer: 야구_감독
  - Selected: 축구_감독

- “김영덕”
  - Answer: 김영덕_(야구인)
  - Selected: 김영덕

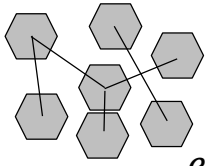
Youngsik Kim, Key-sun Choi, 2015,
Entity Linking Korean Text: An Unsupervised Learning Approach using Semantic Relations, CoNLL

Application on Japanese (Training Process)

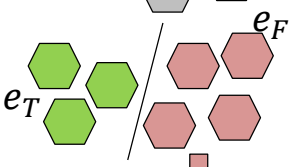
1. Part-of-speech tagging
   - Perform morphological analysis on the given text. (with a Japanese tagger)



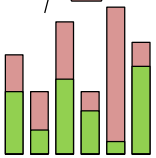
2. Entity candidate detection
   - Parse a list of all possible entity candidates (JP DBpedia) from the given text.



3. Semantic relation search
   - Build a graph of all semantic relations.



4. Entity candidate classification
   - Classify each candidate based on the amount of semantic relations.

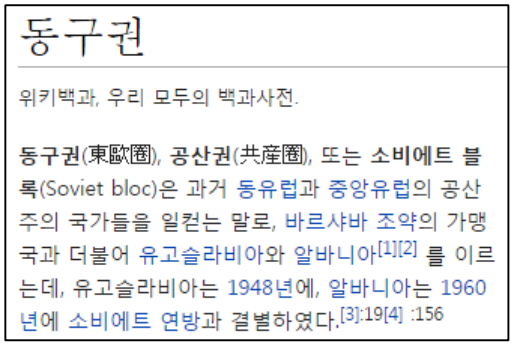


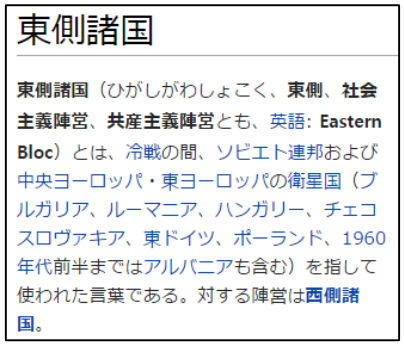
5. Feature training
   - Train features based on the entity candidate classification.

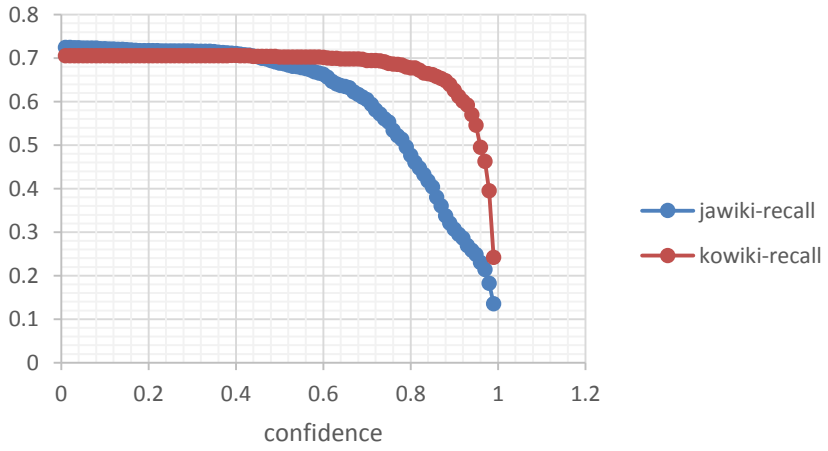
Youngsik Kim, Key-sun Choi, 2015,
Entity Linking Korean Text: An Unsupervised Learning Approach using Semantic Relations, CoNLL

Application on Other Languages: Japanese

- Training data: 4,000 Wikipedia documents.
- Test data: Links within Wikipedia documents.
  - Only recall is reliable.
- Our system was easily configurable to Japanese.
- Performance is relatively lower for Japanese.



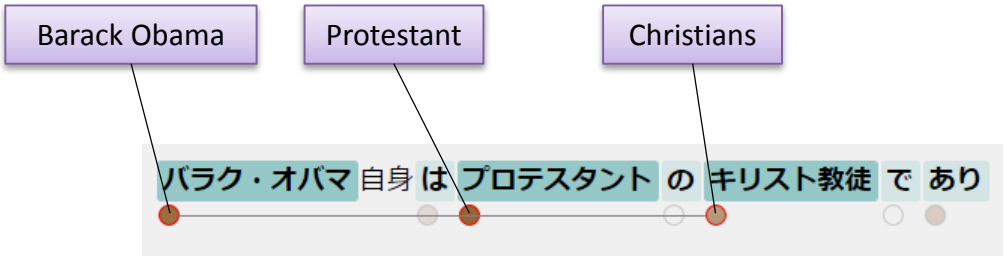




Youngsik Kim, Key-sun Choi, 2015,
Entity Linking Korean Text: An Unsupervised Learning Approach using Semantic Relations, CoNLL

Japanese Entity Linking Example

バラク･オバマ自身はプロテスタントのキリスト教徒であり
(Barack Obama is the Christians in the Protestant himself)



Youngsik Kim, Key-sun Choi, 2015,
Entity Linking Korean Text: An Unsupervised Learning Approach using Semantic Relations, CoNLL

Learning by Reading: Pseudo Entity Linking

Pseudo Entity Linking

- To support Entity Linking Module
  - Entity Linking Moudle is not perfect
  - Missing entity to possible resource
  - Supplement of Data Type Entity
    - String, Integer, Double, Date …

Pseudo Entity Linking

- To support Entity Linking Module
  - Pseudo-Entity Linking
    - Entity Boundary Detecting
      - Using NER, POS tag
      - Example
        - » "대한민국 서울특별시에서" (in Seoul, Korea)
          - Remove postpostion using POS tag("에서", in)
          - Detected boundary: 대한민국 서울특별시
      - Using both NP Chunk and Indivisual NP
        - » 대한민국 서울특별시
          - 대한민국 서울특별시 (Seoul, Korea)
          - 대한민국 (Korea)
          - 서울특별시 (Seoul)

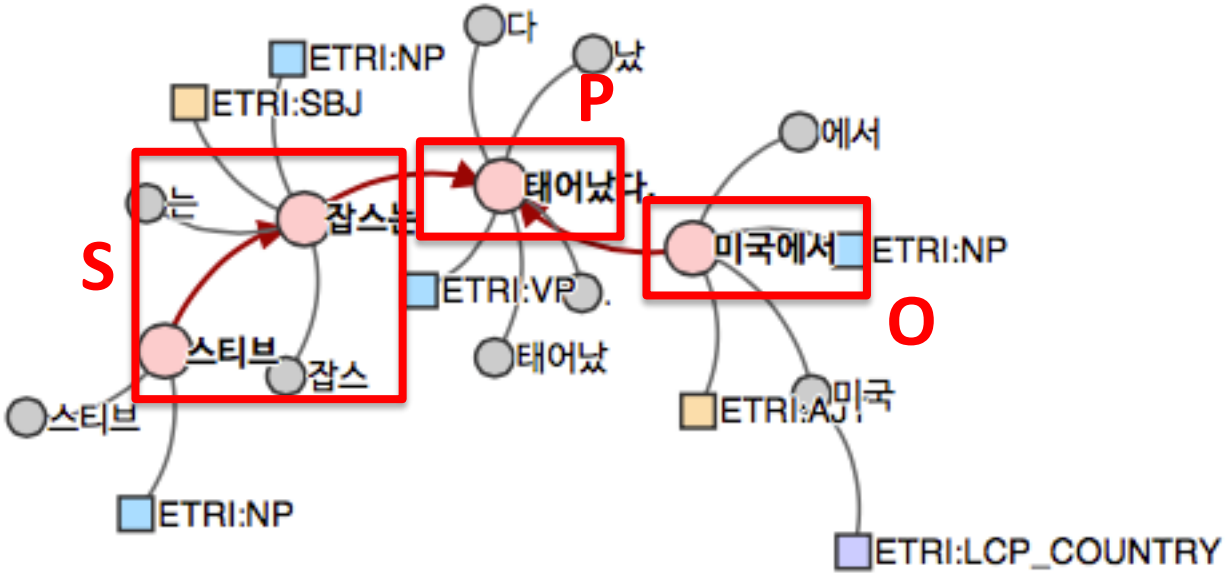
Pseudo Entity Linking

- To support Entity Linking Module
  - Example: Data Type Entity
    - Example
      - Time Representation
        - » "1545년 4월 28일" (in text)
          - Convert to "1545-4-28"
          - "1545-4-28"^^http://www.w3.org/2001/XMLSchema#date
        - » 4월 (in text, April)
          - Convert to "4"
          - "4"^^http://www.w3.org/2001/XMLSchema#integer

TRIPLIZATION

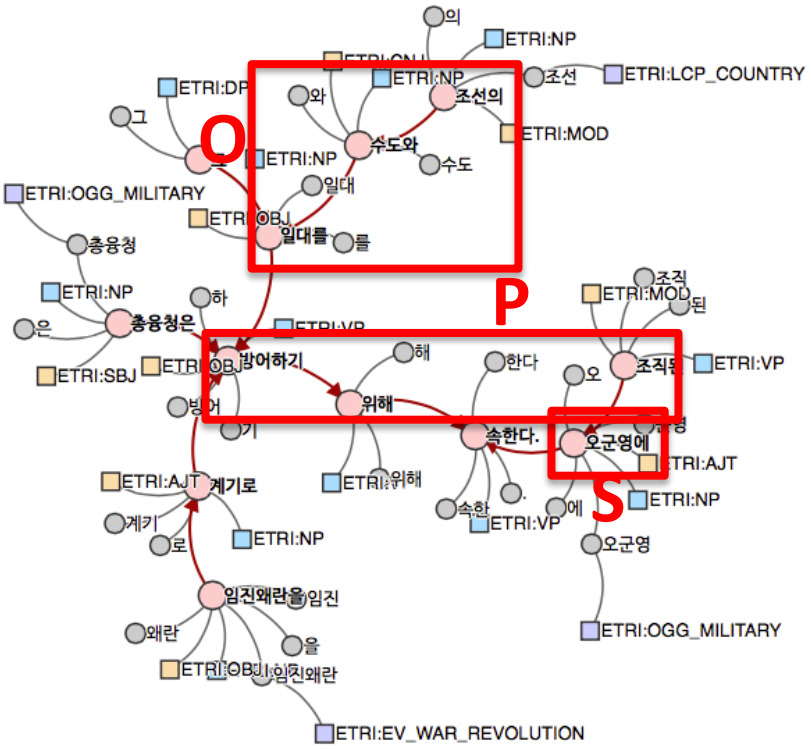
Triplization - example

- 1) S / O / P
  - [스티브 잡스는](S) [미국에서](O) [태어났다](P).



