

비정형 텍스트에서 Knowledge Graph 자동 생성을 위한 엔티티-관계 추출

최준호¹, 김형주²

조선대학교 자유전공학부¹, 조선대학교 컴퓨터공학과²

e-mail : xdman@chosun.ac.kr, snowlisakim@gmail.com

Entity-relation extraction for automatic knowledge graph generation from unstructured text

Jun-Ho Choi¹, Hyung-Ju Kim²

Division of Undeclared Majors, Chosun University¹

Department of Computer Engineering, Chosun University²

요 약

지식베이스 생성 및 관리 분야는 학습데이터의 확보와 학습된 지식의 검증, 그리고 지식베이스의 유지보수 문제 등에 대한 새로운 방법론이 요구되고 있다. 특히, 지식베이스는 시맨틱 레벨에서 지식기반 딥러닝 알고리즘의 학습데이터로 활용할 수 있는 방안이 활발히 연구되고 있다. 기존 지식처리 방식은 비정형 텍스트 데이터에서 자연어 처리기반 통계적 학습을 통한 지식 추출에 머물러 있다. 지식베이스 자동구축은 개념과 관계를 자동으로 추출하여 확장하는 연구가 진행되고 있지만, 텍스트 기반의 통계적인 방법에 국한된 상황이다. 본 논문에서는 비정형 텍스트에서 엔티티와 관계를 추출하여 자동으로 Knowledge Graph를 생성하는 방법을 제안한다.

1. 서 론

지식베이스 생성 및 관리 분야는 학습데이터의 확보와 학습된 지식의 검증, 그리고 지식베이스의 유지보수 문제 등에 대한 새로운 방법론이 요구되고 있다. 특히, 지식베이스는 시맨틱 레벨에서 지식기반 딥러닝 알고리즘의 학습데이터로 활용할 수 있는 방안이 활발히 연구되고 있다 [1]. 약한 인공지능(Artificial Narrow Intelligence(ANI)) 영역의 머신러닝은 특정 분야에서 저차원 데이터 위주의 학습을 수행하는 반면, 강한 인공지능(Artificial General Intelligence(AGI)) 영역에서는 Knowledge Graph를 기반으로 딥러닝을 통해 기존 지식과 통합하여 새로운 지식을 추출하는 기술이 핵심 기술로 떠오르고 있다[2].

기존 지식처리 방식은 비정형 텍스트 데이터에서 자연어 처리기반 통계적 학습을 통한 지식 추출에 머물러 있다. 지식베이스 자동구축은 개념과 관계를 자동으로 추출하여 확장하는 연구가 진행되고 있지만, 텍스트 기반의 통계적인 방법에 국한된 상황이다. Knowledge Graph는 지식 생성 및 추출을 위한 효율적인 접근 방식을 제공하고, 비정형 및 구조화된 데이터를 결합하는 데이터 통합 분야에서 중요하게 활용되고 있는 지식표현 방법이다. 본 논문에서는 비정형 텍스트에서 엔티티와 관계를 추출하여 자동으로 Knowledge Graph를 생성하는 방법을 제안한다.

2. Knowledge Graph 자동 생성을 위한 엔티티-관계 추출

2.1 위키피디아 웹 스크래핑

지식 그래프를 생성하는 방법은 생성하고자 하는 텍스트 데이터의 소스를 수집하는 것이다. 위키피디아는 지식을 생성하는데 필요한 신뢰도가 높은 데이터를 포함하고 있기에 지식 정보를 추출하는 중요한 소스 데이터이다. 본 논문에서는 특정 대상과 관련된 다수의 위키피디아 페이지에서 문서의 제목, 본문, 카테고리 정보, 내부링크, 외부링크 정보 등을 추출하여 사용한다.

