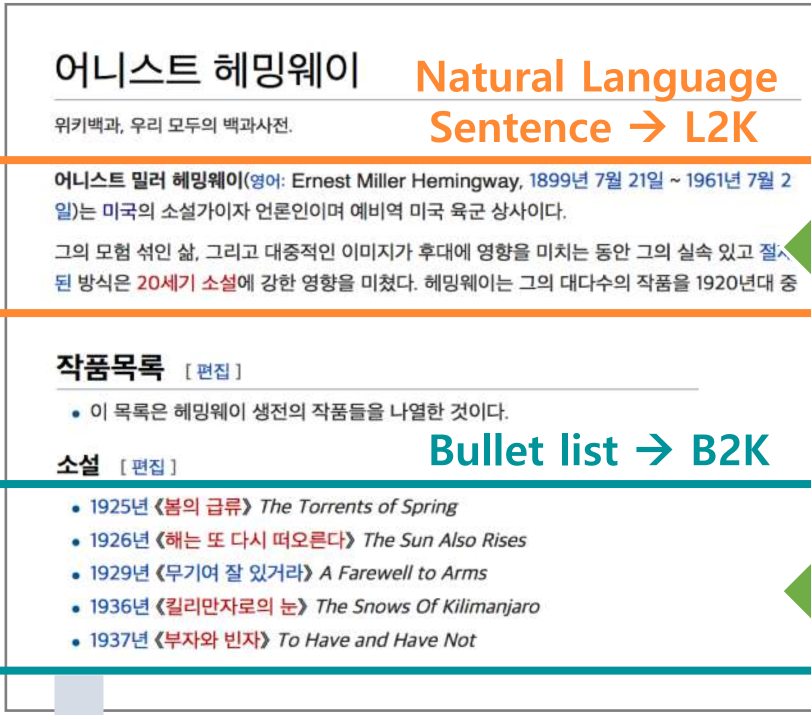
Enriching DBpedia by Knowledge Base Population and Dark Entity Resolution

Key-Sun Choi
kschoi@kaist.ac.kr

School of Computing
Korea Advanced Institute of Science & Technology
(www.kaist.edu)

Overview of Fact Triplet Extraction


어니스트 헤밍웨이

위키백과, 우리 모두의 백과사전.

어니스트 밀러 헤밍웨이(영어: Ernest Miller Hemingway, 1899년 7월 21일 ~ 1961년 7월 2일)는 미국의 소설가이자 언론인이며 예비역 미국 육군 상사이다.

그의 모험 섞인 삶, 그리고 대중적인 이미지가 후대에 영향을 미치는 동안 그의 실속 있고 절제된 방식은 20세기 소설에 강한 영향을 미쳤다. 헤밍웨이는 그의 대다수의 작품을 1920년대 중

작품목록 [편집]

- 이 목록은 헤밍웨이 생전의 작품들을 나열한 것이다.

소설 [편집]

- 1925년 《봄의 급류》 *The Torrents of Spring*
- 1926년 《해는 또 다시 떠오른다》 *The Sun Also Rises*
- 1929년 《무기여 잘 있거라》 *A Farewell to Arms*
- 1936년 《킬리만자로의 눈》 *The Snows Of Kilimanjaro*
- 1937년 《부자와 빈자》 *To Have and Have Not*

Natural Language Sentence → L2K

Bullet list → B2K

Learning

Learning

Knowledge Base

Subject	Relation	Object
Ernest_Hemingway	award	Nobel_Prize_in_Literature
	birthPlace	United_States
	notableWork	The_Old_Man_and_the_Sea
		For_Whom_the_Bells_Tolls
		A_Farewell_to_Arms

Extraction

- nationality(Ernest_Hemingway, Unites_States)
- occupation(Ernest_Hemingway, Novelist)
- occupation(Ernest_Hemingway, Journalist)

- notableWork(Ernest_Hemingway, The_Torrents_of_Spring)
- notableWork(Ernest_Hemingway, The_Sun_Also_Rises)

Goal and Challenges

- Goal
  - Constructing an iterative learning platform based on Distant Supervision to improve knowledge learning quality
- Challenges
  - 1) How to build a good quality & quantity initial knowledge base?
  - 2) How to make a reliable knowledge base population system?

Machine need to learn from text

- Corpus
  - Wikipedia, News, Blog, Twitter, Dialog, …
- Knowledge Base
  - DBpeida, Freebase, Wikidata, …