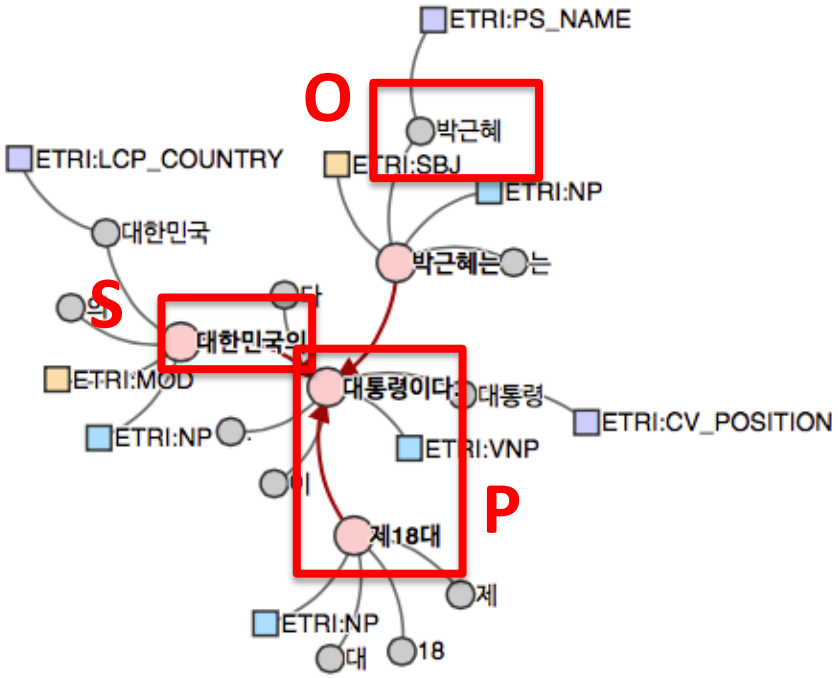
Triplization - example

- 2) O / P / S
  - 총융청은 임진왜란을 계기로 [조선의 수도와 그 일대를](O) [방어하기 위해 조직된](P) 오군영에(S) 속한다.



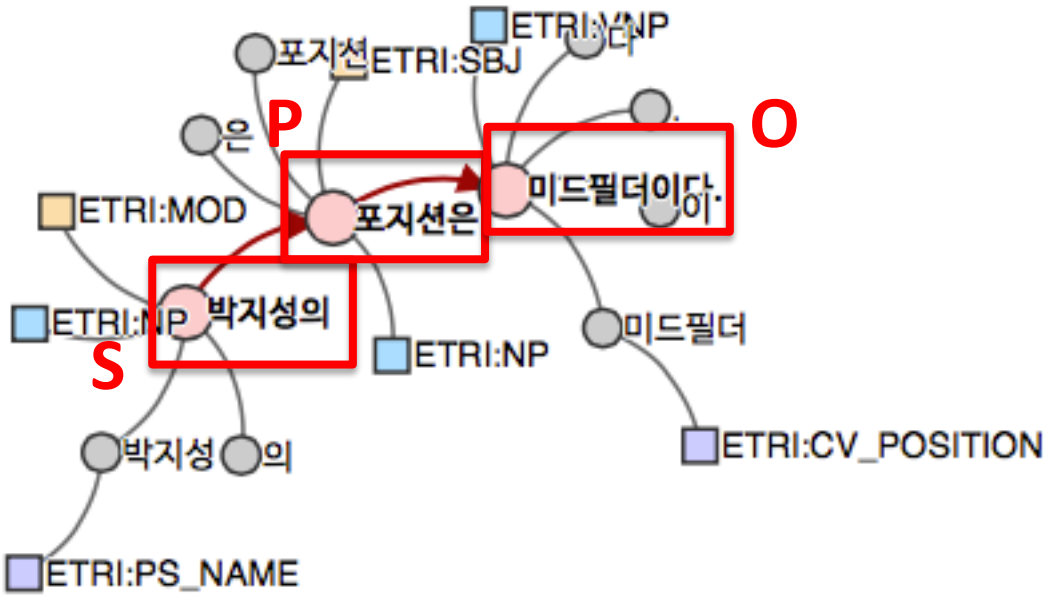
Triplization - example

- 3) O / S / P
  - [박근혜는](O) [대한민국의](S) [제18대 대통령이다](P).



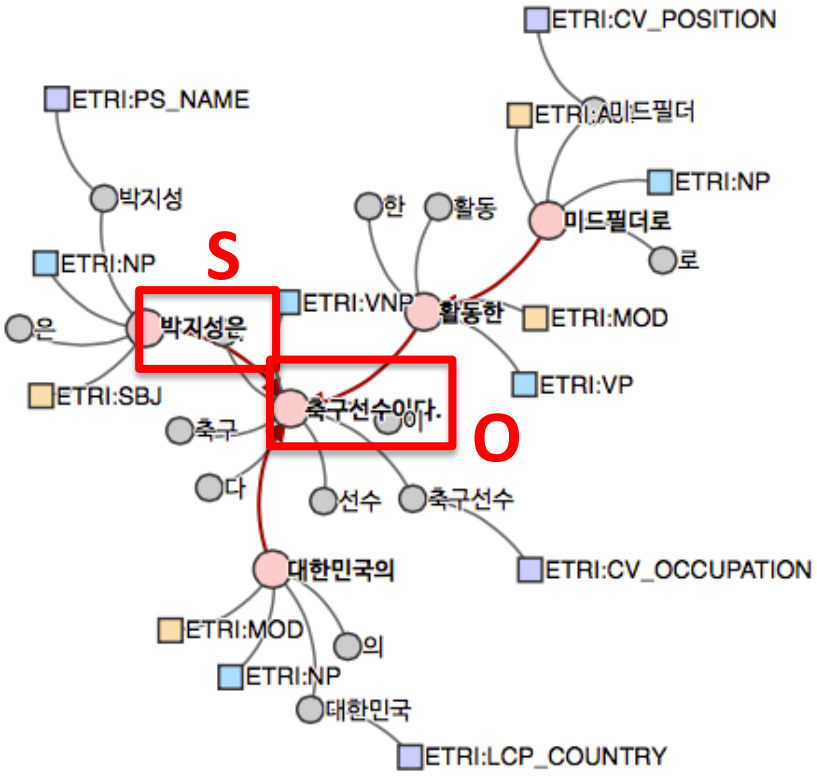
Triplization - example

- 4) S / P / O
  - 박지성의(S) 포지션은(P) 미드필더이다(O).



Triplization - example

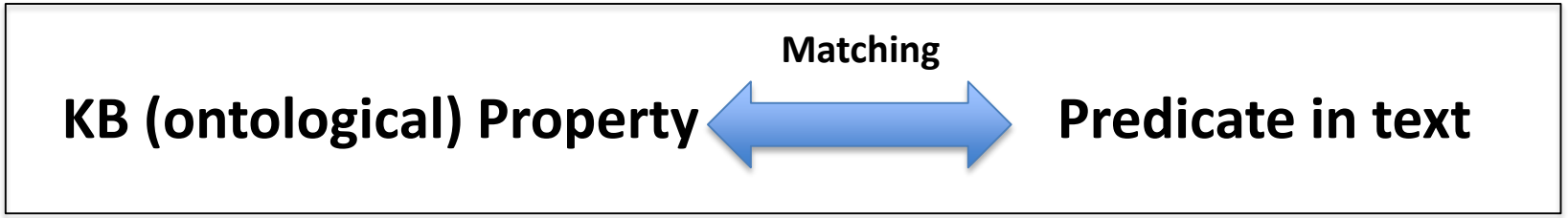
- 5) ISA (is a)
  - 박지성은(S) 미드필더로 활동한 대한민국의 축구선수이다(O).



Learning by Reading: Predicate Linking

Predicate Linking

- **~= Ontology Lexicalization (for Knowledge Base Property)**



- **Prerequisite**
  - **Distant Supervision Assumption**
    - Any sentence that contains a pair of entities that participate in a known DBpedia property is likely to express that relation in some way.
      - » (Distant Supervision for Relation Extraction, ACL 2009, Mike Mintz)
  - **Existence of Wikipedia sentecne corresponding to DBpedia RDF Semantically**
    - **Sentences including both S and O of RDFs has high possiblitiy to get the related predicate with P.**
    - **Predicate mapping to Property**

Input Data

- Korean Wikipedia
  - Dump as of 11 Sep 2014
  - About 0.29 million documents
  - About 2.8 million sentences
- Korean DBpedia
  - Dump 2014
  - 2.2 milion triples using prop-ko property
    - prop-ko property
      - Korean local property
      - About 10000 properties

Approach

- (1) Collect S, O pair for a seed property

Seed property

frequently used 1300 properties
Covering 95% of DBPedia Triples

	DBpedia Property	# of triple
1	prop-ko:이름	95,365
2	prop-ko:제목	47,294
3	prop-ko:그림	37,882
4	prop-ko:출생지	30,851
5	prop-ko:장르	23,511
6	prop-ko:직업	23,084
7	prop-ko:클럽연도	19,644
8	prop-ko:출장수(골)	18,868
9	prop-ko:클럽	18,391
10	prop-ko:웹사이트	17,107
...	...	...

For a certain property (prop-ko:출생지)

DBpedia-owl:birthPlace

Related Triple

트리플 1 S_1: <스티븐_스필버그>, P_1: <prop-ko:출생지>, O_1: <오하이오_주>

트리플 2 S_2: <유재석>, P_2: <prop-ko:출생지>, O_2: <서울_특별시>

트리플 N S_N: ..., P_N: ..., O_N: ...

About 30,000 triples exist (prop-ko:출생지)

Collect all possible S, O Labels

(include Redirection pages)

S	O
스티븐 스필버그	오하이오
스티븐 스필버그	오하이오 주
스티븐 스필버그	오하이오주
스티븐 스필버그	오하오
스필버그	오하이오
스필버그	오하이오 주
스필버그	오하이오주
...	...
유재석	서울 특별시
유재석	서울시
유재석	서울
...	...

Approach

- (2) Search all sentences having both S, O label

All possible S, O Labels collected
(include Redirection pages)

S	O
스티븐 스필버그	오하이오
스티븐 스필버그	오하이오 주
스티븐 스필버그	오하이오주
스티븐 스필버그	오히오
스필버그	오하이오
스필버그	오하이오 주
스필버그	오하이오주
…	…
유재석	서울 특별시
유재석	서울시
유재석	서울
…	…

Search all sentences containing both S, O label among 2.8 million Wikipedia sentences

$Sent_1$ 스티븐 앨런 **스필버그**는 유대인 부모 아놀드 스필버그와 레아 아들러의 아들로 **오하이오**, 신시네티에서 태어났다.

$Sent_2$ 스티븐 **스필버그**는 **오하이오**에서 어린 시절을 보냈다.

