오픈 데이터와 인공지능

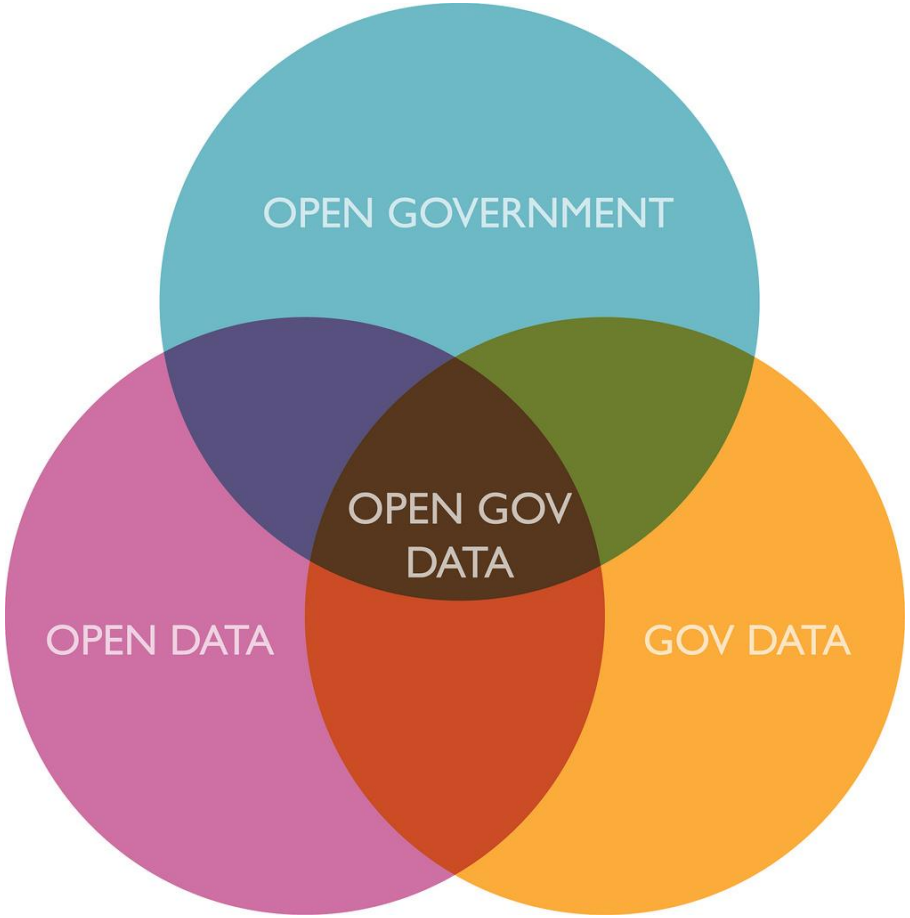
이명진

LiST, Linked Data and Semantic Web Technology

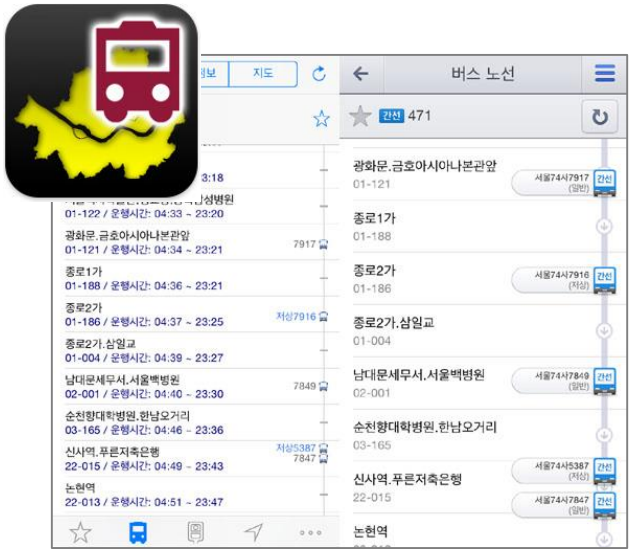
오픈 데이터와 공공데이터

- 오픈 데이터(open data)
  - 저작권이나 특허 등과 같은 제약사항 없이 모든 사람이 자유롭게 사용할 수 있는 데이터

- 공공데이터(open government data)
  - 데이터베이스, 전자화된 파일 등 공공기관이 법령 등에서 정하는 목적을 위하여 생성 또는 취득하여 관리하고 있는 광(光) 또는 전자적 방식으로 처리된 자료 또는 정보
  - “기계 판독이 가능한 형태”로 “제공”