WIKIPEDIA
The Free Encyclopedia

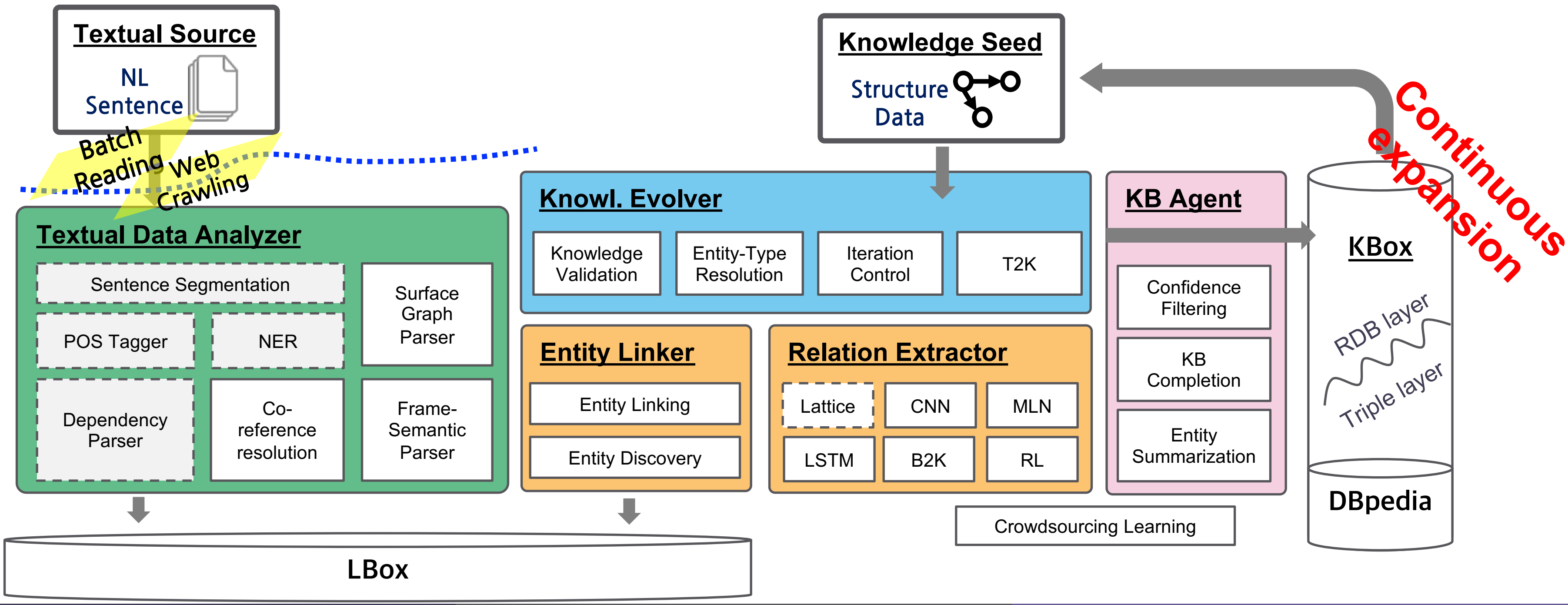
DBpedia

WIKIDATA

Freebase

KBox : Extended DBpedia

- KBox is a new KB that expands Korean DBpedia.
  - follows the DBpedia ontology schema.
  - continually expands much of the knowledge extracted from text using the Korean KBP system.



Kbox Initialization, kbox.kaist.ac.kr
- Korean enriched DBPedia

KBox Initial

KBox Initial	# of triples
Type inference using automatic type inference system	8,874,922
Type inference using **Wikidata**	8,814,984
Triple extraction and merge from **ENG.Dbpedia** using :sameAs link	6,628,484
Convert local property to DBO ontology triple	5,645,729
Korean DBpedia	3,967,566

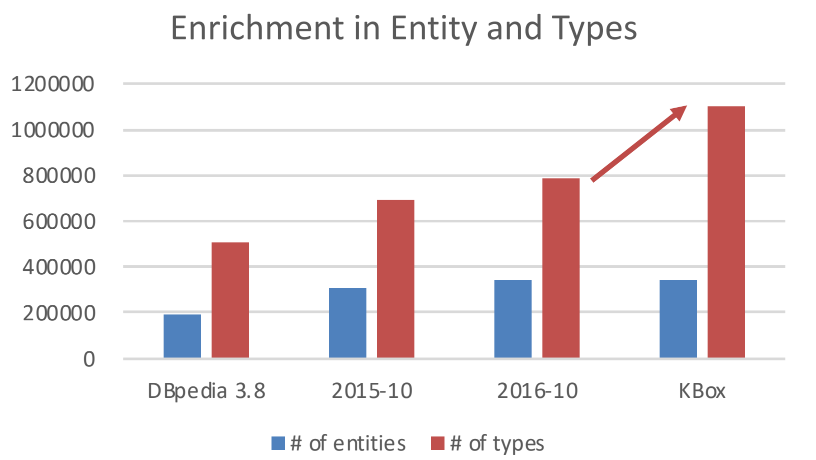
Kbox Initialization

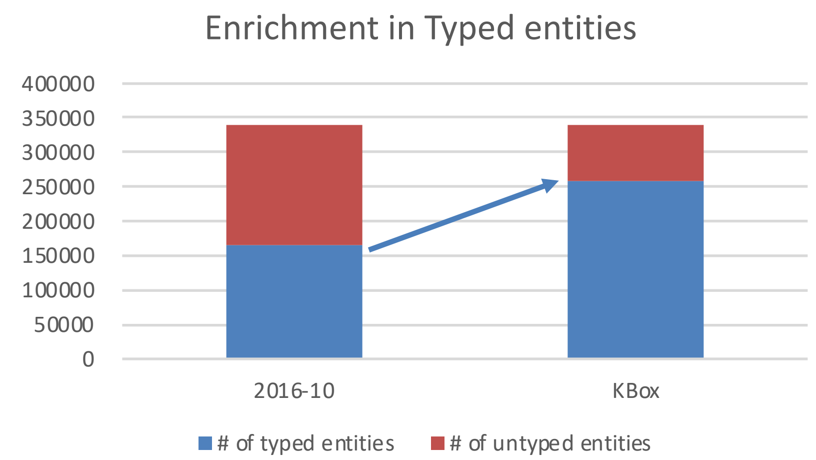
- How to build a good quality & quantity initial knowledge base