…

$Sent_1$ 유재석은 1972년 8월 14일, 서울에서 태어났다.

…

About 10,000 sentences retrieved containing both a certain S and O pair
(Filtering process required for the retrived sentence)

Conclusion: Use in Case

- 서울 지하철 1호선은 시청역을 운행한다.
- (Line #1 of Seoul Metro runs to City Hall station.)
  - **http://ko.dbpedia.org/property/노선**

- 금강은 군산에 흐르는 강이다.
- (Geum River flows through Gunsan.)
  - **<http://ko.dbpedia.org/property/도시>**
  - **(<http://dbpedia.org/ontology/city>)**

- 라디오헤드는 1985년에 영국 옥스퍼드셔 주 애빙던에서 결성된 얼터너티브 록 밴드이다.
- (Radiohead are an English rock band from Abingdon, Oxfordshire, formed in 1985.)
  - **http://ko.dbpedia.org/property/활동시기**
  - **(<http://dbpedia.org/ontology/activeYearsStartYear>)**

Conclusion: Use in Case

- 잡스는 1972년 리드 대학교에 입학하여 철학 공부를 시작하였다.
- (Jobs briefly attended Reed College in 1972)
  - **http://ko.dbpedia.org/property/출신학교**

- 유재석은 1972년 8월 14일, 서울에서 태어났다.
- (Yoo Jae-suk was born on August 14, 1972 in Seoul, South Korea.)
  - **<http://ko.dbpedia.org/property/출생일>**
  - **(<http://dbpedia.org/ontology/birthDate>)**

2. TYPE 2:

QUESTIONS WHOSE ANSWERS “CAN” BE FOUND IN KOREAN DBPEDIA AND WIKIPEDIA IN THE FORM OF “A WORD” OR “A SERIES OF WORDS”

Type 2 : **24 out of 384 (6.25%)**

= Questions whose answers "can" be found in Korean DBpedia and Wikipedia in the form of "a word" or "a series of words".

e.g.,

Ko.NLQ#168 :

대한민국 기획재정부의 산하 기관들은 무엇인가?

What are the affiliated organizations of the Ministry of Strategy and Finance of South Korea?

Answer in Ko.DBpedia & Wikipedia :

기획재정부
Ministry of Strategy and Finance of South Korea
企劃財政部

기획재정부 청사

기획재정부의 상징

설립일	• 1948년 7월 17일 재무부로 설립[1] • 2008년 2월 29일 기획재정부로 개편
전신	• 재정경제부 • 기획예산처
소재지	세종특별자치시 갈매로 477(어진동)
직원 수	977명(본부 950명 + 소속기관 27명)
기관장	최경환[2]
산하 기관 Affiliated Organizations	• 국세청 National Tax Service • 관세청 Korea Customs Service • 조달청 Public Procurement Service • 통계청 Korea National Statistical Office
웹사이트	http://www.mosf.go.kr/

국세청
National Tax Service
國稅廳

국세청 휘장

국세청 본청

설립일	1966년 3월 3일
전신	(없음)
소재지	세종특별자치시 노을6로 8-14(나성동 457)
직원 수	19,981명 (본부 817명 + 소속기관 19,164명)
기관장	임환수[1]
상급 기관	기획재정부 Higher Authority : Ministry of Strategy and Finance
웹사이트	http://www.nts.go.kr/

관세청
Korea Customs Service
關稅廳

관세청의 상징그림

설립일	• 1948년 11월 4일 재무부 세관국으로 설립 • 1970년 8월 27일 관세청을 설치
전신	재무부 세관국
소재지	대전광역시 서구 청사로 189 1동
직원 수	4,588명 (본부 331명 + 4,257명)
기관장	김낙회[1]
상급 기관	기획재정부 Higher Authority : Ministry of Strategy and Finance
웹사이트	http://www.customs.go.kr/

조달청
Public Procurement Service
調達廳

조달청의 상징그림

설립일	• 1949년 1월 17일 임시외자총국을 설치 • 1961년 10월 2일 조달청을 설치
전신	외자청
소재지	대전광역시 서구 청사로 189 정부대전청사 3동
직원 수	968명 (본부 442명 + 소속기관 526명)
기관장	김상규[1]
상급 기관	기획재정부 Higher Authority : Ministry of Strategy and Finance
웹사이트	http://www.pps.go.kr/

통계청
Korea National Statistical Office
統計廳

통계청의 휘장

설립일	• 1948년 11월 6일 공보처 통계국으로 설립 • 1990년 12월 27일 통계청을 설치
전신	경제기획원 조사통계국
소재지	대전광역시 서구 청사로 189 정부대전청사 3동
직원 수	633명 (본부 603명 + 소속기관 30명)
기관장	박형수[1]
상급 기관	기획재정부 Higher Authority : Ministry of Strategy and Finance
웹사이트	http://www.kostat.go.kr/

e.g.,

Ko.NLQ#200 :

대한수중핀수영협회가 한국스쿠버다이빙협회로 개칭한, 세계수중연맹에 가입한, 대한체육회에 가입한, 사단법인 대한수중핀수영협회로 개칭한 년도는 각각 언제인가?

When was the year that the Korea Underwater Association changed its name to "Korea Scuba Diving Association“, joined the Confédération Mondiale des Activités Subaquatiques (CMAS), joined the Korean Olympic Committee, and changed its name to "Korea Underwater Association“ respectively?

Answer in Ko.DBpedia & Wikipedia :

대한수중핀수영협회 Korea Underwater Association

위키백과, 우리 모두의 백과사전.

대한수중핀수영협회(Korea Underwater Association, **KUA**)는 수중 스포츠 보급을 통한 국민체력 향상과 우수 선수 양성으로 국민체육 발전에 이바지할 목적으로 1997년 4월 7일 설립된 대한민국 문화체육관광부 소관의 사단법인이다. 사무실은 서울특별시 송파구 오륜동 88-2 올림픽공원 펜싱경기장 내에 있다.

목차 [숨기기]

연혁 [편집] History

- 1968년 3월 6일 : 한국스킨스쿠버다이빙클럽 발족
- 1969년 7월 20일 : 제1회 전국수중경기대회 개최 및 매년 핀수영, 수중 방향찾기, 수중채집, 수중수렵 경기 주최
- 1973년 3월 2일 : 한국스쿠버다이빙협회 개칭 2 March 1973 : changed its name to "Korea Scuba Diving Association"
- 1981년 11월 20일 : 세계수중연맹(CMAS) 가입을 위해 한국수중협회로 개칭
- 1982년 6월 5일 : 세계수중연맹(CMAS) 가입 5 June 1982 : joined the CMAS
- 1988년 12월 27일 : 아시아수중연맹(AUF) 가입 및 대한체육회 준회원 가입 및 대한수중협회 개칭
- 1992년 2월 24일 : 대한체육회 정식 가입 24 February 1992 : joined the Korean Olympic Committee
- 1997년 4월 7일 : 사단법인 대한수중협회 설립 허가
- 2010년 5월 3일 : 사단법인 대한수중핀수영협회로 명칭 변경 3 May 2010 : changed its name to "Korea Underwater Association"

e.g.,

Ko.NLQ#218 :

소설 '해리 포터'의 부제목들은 무엇인가?

What are the subtitles of the novel 'Harry Potter'?

Answer in Ko.DBpedia & Wikipedia :

해리 포터

위키백과, 우리 모두의 백과사전.
(해리포터에서 넘어옴)

이 문서는 해리포터 책 시리즈에 관한 것입니다. 등장인물에 대해서는 해리 포터 (등장인물) 문서를 참조하십시오.
영화 시리즈에 대해서는 해리 포터 (영화 시리즈) 문서를 참조하십시오.

《해리 포터》(Harry Potter)는 조앤 K. 롤링이 쓴 판타지 소설이다. 이 책의 제목은 주인공인 해리 포터의 이름이다. 《해리 포터》 시리즈의 첫째 작품인 《해리 포터와 마법사의 돌》은 1997년에 출판되었다. 이 작품이 큰 성공을 거두면서 전 세계적으로 알려지게 되어, 영화와 비디오 게임 및 다양한 상품들이 제작되었다. 2007년에 《해리 포터와 죽음의 성물》을 마지막으로 모두 완결되었다.

목차 [숨기기]

해리포터 신드롬 [편집]

해리포터의 작가 조앤 K. 롤링은 해리포터를 쓰기 전에는 국가에서 지원을 받는 저소득층 사람으로 기초유급자 생활을 하고 있었다. 하지만 롤링이 집 근처 카페에서해리포터와 마법사의 돌을 완성하고, 출판사에 책을 출판한 후 이 책은 지금까지 4억 부 이상이 팔렸으며 결국 판타지 소설에서 **해리 포터**는 반지의 제왕, 나니아 연대기와 동등한 위치에 서있게 되었다.

특히 이 책이 나오는 날이면, 독자들이 추운 날씨나 더운 날씨 마다하지 않고 서점이 열 때까지 기다려서 책을 사고, 해리포터 책이 판매 시작되는 날이면 항상 교통이 마비되고, 서점은 늦게까지 문을 연다. 해리 포터 소설은 보통 어린이를 대상으로 한 작품으로 여겨지지만, 거의 모든 연령대에 걸쳐 독자 층을 형성하고 있으며, 그 근거로 어른들도 공공 장소에서 거리낌 없이 읽을 수 있도록 좀 더 '점잖은' 표지로 된 판이 따로 출간된 바 있다.

소설을 바탕으로 워너 브라더스에서 영화로 제작하여 2001년 첫 영화가 개봉되었으며, 2011년까지 여덟 편의 영화가 개봉되었다. 작가 롤링에 따르면 그녀는 맨체스터에서 런던으로 가는 기차 안에서 처음 이 이야기를 떠올려, 에딘버러에 있는 작은 카페에서 어린 딸을 어르며 연작의 첫 권을 썼다. 그 뒤 책의 인세 및 영화나 관련 상품의 로열티를 통해 롤링은 2004년 《포브스》가 발표한 세계의 부자 순위 552위에 올랐으며,[1] 영국 여왕보다도 더 큰 부자가 되었다.[2]

참고로 조앤 K. 롤링은 2007년 포보스지가 조사한 결과에서 세계 부자 순위 663위에 올라 있다.