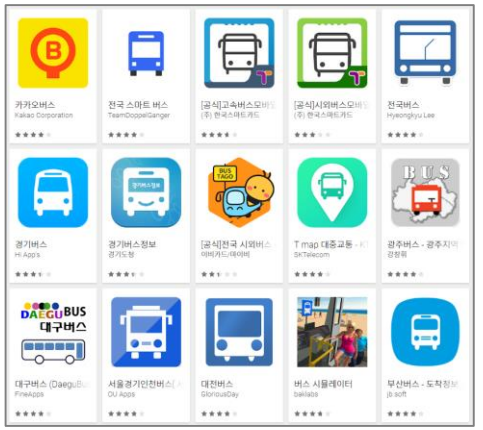


오픈 데이터의 중요성



HTML 스크랩핑 → 경기도 서버에서 앱 차단 → 오픈 API로 개방 → 광고 게재 → 네이버 지원 → 공공데이터의 상업적 이용 제한

- 카카오 인수
- 공공데이터의 제공 및 이용 활성화에 관한 법률 (약칭: 공공데이터법)
  [시행 2013.10.31.] [법률 제11956호, 2013.7.30., 제정]

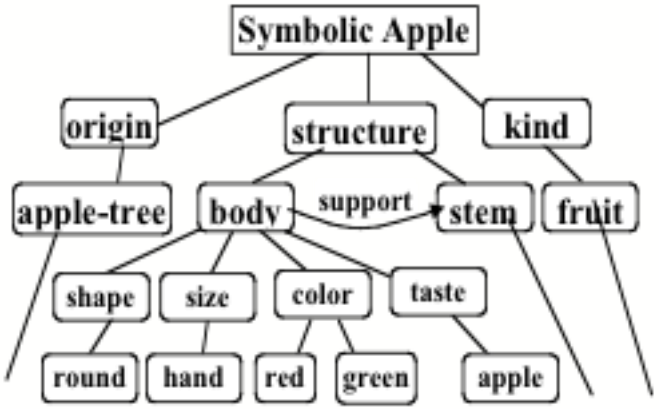


인공지능(Artificial Intelligence)

- 기계로부터 만들어진 인공적인 지능
- 학습과 문제해결과 같은 인간의 인지 기능을 모방하여 만들어진 기계의 지능

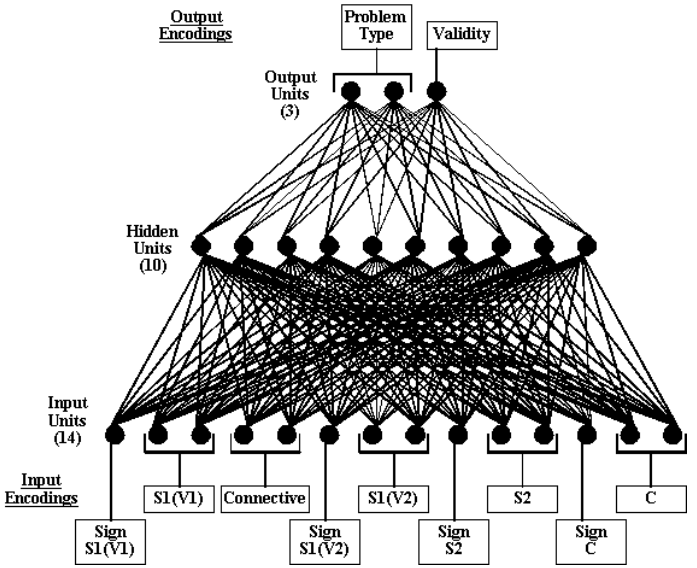
Symbolic

- 인간의 지식을 기호로 표시하고 이를 바탕으로 논리, 검색, 문제 표현 등을 처리
- 1950년대 중반부터 1980년대 후반까지 AI 연구의 지배적 패러다임