page	text	link	categories	topic
0	Microsoft Corporation is an American multinational...	https://en.wikipedia.org/wiki/Microsoft	[2075 establishments in New Mexico, 1886s Microsoft Init...	Microsoft
1	32-bit computing In computer architecture, 32-bit integers, mem...	https://en.wikipedia.org/wiki/32-bit_computing	[all articles needing additional references, A...	Microsoft
2	24/7 service In commerce and industry, 24/7 or 24-h service...	https://en.wikipedia.org/wiki/24/7_service	[Articles with short description, Short descri...	Microsoft
3	.NET Foundation The .NET Foundation is an organization incorpo...	https://en.wikipedia.org/wiki/.NET_Foundation	[.NET, 501(c)(3) nonprofit organizations, All ...	Microsoft
4	32-bit computing In computer architecture, 32-bit integers, mem...	https://en.wikipedia.org/wiki/32-bit_computing	[Articles using small message boxes, Articles Comput...	Microsoft
...	...	...	...	...
932	Category:Software Companies	https://en.wikipedia.org/wiki/Category:Software...	[Commons category link is on Wikidata, Microsoft	Microsoft
933	Category:Use American English from May 2020 This category combines all use American English...	https://en.wikipedia.org/wiki/Category:Use_Ame...	[CekAutoTC generates standard Category TOC, C...	Microsoft
934	Portal:United States	https://en.wikipedia.org/wiki/Portal:United_St...	[All portals, All portals with triaged subpage...	Microsoft
935	Category:Use mdy dates from October 2020 Wikipedia articles tagged in this month sha...	https://en.wikipedia.org/wiki/Category:Use_mdy...	[CekAutoTC generates Large category TOC, C...	Microsoft
936	Portal:Companies	https://en.wikipedia.org/wiki/Portal:Companies	[All portals, All portals with triaged subpage...	Microsoft

912 rows × 5 columns

(그림 1) 'Microsoft'에 대한 위키피디아 수집 정보

(그림 1)은 위키피디아에서 'Microsoft'에 대한 정보를 스크래핑하여 항목별로 분류한 결과이다.

2.2 엔티티-관계 정보 추출

수집된 위키피디아의 본문에서 엔티티-관계 정보를 품사 태깅과 종속 구문 분석을 이용하여 추출한다. 문법적인 패턴은 Subject-Relation-Object 형태로 표현한다. 해당 텍스트에서 명사구를 추출하는 과정에서 대명사, 기호, 접속사, 구두점 등은 제거한다. 엔티티 추출은 복합어 여부, 수식어 여부 등을 고려하였다. 다음 (그림 2-3)은 추출된 Subject와 Object를 출력한 결과이다.

```
[ 'Microsoft Enterprise References', 'External links', '16 bit', 'How Bits', 'se
cular', 'formerly Standard Libraries', 'External links', 'External links', 'also
Wn Companies portal', 'historical records', 'selected', 'External links', 'Offic
ial website', 'External Company website', 'External links', '2019 Ultimate Acorn
Archimedes', 'narrative Games', 'External links', '26 Kevin Take', 'External li
nks', 'External links', 'External links', 'External links', 'other ATT&K secur
ity community', 'External Greenleaf Book links', 'External links', 'External lin
ks', 'External Organization website', 'also Inc.', 'References', '$', 'External
links', 'Christian University References', 'External links', 'External Global Ma
thematics links', 'External Wn Internet links', '12 Wn MITS Altair', '1993 1.0',
'2015 Intel', 'disassembly Bill Gates', '13 References', '2015 Blacklane', 'Jam
es Gosling', 'affiliated legislators', 'small', 'External links', 'same', 'Exter
nal', 'External links', 'Official website', 'also', 'External Citations links',
'11 They', 'Area portal Wn Companies portal', 'How Apple', 'ethnic', 'improbab
```

(그림 2) 추출된 Subject 목록

```
[ 'website Wn Microsoft OpenCorporates', 'ASP.NET AJAX', 'e.g. CD audio', '64 201
B', 'High availability', 'Framework Base Class', '8800 Howard M. Harte', 'Externa
l Automatic Data Processing', '2007 aQuantive', 'historical Minnesota Historical
Society', 'External links', 'Archived Japanese', 'Official ASML Holding', 'damag
es policy', 'Official Wn Accenture OpenCorporates', '2021.ARM ExtensionsQuick Ref
erence CardsInstructions', 'narrative Tex Murphy series', '2020 November', 'also
Achronix', 'Mailbox Android application', 'Corporate website', 'Adobe Inc', 'Offic
ial Activision Blizzard', 'other Persistent Group Pages', 'HowStuffWorks Advanced
Micro Devices', 'Adobe Inc', 'news agencies', 'External Tokyo Akihabara Guide', 'O
fficial Akamai Technologies', 'United States', 'also United States', 'historical N
ew Curlie', 'also Alibaba Cloud ECS', 'Official Align Technology', 'Official Group
```