소설 시리즈 [편집]

해리 포터 (소설 시리즈) [편집]

출간 일자	제목	원제
1997년 6월 26일	《해리 포터와 마법사의 돌》	Harry potter and the Philosopher's Stone
1998년 7월 2일	《해리 포터와 비밀의 방》	Harry Potter and the Chamber of Secrets
1999년 7월 8일	《해리 포터와 아즈카반의 죄수》	Harry Potter and the Prisoner of Azkaban
2000년 7월 8일	《해리 포터와 불의 잔》	Harry Potter and the Goblet of Fire
2003년 6월 1일	《해리 포터와 불사조 기사단》	Harry Potter and the Order of Phoenix
2005년 7월 16일	《해리 포터와 혼혈 왕자》	Harry Potter and the Half-Blood Prince
2007년 7월 21일	《해리 포터와 죽음의 성물》	Harry Potter and the Deathly Hallows

해리 포터

Harry Potter

미국판 소설에서 처음 사용된 《해리 포터》 로고

해리 포터와 마법사의 돌
해리 포터와 비밀의 방
해리 포터와 아즈카반의 죄수
해리 포터와 불의 잔
해리 포터와 불사조 기사단
해리 포터와 혼혈 왕자
해리 포터와 죽음의 성물

작가	조앤 K. 롤링
원제	Harry Potter
국가	영국
언어	영어
장르	판타지 소설
출판사	블룸스베리 퍼블리싱 (Bloomsberry pubilicing)
출판일	1997-6-28
한국어 출판일	2005년 11월 2일
미디어 유형	도서

호그와트 심볼

C2K
for TYPE 2

- Category to Knowledge

- Jiseong Kim, Eun-Kyung Kim, Yousung Won, Sangha Nam and Key-Sun Choi
- The Association Rule Mining System for Acquiring Knowledge of DBpedia from Wikipedia Categories
- NLP&DBpedia 2015, ISWC2015

Wikipedia: The Rich Source of Knowledge

- Wikipedia describes such information with various ways:

Articles:

Alan Turing

From Wikipedia, the free encyclopedia

"Turing" redirects here. For other uses, see Turing (disambiguation).

Alan Mathison Turing, OBE, FRS (/ˈtjʊərɪŋ/; 23 June 1912 – 7 June 1954) was a British pioneering computer scientist, mathematician, logician, cryptanalyst, philosopher, mathematical biologist, and marathon and ultra distance runner. He was highly influential in the development of computer science, providing a formalisation of the concepts of algorithm and computation with the Turing machine, which can be considered a model of a general purpose computer.[2][3][4] Turing is widely considered to be the father of theoretical computer science and artificial intelligence.[5]

During the Second World War, Turing worked for the Government Code and Cypher School (GC&CS) at Bletchley Park, Britain's codebreaking centre. For a time he led Hut 8, the section responsible for German naval cryptanalysis. He devised a number of techniques for breaking German ciphers, including improvements to the pre-war Polish bombe method, an electromechanical machine that could find settings for the Enigma machine. Turing played a pivotal role in cracking intercepted coded messages that enabled the Allies to defeat the Nazis in many crucial engagements, including the Battle of the Atlantic; it has been estimated that this work shortened the war in Europe by as many as two to four years.[6]

→ Related researches:
Learning by Reading

Infoboxes:

Alan Turing

Turing aged 16

Born	Alan Mathison Turing 23 June 1912 Maida Vale, London, England
Died	7 June 1954 (aged 41) Wilmslow, Cheshire, England
Residence	Wilmslow, Cheshire, England
Nationality	British

→ Related researches:
DBpedia, Yago

Categories:

⇐ Our target

Categories: Alan Turing | 1912 births | 1954 deaths | 20th-century mathematicians | 20th-century philosophers | Academics of the University of Manchester Institute of Science and Technology | Alumni of King's College, Cambridge | Artificial intelligence researchers | Atheist philosophers | Bayesian statisticians | British cryptographers | British long-distance runners | British male athletes | British people of World War II | Computability theorists | Computer designers | Deaths in Cheshire | English atheists

→ Related researches: Yago, Catriple

Jiseong Kim, Eun-Kyung Kim, Yousung Won, Sangha Nam and Key-Sun Choi,
The Association Rule Mining System for Acquiring Knowledge of DBpedia from Wikipedia Categories, NLP&DBpedia 2015, ISWC

The Goal of C2K

- To extract DBpedia triples from Wikipedia categories
- (S, Wikipedia category) → ⟨S, P, O⟩

e.g., (Albert Einstein, Category: Swiss emigrants to the United States)
→ <Albert Einstein, *nationality*, the United States>
→ <Albert Einstein, *naturalization*, the United States>
→ …

Jiseong Kim, Eun-Kyung Kim, Yousung Won, Sangha Nam and Key-Sun Choi,
The Association Rule Mining System for Acquiring Knowledge of DBpedia from Wikipedia Categories, NLP&DBpedia 2015, ISWC

Characteristics of Wikipedia Categories

- Categories have plentiful lexical patterns of each relation
-

  e.g., (Albert Einstein, Category: Swiss emigrants to the United States)
  → <Albert Einstein, *nationality*, the United States>
  → <Albert Einstein, *naturalization*, the United States>
  → …

- They are organized in the category network and have siblings with similar lexical patterns.

  e.g., Swiss emigrants to the United States
  Swiss emigrants to Brazil
  Swiss emigrants to Canada
  …

Jiseong Kim, Eun-Kyung Kim, Yousung Won, Sangha Nam and Key-Sun Choi,
The Association Rule Mining System for Acquiring Knowledge of DBpedia from Wikipedia Categories, NLP&DBpedia 2015, ISWC

Our Approach: Step 1

- Lexical pattern learning:

- An entity: Paul de Castella

- Categories: Swiss emigrants to Australia, …

- DBpedia infobox triples: <Paul de Castella, *nationality*, Australia>, …

- Lexical patterns of the relation “nationality”: Swiss emigrants to x, …

exact matching on a basis of a word

Jiseong Kim, Eun-Kyung Kim, Yousung Won, Sangha Nam and Key-Sun Choi,
The Association Rule Mining System for Acquiring Knowledge of DBpedia from Wikipedia Categories, NLP&DBpedia 2015, ISWC

Our Approach: Step 2

- Propagating learned lexical patterns:

The learned lexical pattern: Swiss emigrants to x, …

Siblings of the category "Swiss emigrants to Australia":

- Swiss emigrants to the United States
- Swiss emigrants to Brazil
- Swiss emigrants to Canada
- …

Propagating the learned pattern and mining association rules:

- <e, *belongsTo*, Swiss emigrants to the United States>
  → <e, *nationality*, the United States>
- <e, *belongsTo*, Swiss emigrants to Brazil>
  → <e, *nationality*, Brazil >
- …

Jiseong Kim, Eun-Kyung Kim, Yousung Won, Sangha Nam and Key-Sun Choi,
The Association Rule Mining System for Acquiring Knowledge of DBpedia from Wikipedia Categories, NLP&DBpedia 2015, ISWC

Normalization for Korean Words

- Our approach run on a basis of a word for a few reasons.
  - To lower the overall error rates
  - To enhance time efficiency of the learning procedure
- Korean words have plentiful postfixes like
  - "ui"(genitive), "eseo"(locative), "nyeon"(unit of year).
- To enhance the quality of results, we should separate the postfixes from each word.

year born-being

e.g., Category:1989년_태어남 (1989_births)

→ The normalized one: Category:1989_년_태어남

separated

Jiseong Kim, Eun-Kyung Kim, Yousung Won, Sangha Nam and Key-Sun Choi,
The Association Rule Mining System for Acquiring Knowledge of DBpedia from Wikipedia Categories, NLP&DBpedia 2015, ISWC

Categories vs. Infobox Triples

- Liu *et al.* (2008) said that
  - The article-coverage of infobox triples = 44.2%
  - The article-coverage of categories = 80.6%

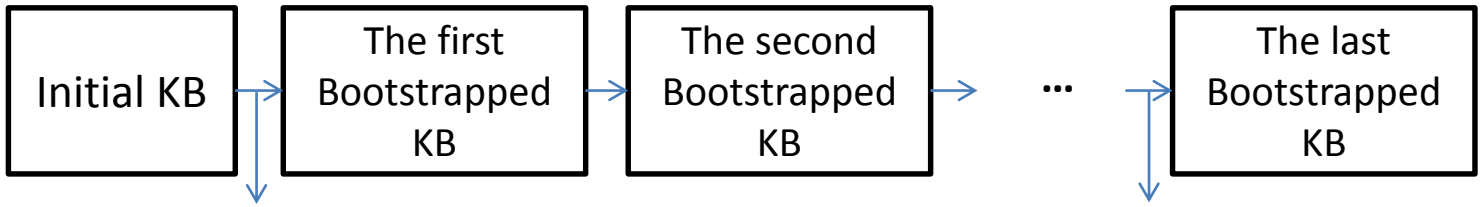
- There would be many entities
  - categorized in Wikipedia categories
  - without DBpedia infobox triples

- The fact encouraged us to bootstrap our results iteratively!

Jiseong Kim, Eun-Kyung Kim, Yousung Won, Sangha Nam and Key-Sun Choi,
The Association Rule Mining System for Acquiring Knowledge of DBpedia from Wikipedia Categories, NLP&DBpedia 2015, ISWC

Our Approach: Bootstrapping

- We bootstrap mined rules by the next flow:



+ Predictions
of learned rules
with high confidence

If the ratio of increase in
learned rules < threshold,
then we stop bootstrapping

Output: the last bootstrapped rules

Jiseong Kim, Eun-Kyung Kim, Yousung Won, Sangha Nam and Key-Sun Choi,
The Association Rule Mining System for Acquiring Knowledge of DBpedia from Wikipedia Categories, NLP&DBpedia 2015, ISWC

Experiments and Analysis

- Korean DBpedia 2014:
  - Category pair # = 728,433
  - DBpedia seed triple # = 2,527,466

- Results:
  - Mined rule # = 71,035
  - Predicted triple # = 1,530,590
  - Iteration # = 3

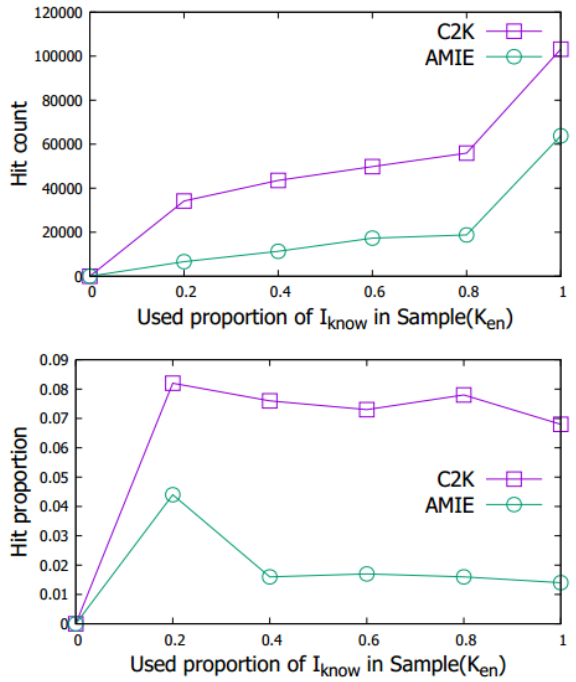
Seg. #	Precision	Min conf.	Max conf.
1	0.75	553.0	3352.0
2	0.7	105.0	536.0
3	0.7	12.0	93.0
4	0.65	2.0	9.0
5	0.6	1.0	2.0
6	0.2	0.0	1.0
7	0.3	0.0	0.0
8	0.15	0.0	0.0
9	0.1	0.0	0.0
10	0.25	0.0	0.0