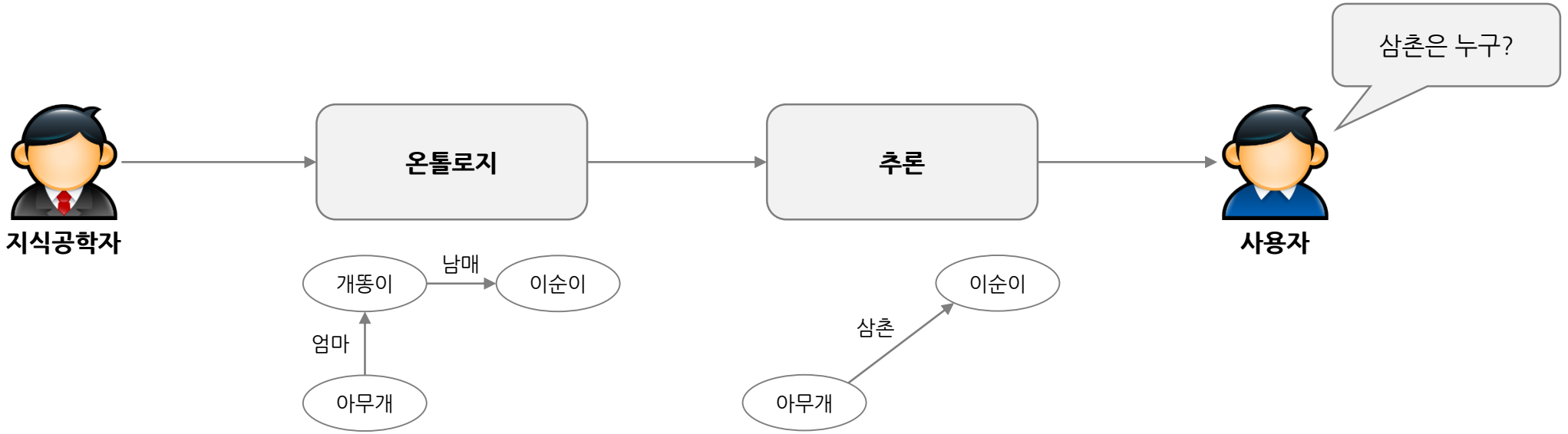
Sub-symbolic

- 지식의 표현 없이 기계학습과 같은 방법을 이용하여 학습, 패턴 인식과 같은 분야에 활용
- 1980년대부터 본격적인 연구가 시작



지식베이스를 활용한 인공지능

- 지식을 수집하고
- 이를 컴퓨터가 처리 및 이해할 수 있는 형태로 구축 및 저장한 후
- 추론, 검색 등과 같은 처리를 통해
- 사용자의 의사결정, 질의응답과 같은 지능형 처리를 지원