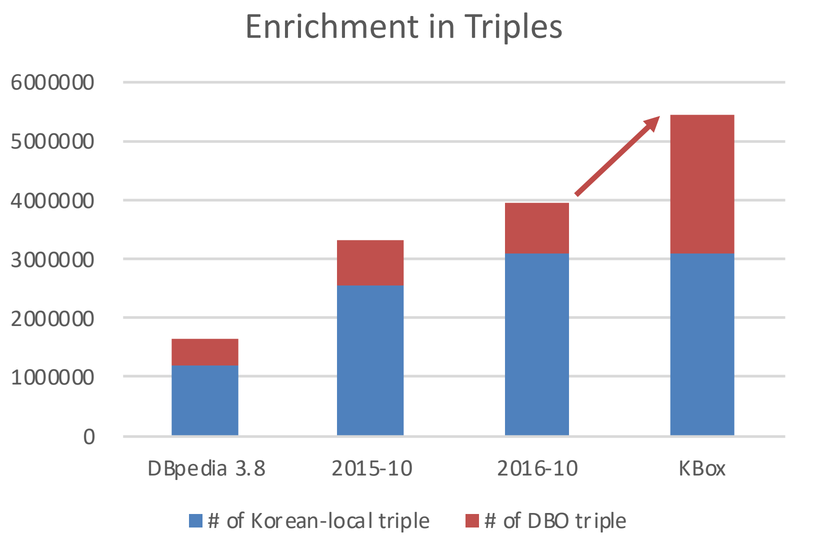
- 2-layer Storage
  - RDB (MySQL) : All the information about all the KBox triples, such as triple score, extraction module, source sentence.
  - Triple Store (Stardog, Virtuoso) : Reliable triples
    - 1) The initial triples extracted from the DBpedia and Wikidata
      - 1-1) Convert local property (prop-ko) to DBO property (출생지 → birthPlace)
      - 1-2) Type inference using :sameAs link from EN.DBpedia
      - 1-3) Triple generation using :sameAs link from EN.Dbpedia and Wikidata
    - 2) Automatically extracted triples using the Korean KB Population with a confidence score above 0.9

KBox Statistics

(How to build a good quality & quantity initial knowledge base?)