- 0.72M category pairs → 1.5M triples predicted

Jiseong Kim, Eun-Kyung Kim, Yousung Won, Sangha Nam and Key-Sun Choi,
The Association Rule Mining System for Acquiring Knowledge of DBpedia from Wikipedia Categories, NLP&DBpedia 2015, ISWC

Experiments on English Datasets

- Sampled English DBpedia 2.0
  - 1,027,213 category pairs → 1,530,253 triples predicted



<Our approach vs. AMIE>

Prop. I_{know}	C2K	AMIE
1.0	$103,177_{/1,528,253}$	$63,819_{/4,708,163}$
0.8	$55,945_{/721,611}$	$18,749_{/1,164,978}$
0.6	$49,813_{/678,008}$	$17,308_{/1,047,907}$
0.4	$43,526_{/574,491}$	$11,317_{/714,909}$
0.2	$34,185_{/415,676}$	$6,603_{/150,529}$

Seg. #	Precision	Min conf.	Max conf.
1	0.85	227.0	41185.0
2	0.7	56.0	222.0
3	0.45	10.0	38.0
4	0.55	1.0	10.0
5	0.55	1.0	1.0
6	0.6	1.0	1.0
7	0.8	1.0	1.0
8	0.45	0.0	1.0
9	0.25	0.0	0.0
10	0.35	0.0	0.0

Jiseong Kim, Eun-Kyung Kim, Yousung Won, Sangha Nam and Key-Sun Choi,
The Association Rule Mining System for Acquiring Knowledge of DBpedia from Wikipedia Categories, NLP&DBpedia 2015, ISWC

Conclusion and Future Work

- We successfully extracted new triples from the Wikipedia categories using the DBpedia triples.

- Our approach and the results can be used
  - in enriching DBpedia automatically.
  - to enhance the performance of QA systems.

Jiseong Kim, Eun-Kyung Kim, Yousung Won, Sangha Nam and Key-Sun Choi,
The Association Rule Mining System for Acquiring Knowledge of DBpedia from Wikipedia Categories, NLP&DBpedia 2015, ISWC

TYPE 3

Subtype 3-1 : **14 out of 384 (3.64%)**

= Questions whose answers "cannot" be found in Korean DBpedia and Wikipedia "partially", because "some portions of information are lacked in the KBs". e.g.,

Ko.NLQ#114 :

남이장군, 겨울연가, 연가지가와 관련 있는 섬은 무엇인가?

What is the island related to General Nami, Winter Sonata and Yeongajiga?

Answer in Ko.DBpedia & Wikipedia :

Namisum is a tiny half-moon shaped island located in Chuncheon, South Korea, …. Its name originated from General Nami, …. The island, especially the Metasequoia path, was one of the main filming locations of the Korean Broadcasting System 2002 television drama series Winter Sonata, starring Bae Yong-joon and Choi Ji-woo.

Korean DBpedia and Wikipedia are not inclusive of any information about "Yeongajiga".

4. OTHER EFFORTS FOR BETTER KB

Triple-based Similarity Propagation for Linked Data Matching

Problem Scope

- Property Alignment between Linked Data using Similarity Flooding
  - Scope in Ontology Alignment
    - matching elements of two ontologies
      - Schema matching
        - » matching concepts of two ontologies
        - » ***matching properties of two ontologies***
      - Instance matching
        - » matching instances of two ontologies

Eun-Kyung Kim, Sangha Nam, Jongsung Woo, Senjin Nam and Key-sun Choi, 2015,
Triple-based Similarity Propagation for Linked Data Matching, The 10th International Workshop on Ontology Matching ISWC

Introduction: Problem Definition 1

- Property Alignment between Datasets
  - OWL defines the concept of property equivalence (**owl**:**equivalentProperty**) as two properties having the same property extension

$$
\begin{aligned}
&\text{Align}(O_1, O_2) = \\
&\{ (e_{i1}, e_{i2}, c_i, r_i) \mid \\
&\quad e_{i1} \in O_1 \ , \\
&\quad e_{i2} \in O_2 \ , \\
&\quad c_i \in [0,1] \ , \\
&\quad r_i \in [\ \textbf{equivalence(=)}, \\
&\qquad \text{generic/broader}(\sqsupseteq)], \\
&\qquad \text{specific/narrower}(\sqsubseteq)], \\
&\qquad \text{overlap}(\sqcap)] \ \}
\end{aligned}
$$

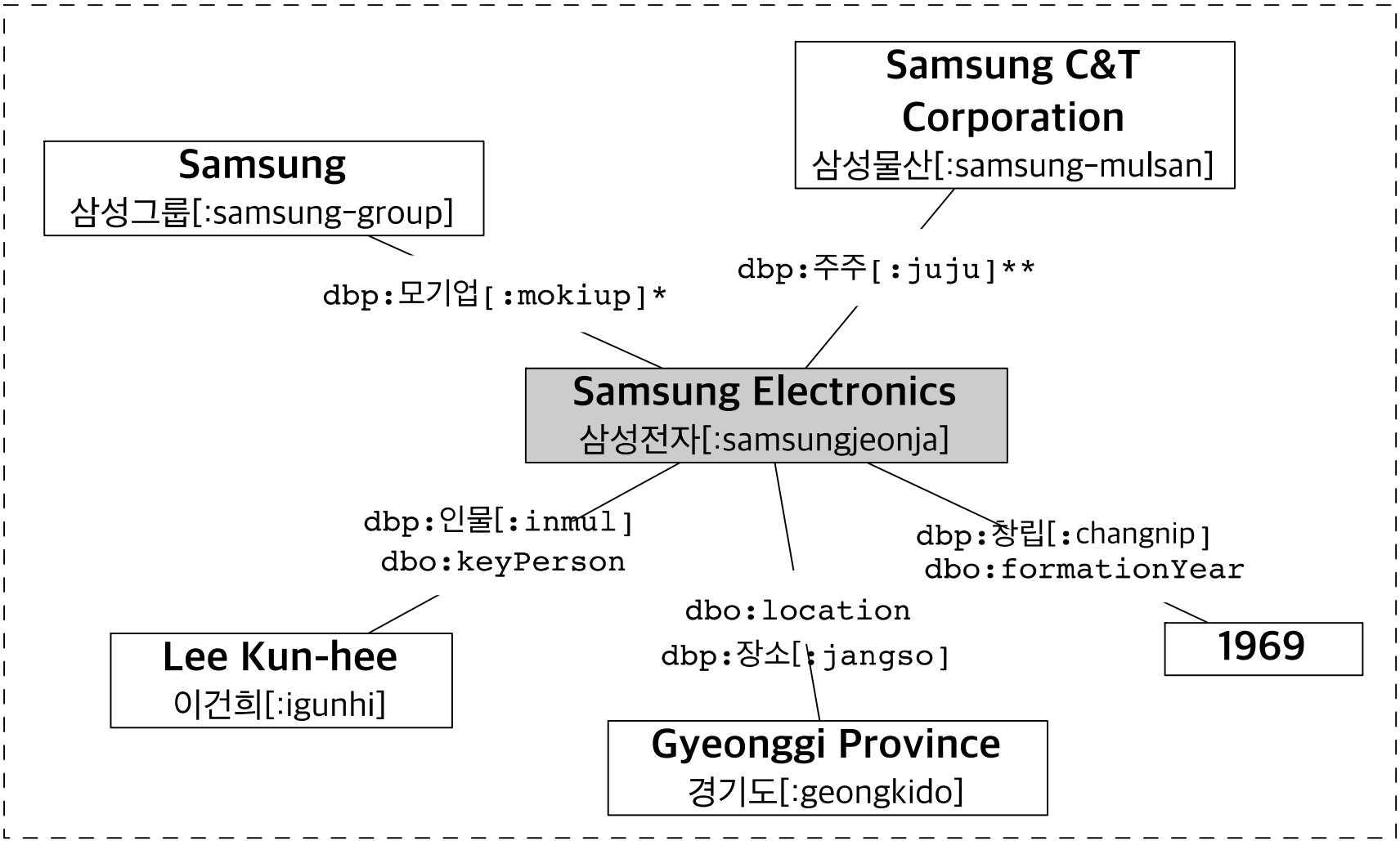
Eun-Kyung Kim, Sangha Nam, Jongsung Woo, Senjin Nam and Key-sun Choi, 2015,
Triple-based Similarity Propagation for Linked Data Matching, The 10th International Workshop on Ontology Matching ISWC

Introduction: Problem Definition 2

- The objective of property alignments
  - identify pairs (P_1, P_2), which may be in the different datasets
    - determine related properties P_1 and P_2 by finding similar triple patterns across datasets by matching subject S and object O values in triples of the form <S, P_1, O> and <S, P_2, O>
      - the given two matching subject and object pairs that are exactly same or connected by owl:sameAs link
      - the associated property names have the potential to be equivalent
    - propagate similarity of equivalent properties
      - current similarity can propagate to siblings in the graph to get more similarity results

Eun-Kyung Kim, Sangha Nam, Jongsung Woo, Senjin Nam and Key-sun Choi, 2015,
Triple-based Similarity Propagation for Linked Data Matching, The 10th International Workshop on Ontology Matching ISWC

Example of different properties



* means a *parent company* in English

** means a *stockholder* in English

Eun-Kyung Kim, Sangha Nam, Jongsung Woo, Senjin Nam and Key-sun Choi, 2015,
Triple-based Similarity Propagation for Linked Data Matching, The 10th International Workshop on Ontology Matching ISWC

Scenario Example

- TargetData
  - <구글, formationYear, 1998>
  - <금호타이어, parentCompany, 금호아시아나그룹>

- SourceData
  - <구글, 창립, 1998>
  - <속리산고속, 창립, 1968>
  - <속리산고속, 모회사, 금호아시아나그룹>

Instance-based Alignment

- TargetData
  - <구글, formationYear, 1998>
  - <금호타이어, parentCompany, 금호아시아나그룹>

- SourceData
  - <구글, 창립, 1998>
  - <속리산고속, 창립, 1968>
  - <속리산고속, 모회사, 금호아시아나그룹>

Eun-Kyung Kim, Sangha Nam, Jongsung Woo, Senjin Nam and Key-sun Choi, 2015,
Triple-based Similarity Propagation for Linked Data Matching, The 10th International Workshop on Ontology Matching ISWC