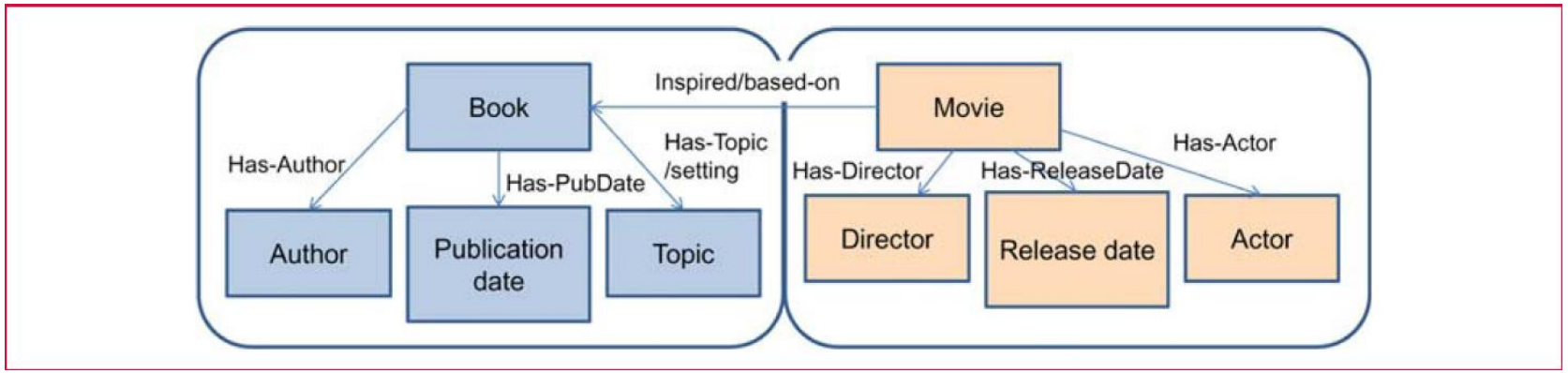


IBM Watson

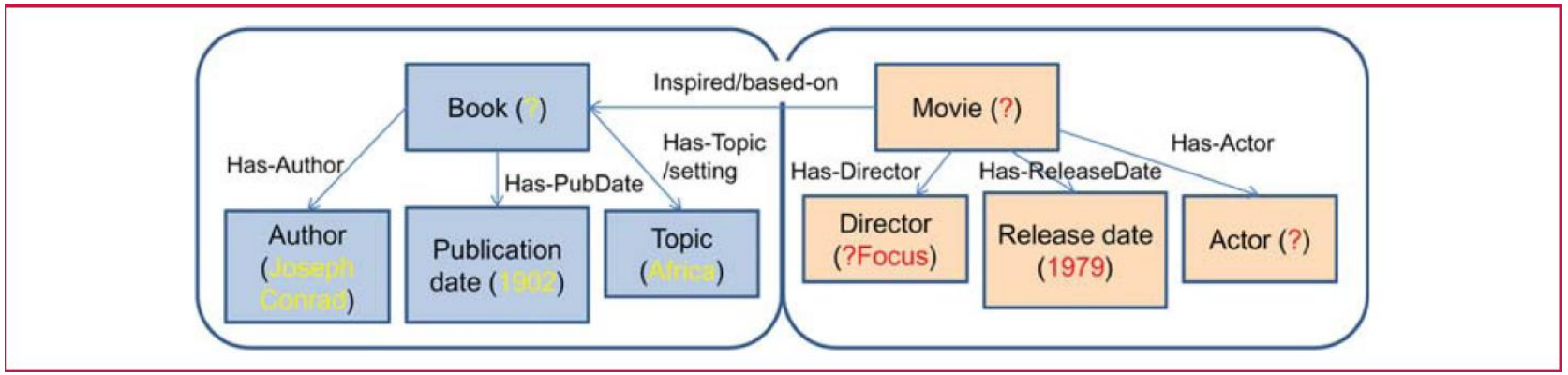
- 자연어 형식으로 된 질문들에 응답할 수 있는 질의응답 인공지능 컴퓨터 시스템
- 2011년에 Jeopardy! 퀴즈쇼에서 우승

Ken Jennings,
74번 Jeopardy! 챔피언으로
$3,422,700 번 남자

Brad Rutter,
Jeopardy!에서
가장 돈 많이 번 남자
($4,455,102)



WAR MOVIES: A 1902 Joseph Conrad work set in Africa inspired this director to create a controversial 1979 war film.