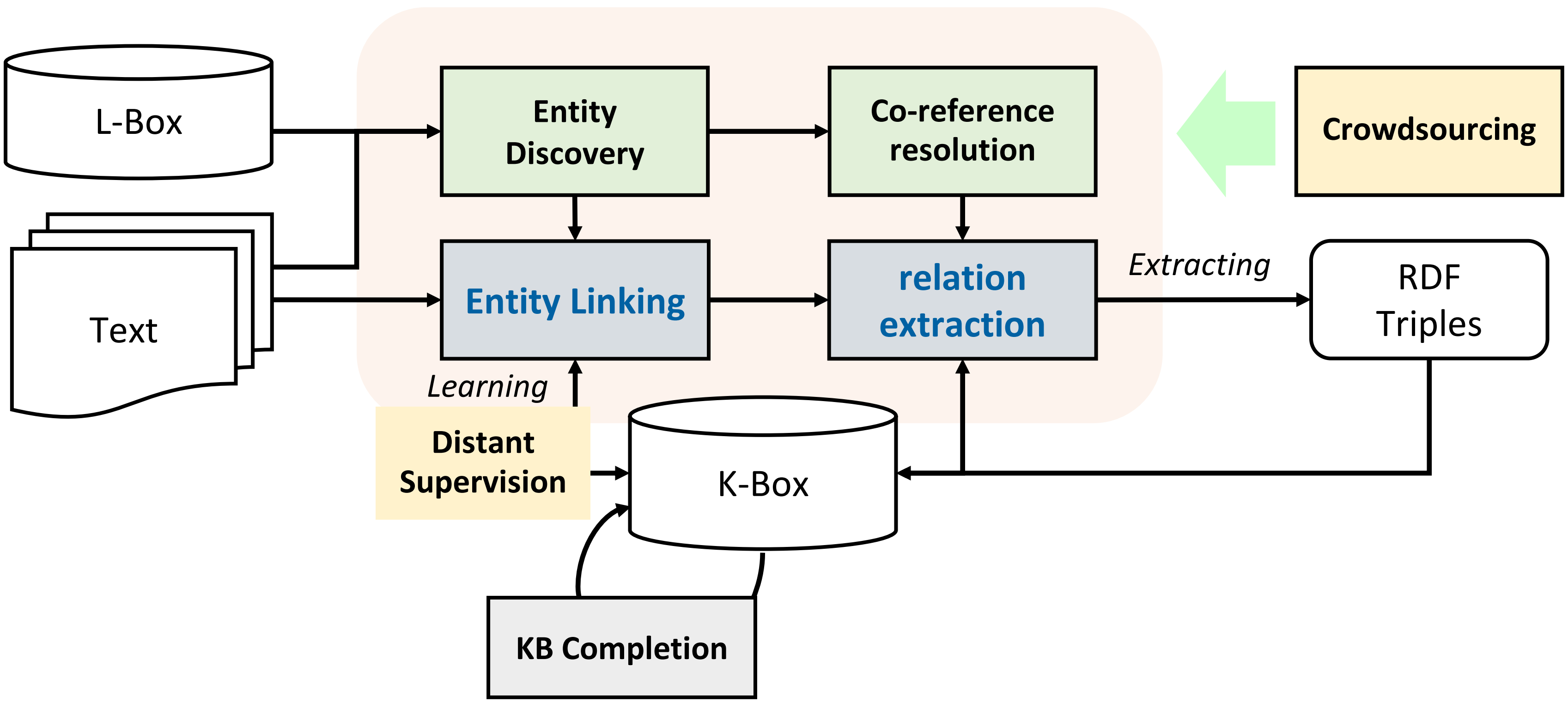






Version	# of Korean local triple	# of DBO triple	# of relational triples
DBpedia 3.8	1,219,355	432,600	1,651,649
2015-10	2,569,432	738,599	3,308,031
2016-10	3,077,297	890,269	3,967,383
KBox	**3,077,297**	**2,350,292**	**5,426,857**

Essential Tasks for KB Population



Example of Knowledge Learning / Extracting

Extracted Knowledge

- Agent Person Artist
  - 마사_겔혼 (Martha_Gellhorn) ★
    - *spouse* 어니스트_헤밍웨이 (Ernest_Hemingway)
    - *is spouse of* 어니스트_헤밍웨이 (Ernest_Hemingway)
  - 어니스트_헤밍웨이 (Ernest_Hemingway) ★
    - *award* 노벨_문학상 (Nobel_Prize_in_Literature)
    - *birthPlace* 미국 (United_States)
    - *birthPlace* 일리노이_주 (Illinois)
    - *deathPlace* 미국 (United_States)
    - *deathPlace* 아이다호_주 (Idaho)
    - *notableWork* 노인과_바다 (The_Old_Man_and_the_Sea)
    - *notableWork* 누구를_위하여_종은_울리나 (For_Whom_the_Bell_Tolls)
    - *notableWork* 무기여_잘_있거라 (A_Farewell_to_Arms)
    - *occupation* 소설가 (Novelist)

어니스트 헤밍웨이는 미국의 소설가이자 저널리스트이다. 1854년 노벨 문학상을 수상하였다. 헤밍웨이는 1899년 7월 21일 일리노이주에서 태어났다. 헤밍웨이는 풀린 파이퍼와 이혼한 뒤 마사 겔혼과 재혼하였다. 19 다호 주에서 엽총으로 62세의 나이에 자살했다.

Type: Artist
String: 헤밍웨이
Kor_Entity: 어니스트_헤밍웨이
Eng_Entity: Ernest_Hemingway

- 1926년 《해는 또다시 떠오른다》 The Sun Also Rises.
- 1929년 《무기여 잘 있거라》 A Farewell to Arms.
- 1940년 《누구를 위하여 종은 울리나》For Whom the Bell Tolls.
- 1950년 《강 건너 숲속으로》 Across the River and Into the Trees.
- 1952년 《노인과 바다》 The Old Man and the Sea.

Relation Extraction ← **Entity Linking**

Problem and solutions in KB Population

How to make a reliable knowledge base population system?