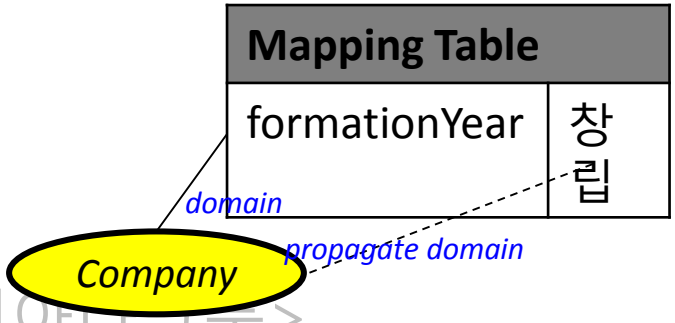
Instance-Concept Translation

- TargetData
  - <구글→*Company*, formationYear, 1998>
  - <금호타이어→*Company*, parentCompany, 금호아시아나그룹>
- SourceData
  - <구글→*Company*, 창립, 1998>
  - <속리산고속→*?*, 창립, 1968>
  - <속리산고속→*?*, 모회사, 금호아시아나그룹>

Eun-Kyung Kim, Sangha Nam, Jongsung Woo, Senjin Nam and Key-sun Choi, 2015,
Triple-based Similarity Propagation for Linked Data Matching, The 10th International Workshop on Ontology Matching ISWC

Type Propagation

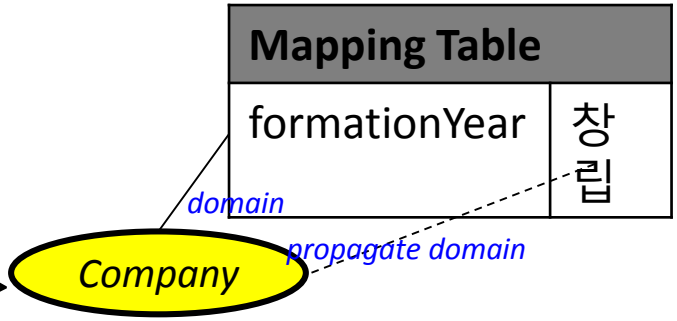
- TargetData
  - <구글, formationYear, 1998>
  - <금호타이어, parentCompany, 금호아시아나그룹>
- SourceData
  - <구글, 창립, 1998>
  - <속리산고속, 창립, 1968>
  - <속리산고속, 모회사, 금호아시아나그룹>



Eun-Kyung Kim, Sangha Nam, Jongsung Woo, Senjin Nam and Key-sun Choi, 2015,
Triple-based Similarity Propagation for Linked Data Matching, The 10th International Workshop on Ontology Matching ISWC

Instance-Concept Translation

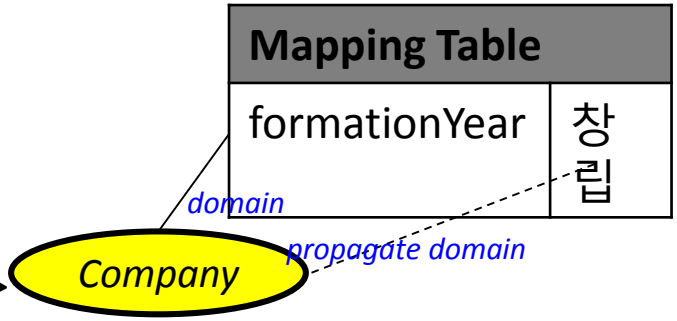
- TargetData
  - <구글→*Company*, formationYear, 1998>
  - <금호타이어→*Company*, parentCompany, 금호아시아나그룹>
- SourceData
  - <구글→*Company*, 창립, 1998>
  - <속리산고속→*?*, 창립, 1968>
  - <속리산고속→*?*, 모회사, 금호아시아나그룹>



Eun-Kyung Kim, Sangha Nam, Jongsung Woo, Senjin Nam and Key-sun Choi, 2015,
Triple-based Similarity Propagation for Linked Data Matching, The 10th International Workshop on Ontology Matching ISWC

Instance-Concept Translation

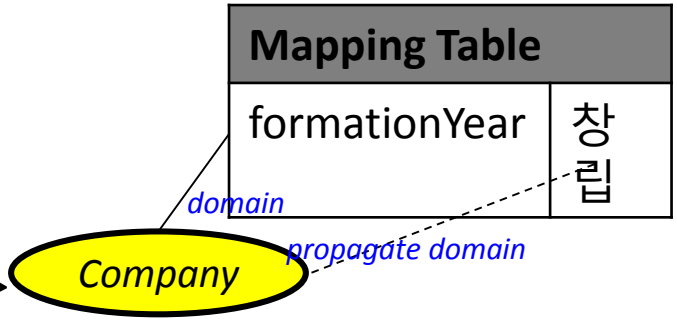
- TargetData
  - <구글→*Company*, formationYear, 1998>
  - <금호타이어→*Company*, parentCompany, 금호아시아나그룹>
- SourceData
  - <구글→*Company*, 창립, 1998>
  - <속리산고속→*Company*, 창립, 1968>
  - <속리산고속→*?*, 모회사, 금호아시아나그룹>



Eun-Kyung Kim, Sangha Nam, Jongsung Woo, Senjin Nam and Key-sun Choi, 2015,
Triple-based Similarity Propagation for Linked Data Matching, The 10th International Workshop on Ontology Matching ISWC

Instance-Concept Translation

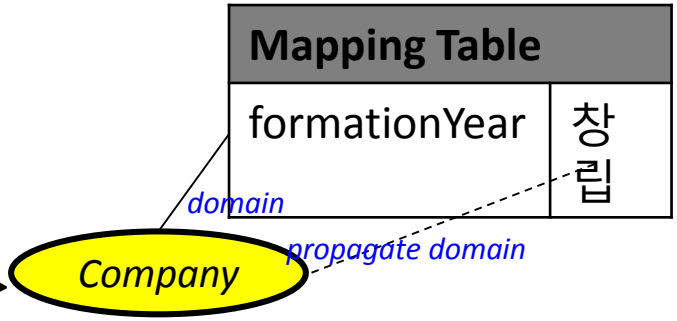
- TargetData
  - <구글→*Company*, formationYear, 1998>
  - <금호타이어→*Company*, parentCompany, 금호아시아나그룹>
- SourceData
  - <구글→*Company*, 창립, 1998>
  - <속리산고속→*Company*, 창립, 1968>
  - <속리산고속→*?*, 모회사, 금호아시아나그룹>



Eun-Kyung Kim, Sangha Nam, Jongsung Woo, Senjin Nam and Key-sun Choi, 2015,
Triple-based Similarity Propagation for Linked Data Matching, The 10th International Workshop on Ontology Matching ISWC

Instance-Concept Translation

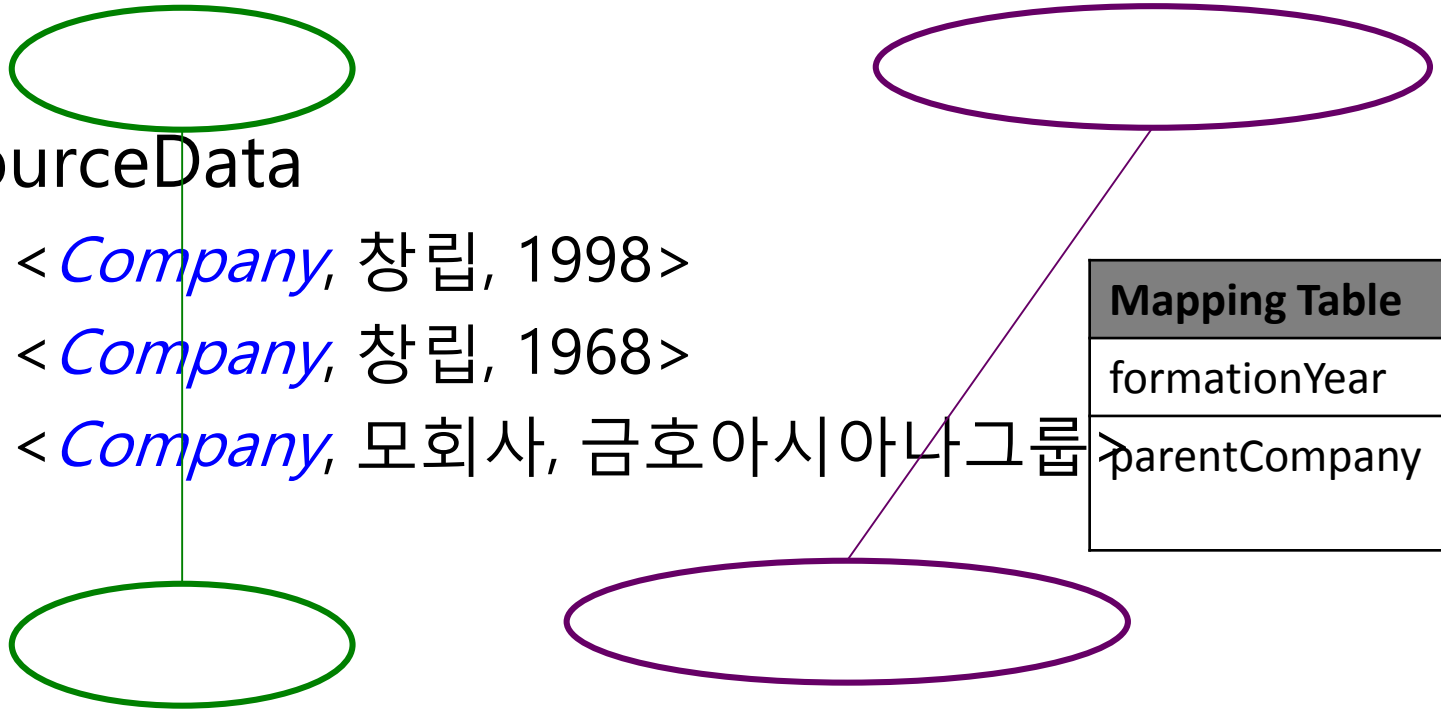
- TargetData
  - <구글→*Company*, formationYear, 1998>
  - <금호타이어→*Company*, parentCompany, 금호아시아나그룹>
- SourceData
  - <구글→*Company*, 창립, 1998>
  - <속리산고속→*Company*, 창립, 1968>
  - <속리산고속→*Company*, 모회사, 금호아시아나그룹>



Eun-Kyung Kim, Sangha Nam, Jongsung Woo, Senjin Nam and Key-sun Choi, 2015,
Triple-based Similarity Propagation for Linked Data Matching, The 10th International Workshop on Ontology Matching ISWC