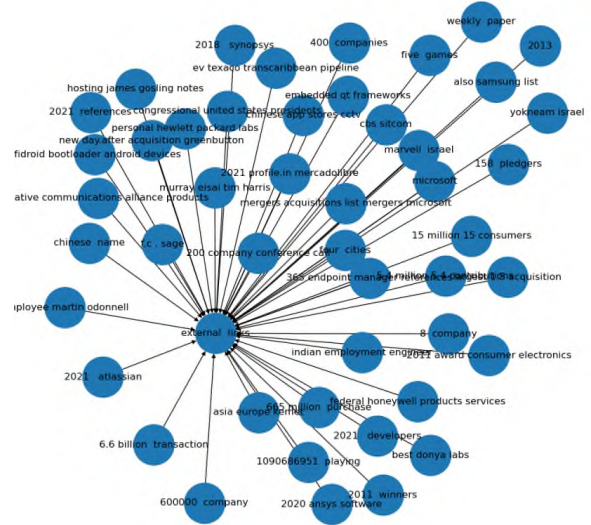
(그림 3) 추출된 Object 목록

Subject, Object 목록을 토대로 문장에서 Subject, Object를 모두 포함하는 엔티티를 추출하고, 나머지는 제거한다. 관계 추출은 규칙 기반 매칭을 이용한다. 본 논문에서 관계 추출을 위해 정의한 패턴은 $\{('DEP': 'ROOT'), ('DEP': 'prep', 'OP': '?'), ('DEP': 'agent', 'OP': '?'), ('DEP': 'ADJ', 'OP': '?')\}$ 와 같이 정의한다.

	source	target	edge
0	official microsoft companies	official opencorporates	produces
1	special list 16bit cpus	. cd audio	addresses
2	bytes	64bit windows	addresses
3	external links	1.0 altair	developed
4	2012 publisher solutions microsoft	also acquisition companies	
5	external links	external ascii corporation page	became
6	2020 acer unveils	reviewcentral	inc
7	3112013 alexander	programmable paul planet	china
8	electrical town organization website	electrical tokyo akihabara guide	tokyo
9	case appeal alcatellucent	states microsoft references	known
10	500 companies	official data align technology	invisalign
11	altair clone truetolife mits	8800 atmel avr	designed
12	first	first market	founded
13	2015 intel	16.7 billion 16.7 cash	headquartered
14	microsoft bill	msdn blogs	ran
15	hosting james gosling notes	external links	apis
16	renewable none bills	renewable 2013	owns
17	also uk patent case	also case employment law	inc
18	2006 audioasics	computer labs	known
19	2020 ansys software	external links	based

(그림 4) 엔티티-관계 정보 추출

추출된 관계 정보를 이용하여 지식 그래프로 생성한 후, 엔티티-관계를 그래프로 도식화하면 (그림 5)와 같은 결과를 표현할 수 있다.



(그림 5) 엔티티-관계 기반 Knowledge Graph 생성

3. 결론

비정형 텍스트 데이터에서 자연어 처리 기반 지식 추출 방법은 지식베이스 자동구축 분야에서 중요한 연구로 인식되고 있다. Knowledge Graph는 지식 생성 및 추출을 위한 효율적인 접근 방식을 제공하고, 비정형 및 구조화된 데이터를 결합하는 데이터 통합 분야에서 중요하게 활용되고 있는 지식표현 방법이다. 본 논문에서는 비정형 텍스트에서 엔티티와 관계를 추출하여 자동으로 Knowledge Graph를 생성하는 방법을 제안하였다. 향후, 종속 트리를 이용하여 문장에 대한 세부적인 규칙을 정의한다면 더욱 정교한 관계를 추출할 수 있을 것으로 기대된다.

감사의 글

본 논문은 2021년 교육부와 한국연구재단 지역대학우수과학자지원사업의 지원을 받아 수행된 연구임(과제번호: NRF-2021R1I1A3049061).

참고문헌

- [1] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu. Learning entity and relation embeddings for knowledge graph completion. In Proceedings of the 29th AAAI Conference on Artificial Intelligence, pp. 2181 - 2187, 2015.
- [2] B. Y. Lin, X. Chen, J. Chen, and X. Ren. Kagnet: Knowledge-aware graph networks for commonsense reasoning. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 2829 - 2839, 2019.