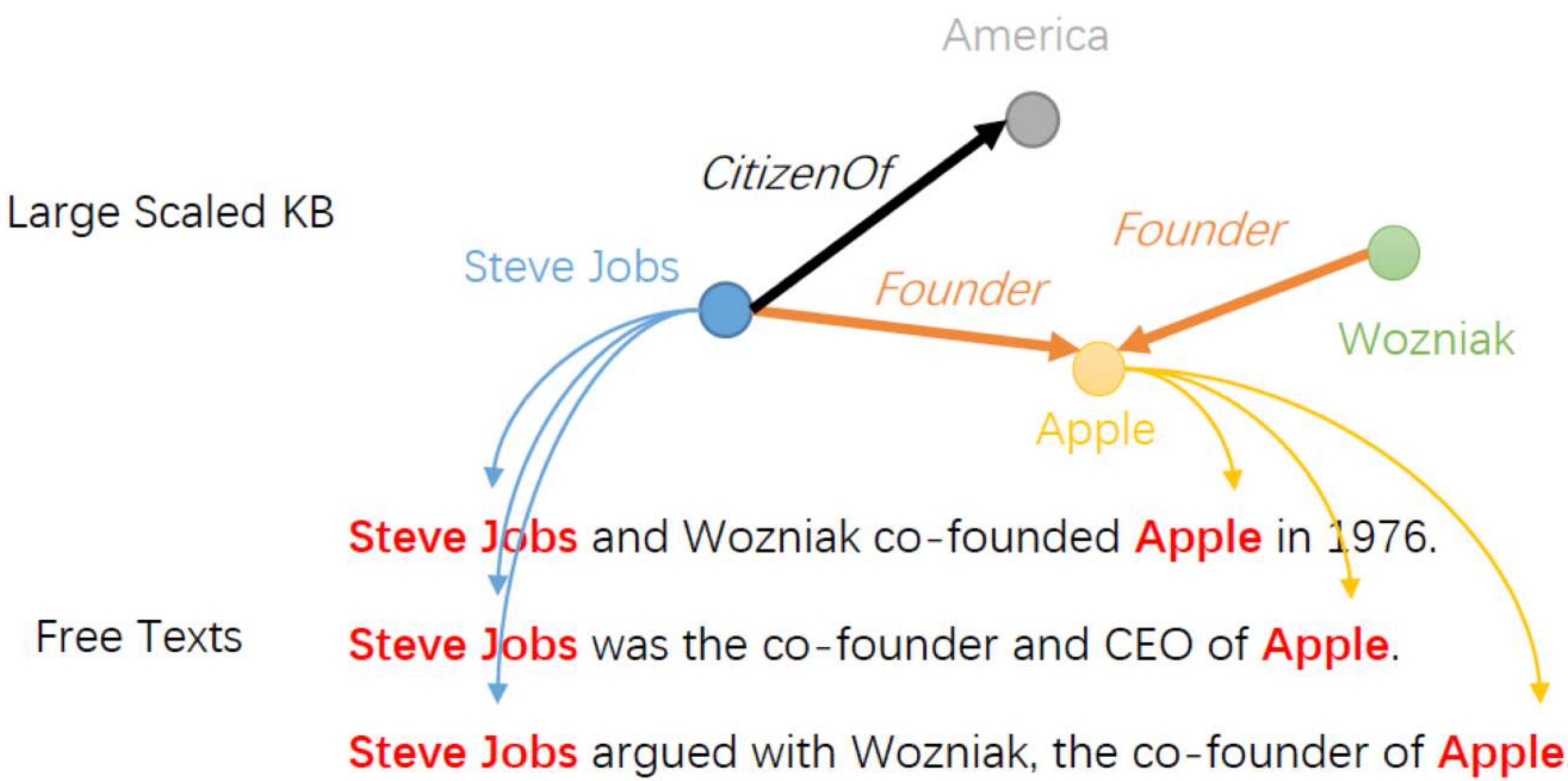
Problems:	Solutions:
Noise data on distant supervision	Enhanced noise filtering using crowdsourcing and reinforcement learning → **Increase precision**
Hard to understand a long and complex sentence	Diversify relation extraction model, Co-reference resolution and Zero anaphora detection
Difficulty in interpreting ontology relation for all collected sentences	Full-text graph generation → **Increase recall**
Incompleteness of knowledge base	Dark entity, Iterative Learning, Knowledge Completion and Summarization

Distant Supervision (Relation Extraction)

- Distant Supervision (Mintz et al. 2009)
  - For a triple fact r(e1, e2) in a KB, all sentences that mention both entities e1 and e2 are aligned with relation r.

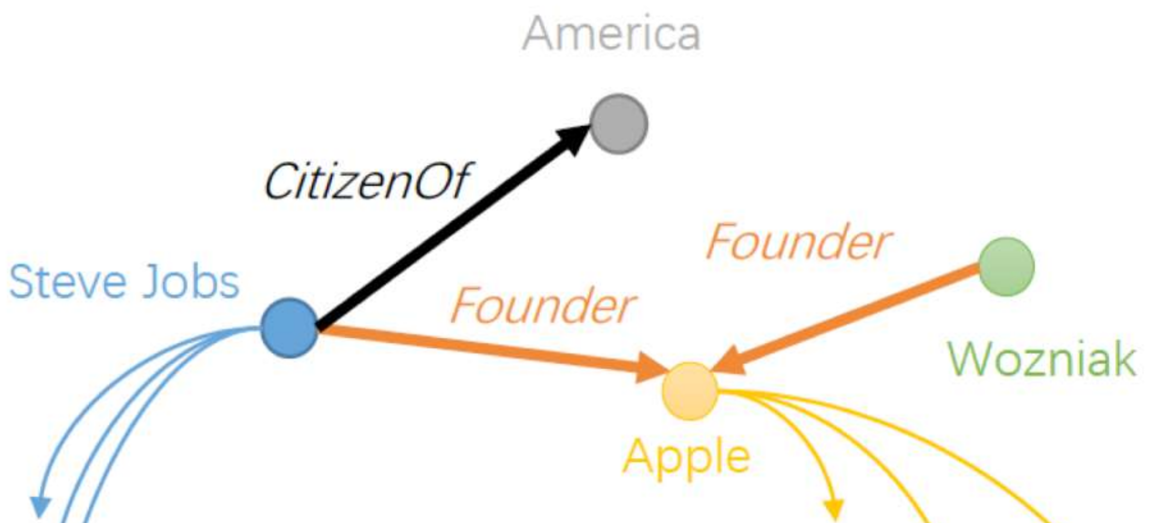
Sentences	Relation
1. **Steve Jobs** and Wozniak co-founded **Apple** in 1976.	*Founder*
2. **Michael Jordan** is an American retired professional **basketball player**.	*Career*
3. **Washington D.C.** is the capital of **United states**.	*CapitalOf*
......	......