Concept-based Alignment with type propagation

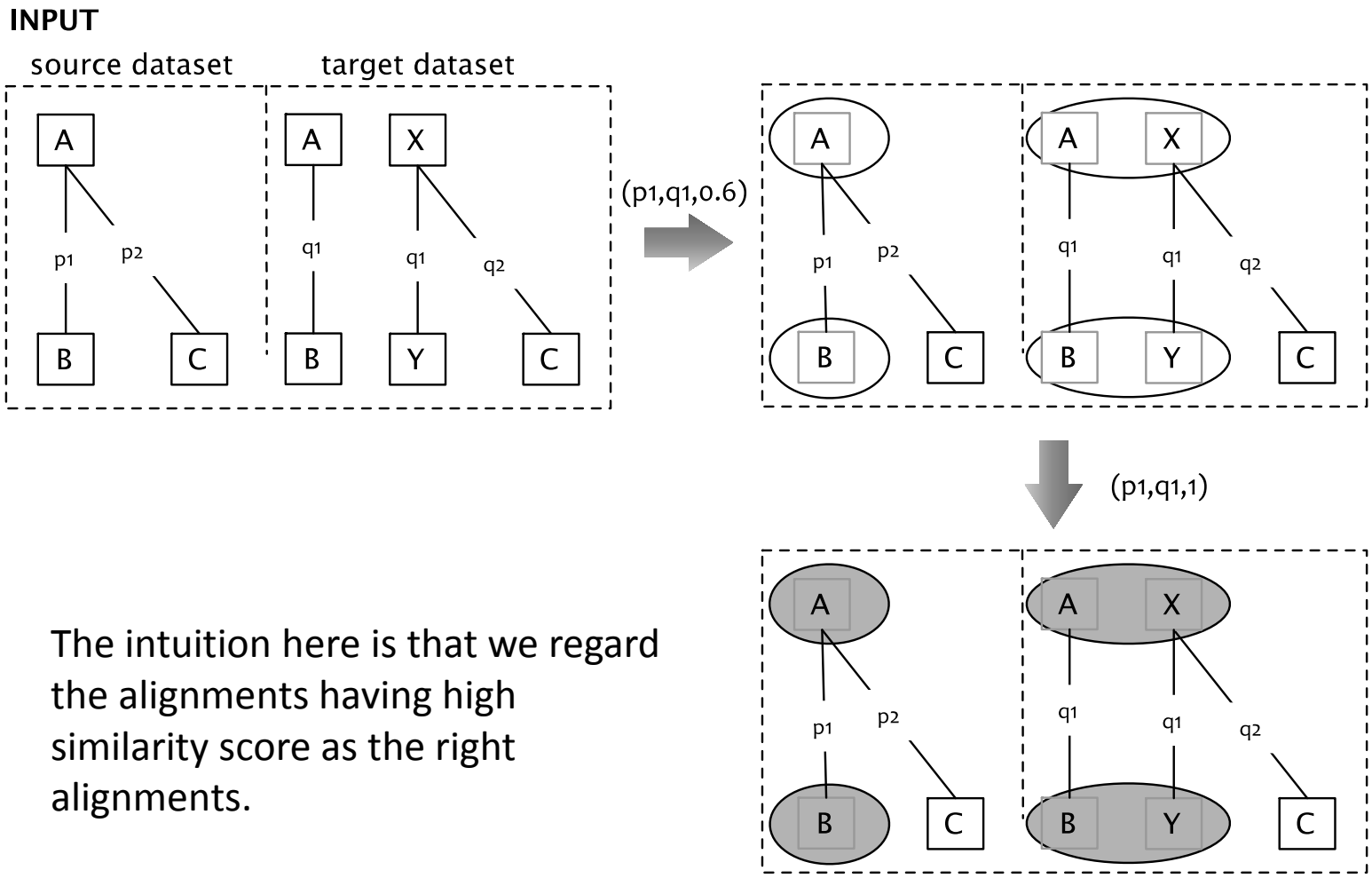
- TargetData
  - <*Company*, formationYear, 1998>
  - <*Company*, parentCompany, 금호아시아나그룹>
- SourceData
  - <*Company*, 창립, 1998>
  - <*Company*, 창립, 1968>
  - <*Company*, 모회사, 금호아시아나그룹>



Mapping Table	
formationYear	창립
parentCompany	모회사

Eun-Kyung Kim, Sangha Nam, Jongsung Woo, Senjin Nam and Key-sun Choi, 2015,
Triple-based Similarity Propagation for Linked Data Matching, The 10th International Workshop on Ontology Matching ISWC

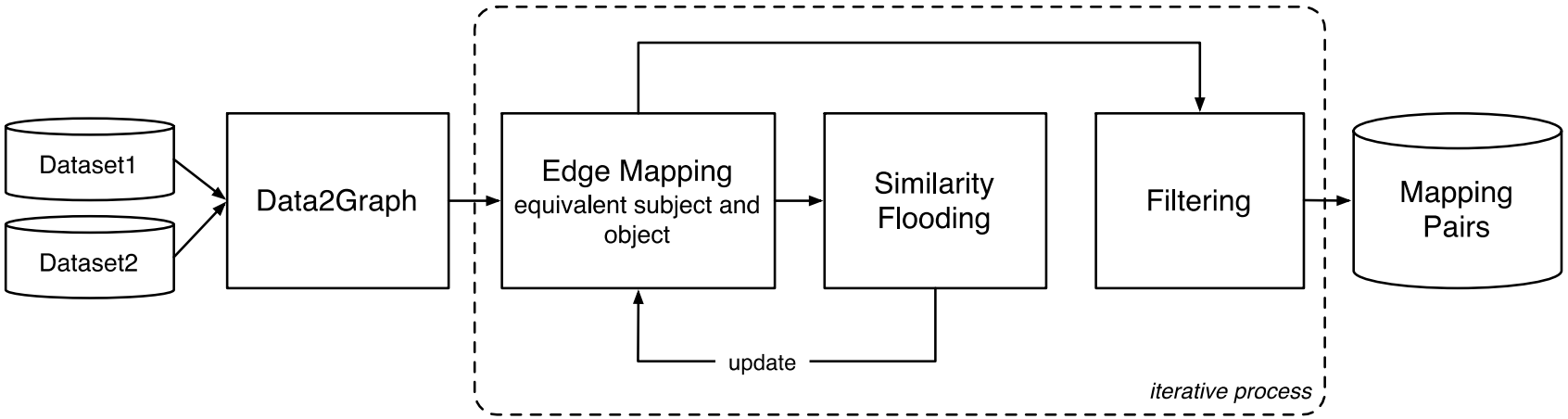
Proposed Approach: Similarity Propagation



The intuition here is that we regard the alignments having high similarity score as the right alignments.

Eun-Kyung Kim, Sangha Nam, Jongsung Woo, Senjin Nam and Key-sun Choi, 2015,
Triple-based Similarity Propagation for Linked Data Matching, The 10th International Workshop on Ontology Matching ISWC

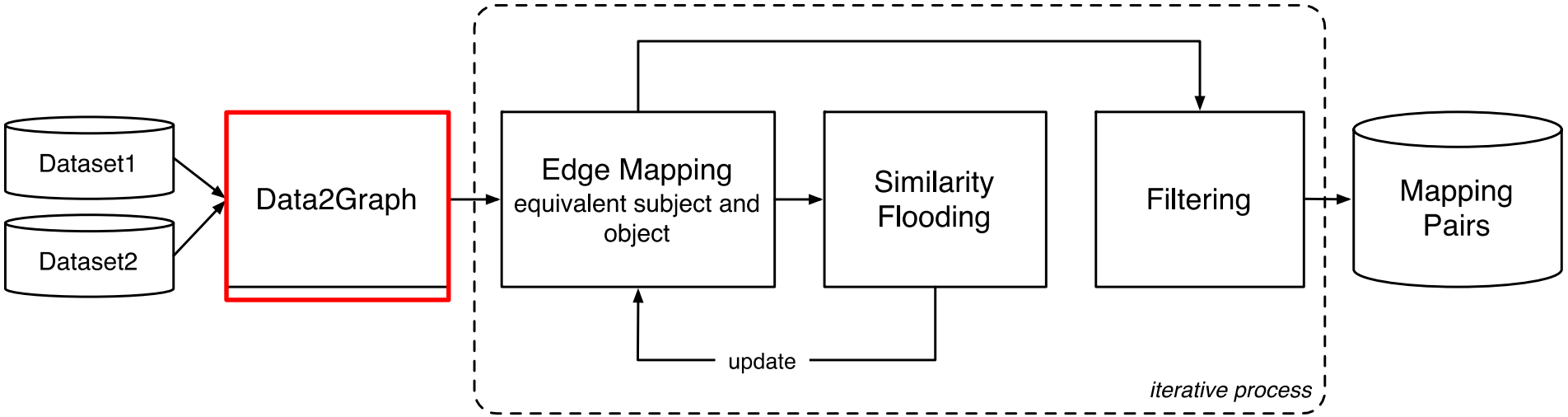
System: MAKARON



- Requirements of credible property pairs (as seeds) to propagate similarity
  - some initial correct alignment can not be changed during propagation
  - avoiding some unnecessary similarity propagation computing
    - reduce the propagation cost
    - decrease the negative affection in propagation

Eun-Kyung Kim, Sangha Nam, Jongsung Woo, Senjin Nam and Key-sun Choi, 2015,
Triple-based Similarity Propagation for Linked Data Matching, The 10th International Workshop on Ontology Matching ISWC

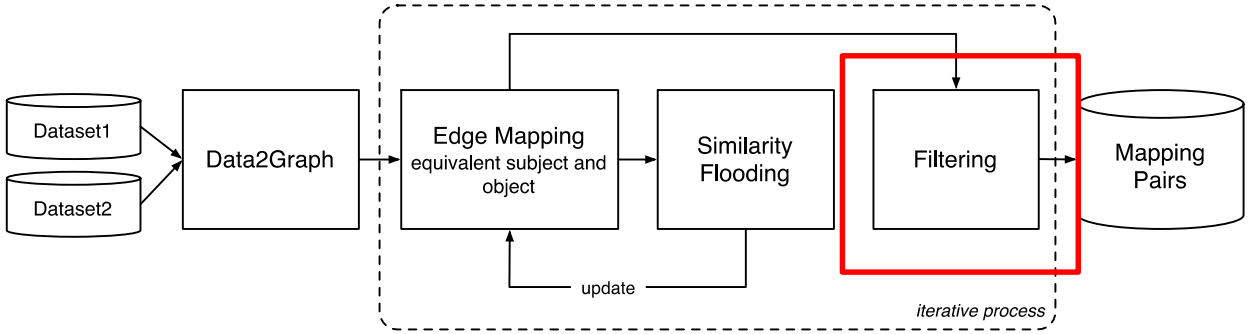
System: MAKARON



- Data2Graph: Given two sets of RDF triples as the input, MAKARON converts these to the RDF graphs G1 and G2.

Eun-Kyung Kim, Sangha Nam, Jongsung Woo, Senjin Nam and Key-sun Choi, 2015,
Triple-based Similarity Propagation for Linked Data Matching, The 10th International Workshop on Ontology Matching ISWC

System: MAKARON



- Filtering
  - Pick out the properties to be fine alignments by examining whether their similarity is higher than a predetermined threshold θ.