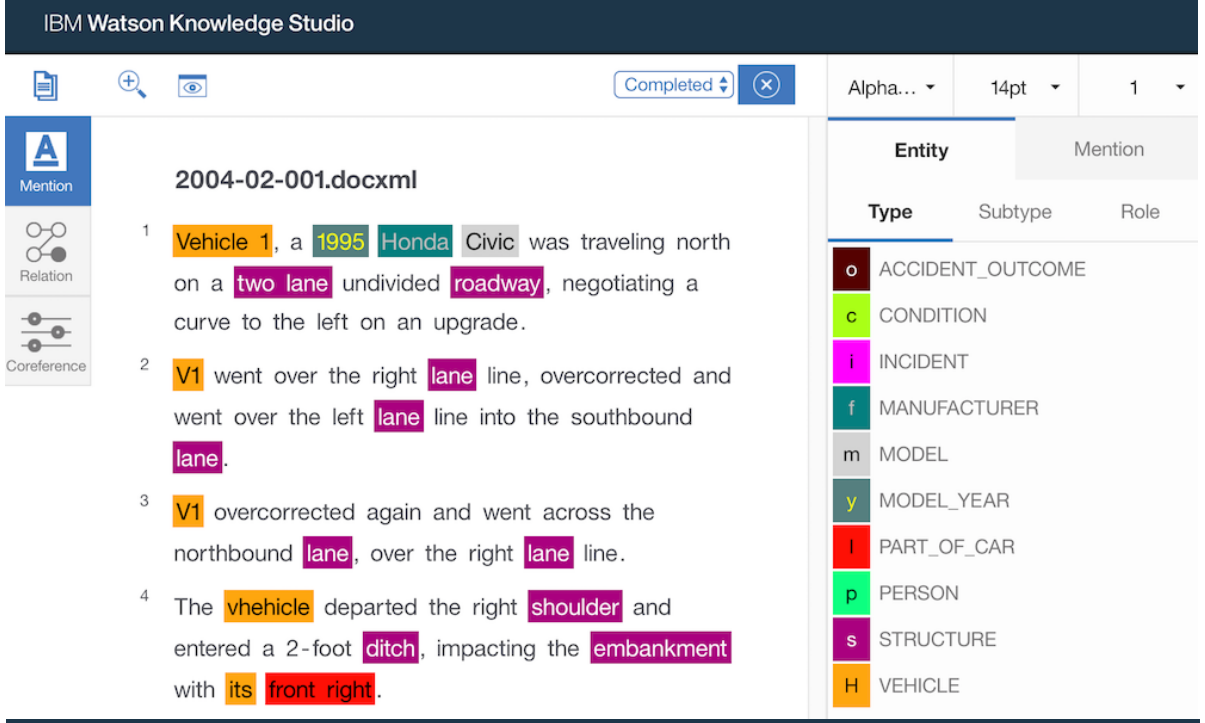


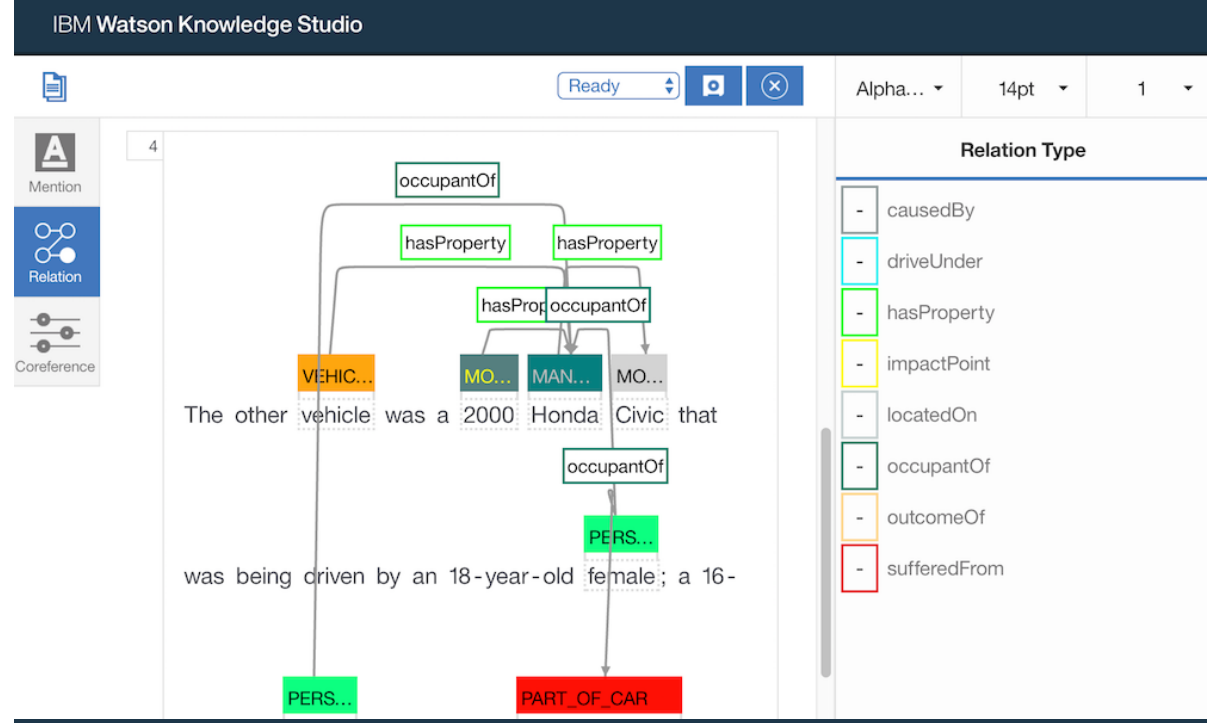
answer: “Francis Ford Coppola”