Automatic labeling between KB and sentence



Problem : Noise data

Large Scaled KB



Free Texts

Steve Jobs and Wozniak co-founded **Apple** in 1976.

Steve Jobs was the co-founder and CEO of **Apple**.

Steve Jobs argued with Wozniak, the co-founder of **Apple**.

Wrong label

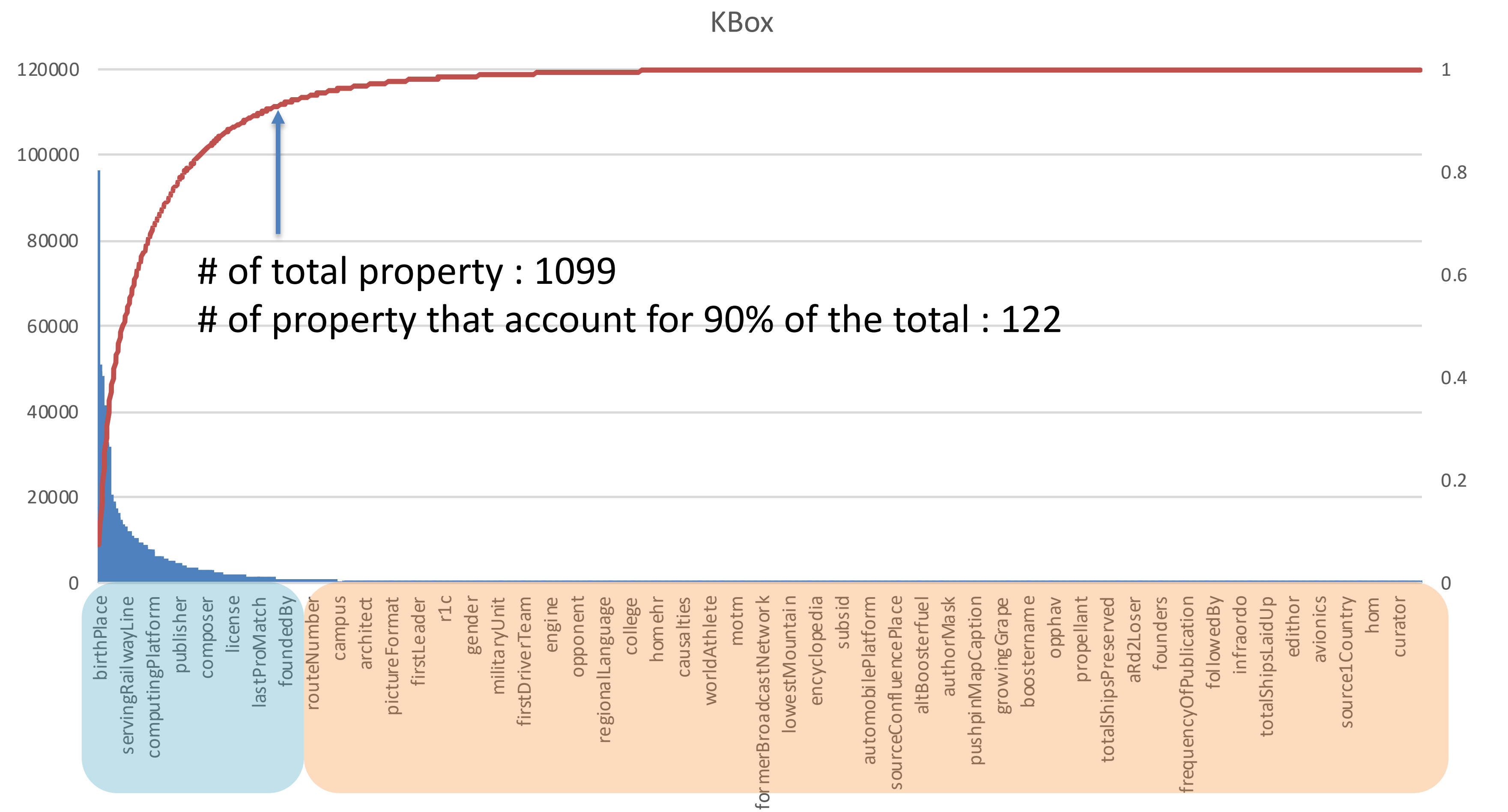
Problem : Noise data

- Average error rate reported in this paper[1]
  - 74.1% (DS data from Wikipedia – YAGO)
  - 31.0% (DS data from NYT News – Freebase)

- **Intent** : Up to 20% performance improvement by learning models with noise-free data (in our experiments)

1. Ru, C., Tang, J., Li, S., Xie, S., & Wang, T. (2018). Using semantic similarity to reduce wrong labels in distant supervision for relation extraction. Information Processing & Management.