Evaluation Result

- Manual validation of the mapping result between DBOP (DBpedia Ontology property in English) a nd DBKP (DBpedia Korean property)
  - 3 separate blinded human annotators aligned 1,000 DBKP to DBO
    - if the meaning of two properties was similar
    - We used the majority vote to determine the correct mapping results

Eun-Kyung Kim, Sangha Nam, Jongsung Woo, Senjin Nam and Key-sun Choi, 2015,
Triple-based Similarity Propagation for Linked Data Matching, The 10th International Workshop on Ontology Matching ISWC

Conclusion

- Aligning the properties of different RDF datasets
  using a mapping similarity propagation which can be adapted
  to find new connectives
  between corresponding properties of interest

- English and the Korean DBpedia has shown that the propagated
  connectives improve the recall and F1-score measures required
  to find mapping pairs of properties
  by taking into account instance types
  in order to discover new mapping candidates

- We see this as the initial step towards enhancing multilingualism
  in Linked Open Data

Eun-Kyung Kim, Sangha Nam, Jongsung Woo, Senjin Nam and Key-sun Choi, 2015,
Triple-based Similarity Propagation for Linked Data Matching, The 10th International Workshop on Ontology Matching ISWC

5. WHAT'S NEXT?

- Our previous research
  - KB construction from text
  - Wikipedia and DBpedia
  - General Domain QA (NLQ400)

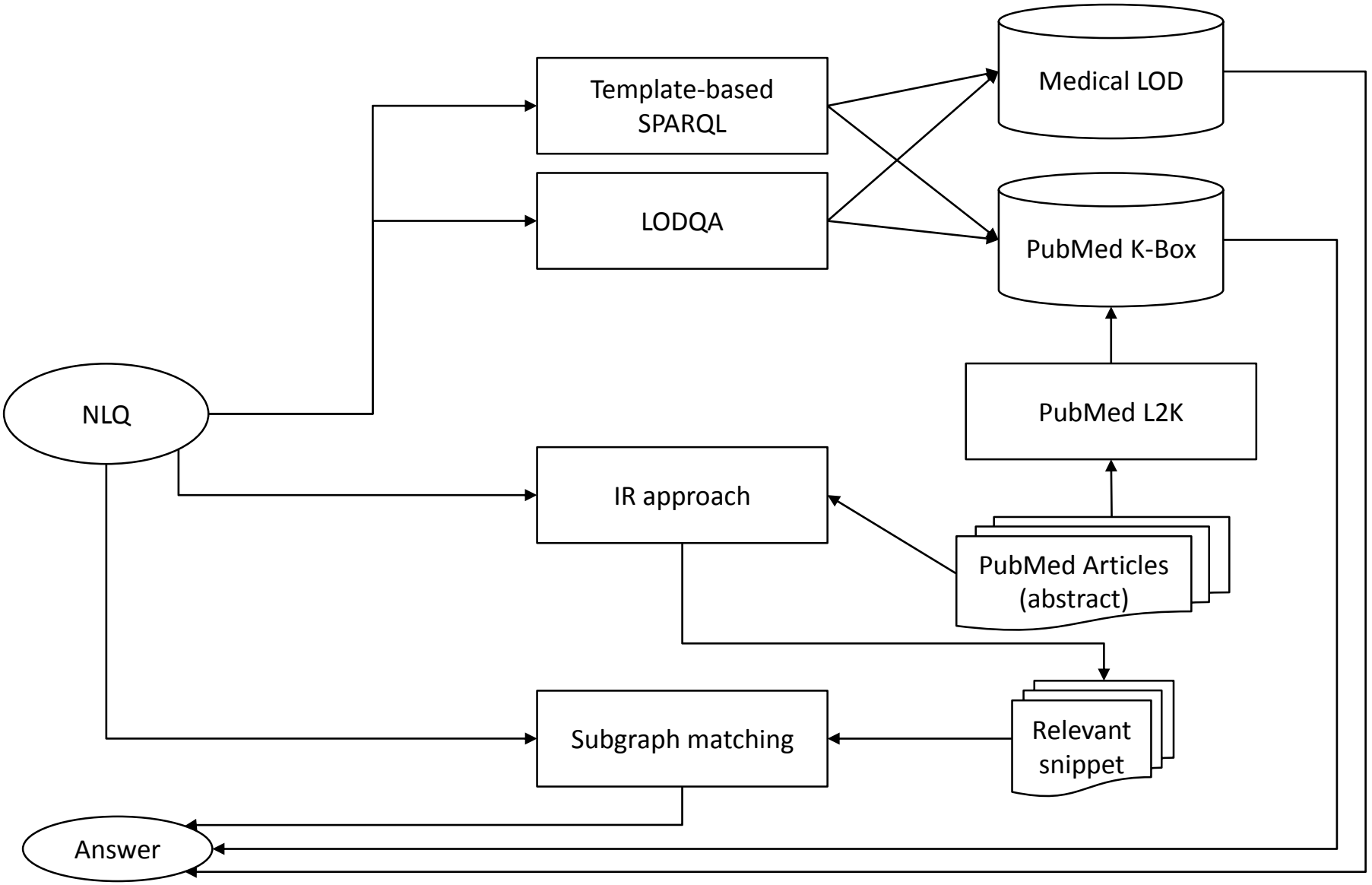
- Our next step?
  - REAL QA system (not only KB, but also Open QA)
  - Domain-specific QA (e.g. Bio medical)

OKBQA Hackathon, http://2015.okbqa.org 2015.8.24 – 28 in Jeju

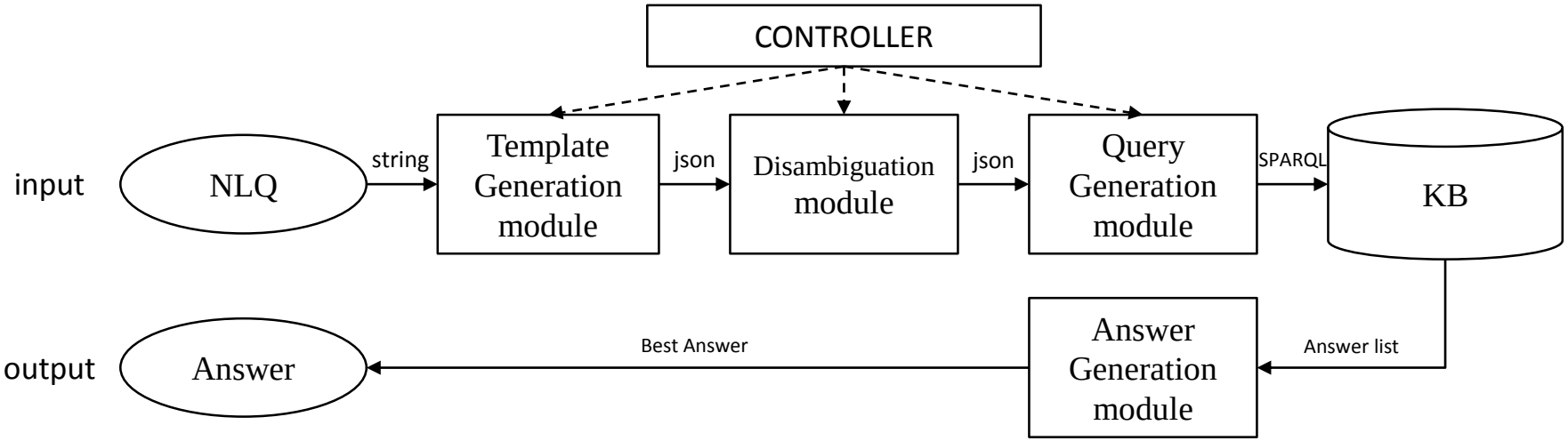
- 3rd Open Knowledge Base and Question-Answering
  - Goal:
    - Knowledge Base Construction
    - Real QA system
    - Special Track: Biomedical domain
      - infectious disease, brain tumor(optional)
  - Members

Axel Ngonga	ULeipzig	BioASQ
Christina Unger	UBielefeld	QALD
Jin-Dong Kim	DBCLS	LODQA
André Freitas	UPassau	
Kousaku Okubo	NIG	

OKBQA – BioMedical QA



OKBQA – REAL QA system



- NL query understanding
- SPARQL query generation
- Federated search & Answer rendering
- Shared task setting

MEMBERS internal

Professor

Key-sun Choi
(kschoi@kaist.ac.kr)

SWRC Leader

Ph.d student

Eun-kyung Kim
(kekeeo@kaist.ac.kr)

Wikipedia analysis for
common ontology, multilinguality

Masters student

Youngsik Kim
(twilight@kaist.ac.kr)

Entity Linking, URI disambiguation

Masters student

Jongsung Woo
(woo88@kaist.ac.kr)

Ontology alignment,

Ph.d student

Younggyun Hahm
(hahmyg@kaist.ac.kr)

Language to Knowledge,
Korean FrameNet

Ph.d student

Khiati Abdel Ilah Zakaria
(zakria.ai@gmail.com)

Multilingual Entity Linking

Masters student

Yousung Won
(styner0305@kaist.ac.kr)

NLP2RDF, Sentence simplification

Masters student

Jiseong Kim
(jiseong@kaist.ac.kr)

Enriching KB (C2K),
Korean WordNet

MEMBER internal

Masters student

Jiwoo Seo
(jiwoo35@kaist.ac.kr)

Ontology linking, NLP2RDF

Researcher

Sanha Nam
(nam.sangha@kaist.ac.kr)

Open IE

Researcher

Hejing Jeong
(hjjeong@world.kaist.ac.kr)

Principal Administrator

Masters student

Jung-wook Kim
(prismriver@kaist.ac.kr)

Korean FrameNet parser

Researcher

Sejin Nam
(namsejin@world.kaist.ac.kr)

Language Resource Management

REFERENCE

1. Youngsik Kim, Key-sun Choi, 2015, *Entity Linking Korean Text: An Unsupervised Learning Approach using Semantic Relations*, CoNLL 2015
2. Yousung Won, Jongsung Woo, Jisung Kim, Younggyun Hahm, Key-sun Choi, 2015, *한국어 서술어와 지식베이스 프로퍼티의 연결에 관한 연구*, KCC 2015
3. Jiseong Kim, Eun-Kyung Kim, Yousung Won, Sangha Nam and Key-Sun Choi, *The Association Rule Mining System for Acquiring Knowledge of DBpedia from Wikipedia Categories*, NLP&DBpedia 2015, ISWC
4. Eun-Kyung Kim, Sangha Nam, Jongsung Woo, Senjin Nam and Key-sun Choi, 2015, *Triple-based Similarity Propagation for Linked Data Matching*, The 10th International Workshop on Ontology Matching, ISWC

감 사 합 니 다

Presenter: 함영균
Contact: hahmyg@kaist.ac.kr