Sentence collecting problem and Imbalanced-labeled data

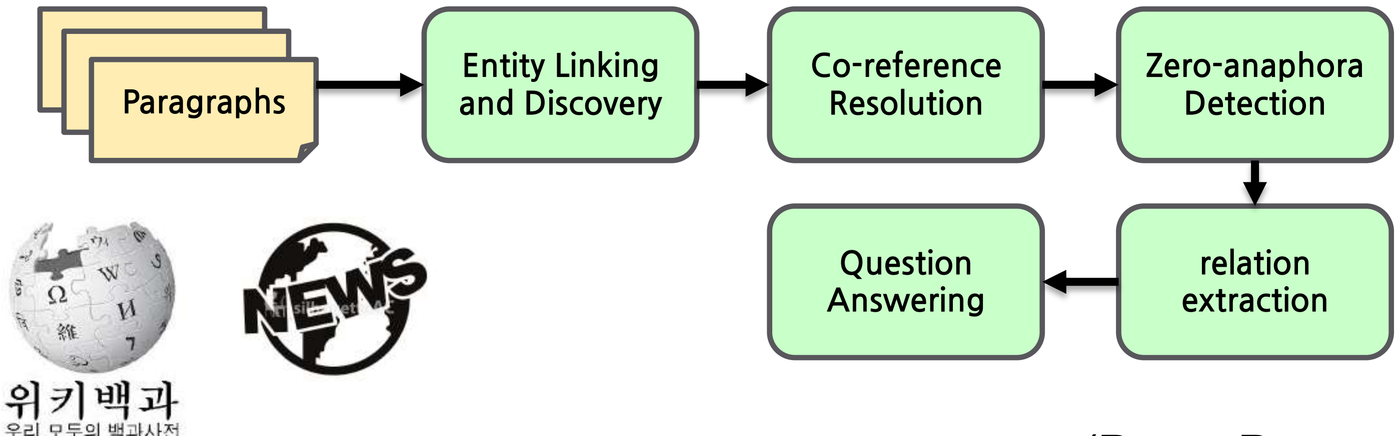


Deep-learning Approach

Rule-based Approach

Solution 1 : Crowdsourcing

- Goal : Improving the performance of knowledge learning models with human-annotated dataset on 5 tasks



(Par. = Paragraph)

Corpus	Target Tasks	Now	Goal
Wikipedia	Task 1-5	500 Par.	40K Par.
News	Task 4 relation extraction	5,482 Par.	60K Par.

What makes increase Recall

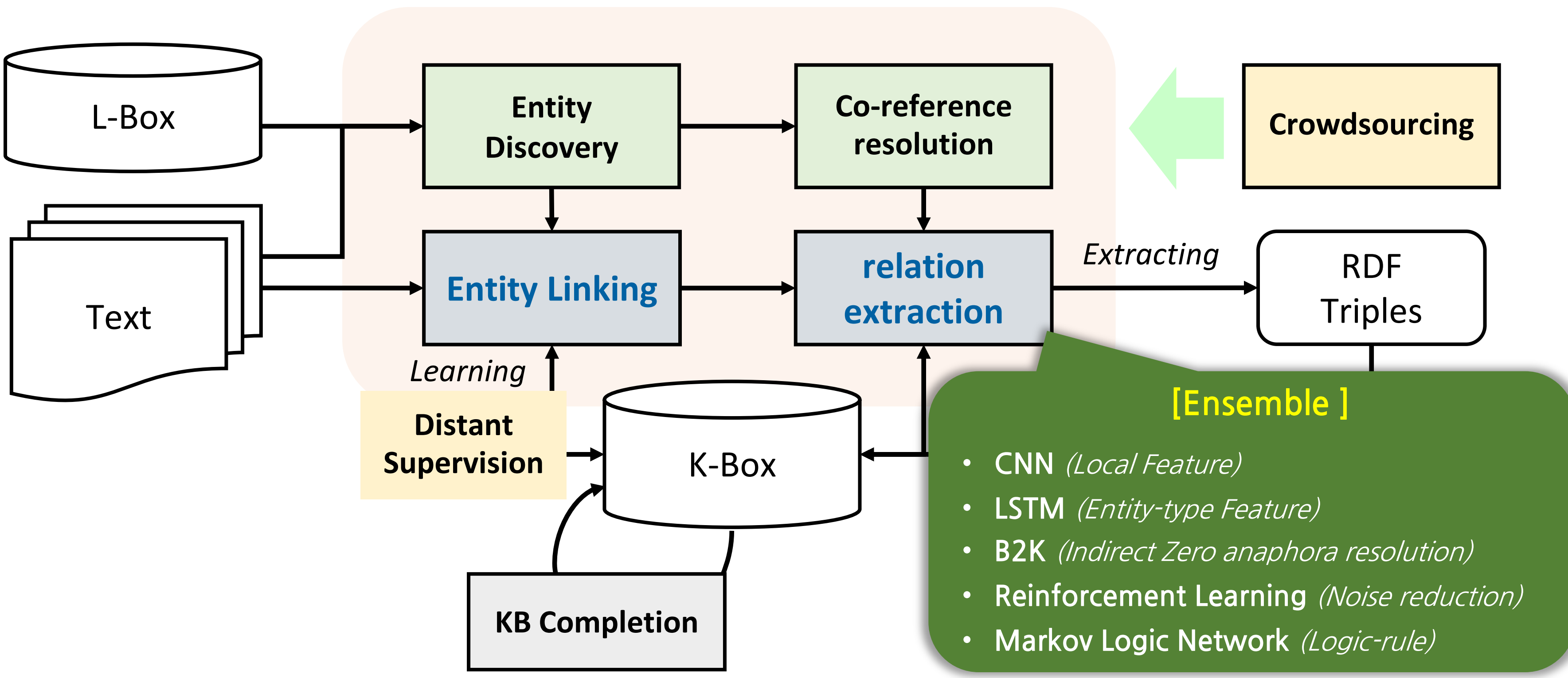
- Entity Discovery
  - Finding new dark entities that can be registered to the KB
- Co-reference resolution
  - **Intent** : Up to 21.6% improved knowledge extraction coverage using Entity Discovery and Co-reference resolution

Crowdsourced Gold Set Analysis Result

# of sentence	Sentence with no or only one entity. (Exception from knowledge extraction target)	Sentences that are included in knowledge extraction targets with Entity Discover / Coreference Resolution	Coverage
1,435	445	96	**21.6% (can be improved)**

Solution 2 : Ensemble

- Even a state-of-the-art relation extraction model shows low performance (F1-score, 40-50%).



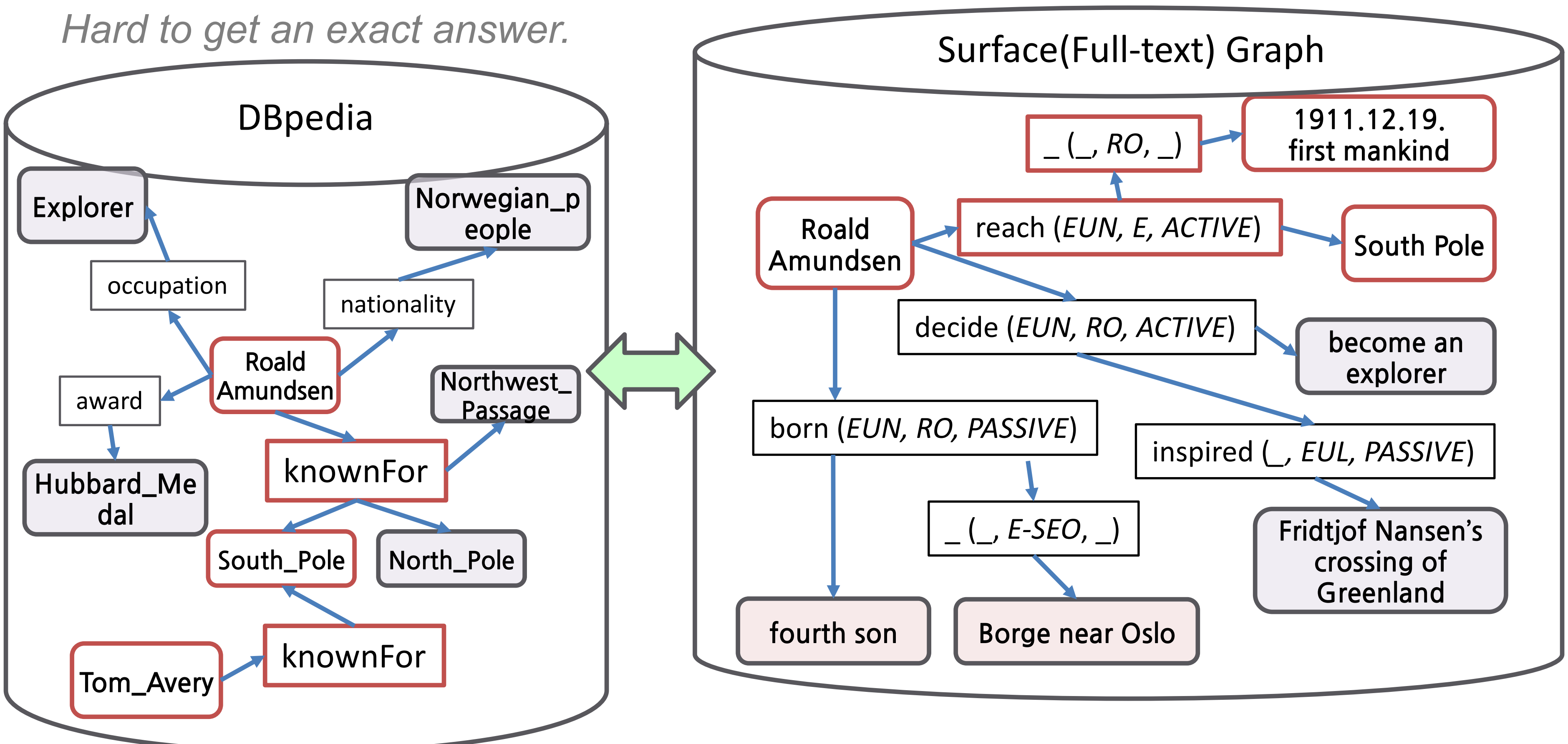
Full-text Graph

- Purpose
  - KBP/QA coverage expansion by combining ontological KB and full-text graph
- Lack of expression
  - There is a lot of knowledge that can not be expressed in an ontological relation.
- ISO/TC37/SC4 Proposal: Surface Knowledge graph with Linguistic Content

Full-text Graph: Example of Application

Question : Who was the first person to reach the South Pole? (남극점에 최초로 도달한 사람은 누구인가?)
<u>Answer : Roald Amundsen</u>

Hard to get an exact answer.



Application: www.OKBQA.org

(Open Knowledge Base and Question Answering)