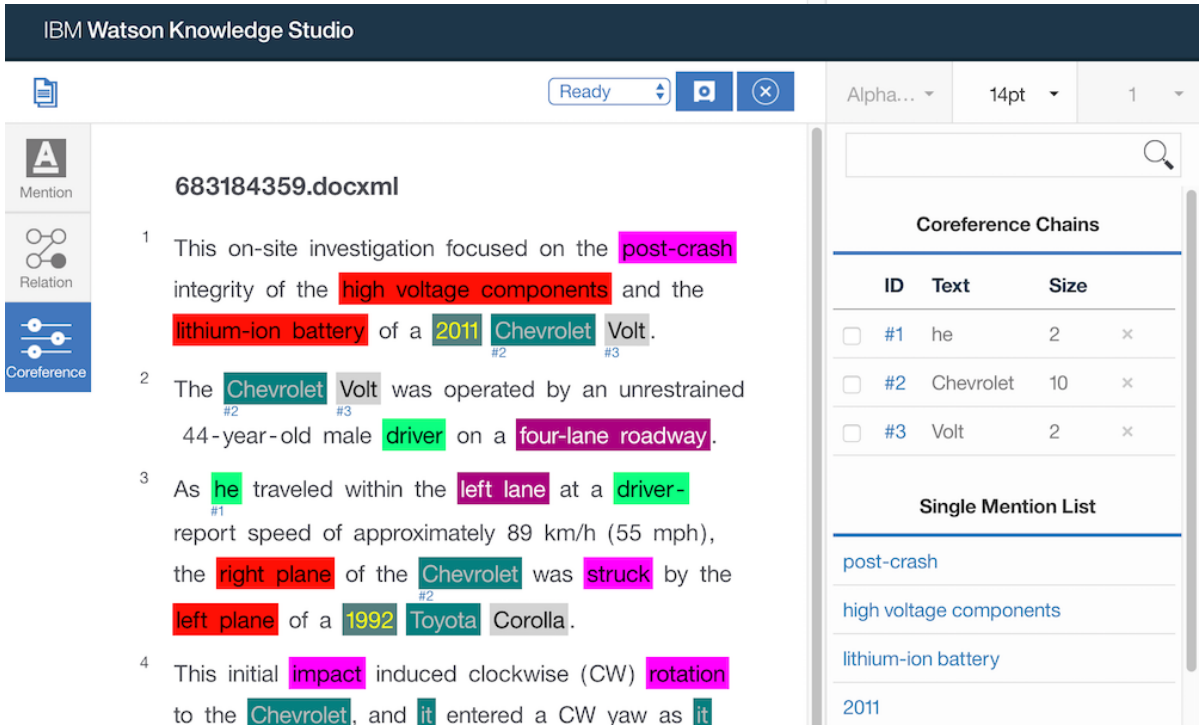
IBM Watson을 위한 데이터

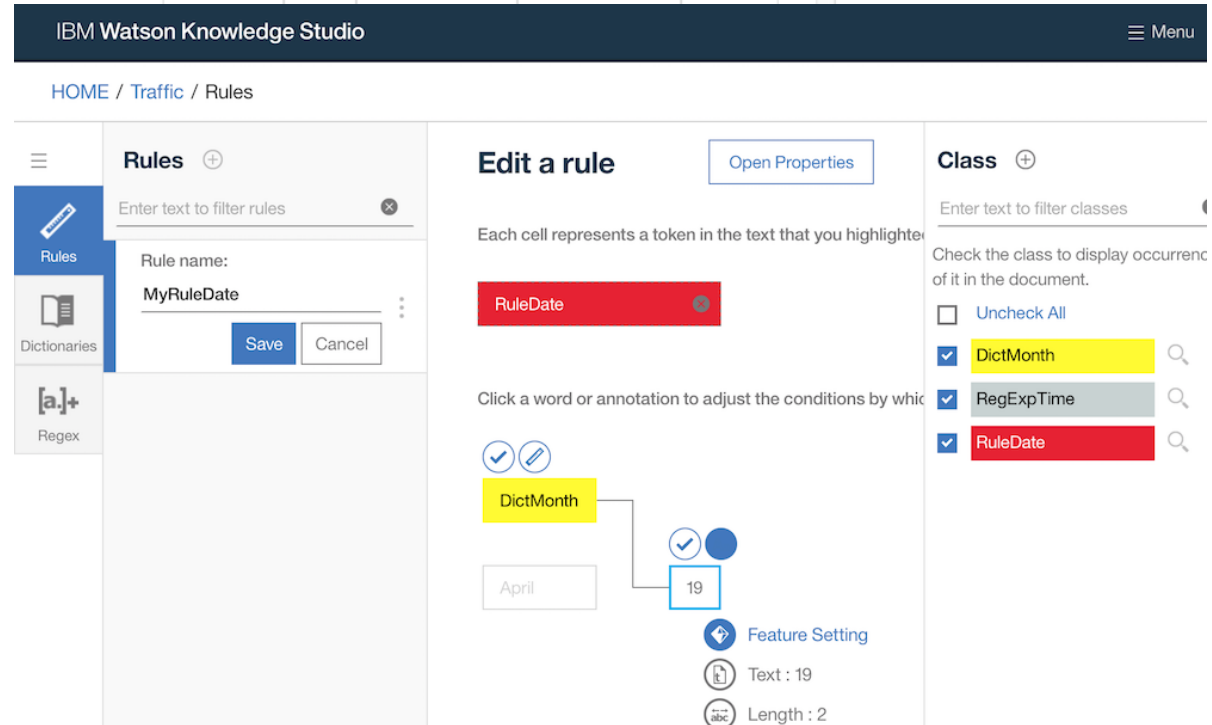
- 소프트웨어
  - Watson uses IBM's DeepQA software and the Apache UIMA (Unstructured Information Management Architecture) framework. The system was written in various languages, including Java, C++, and Prolog, and runs on the SUSE Linux Enterprise Server 11 operating system using Apache Hadoop framework to provide distributed computing.

- 데이터
  - The sources of information for Watson include encyclopedias, dictionaries, thesauri, newswire articles, and literary works. Watson also used databases, taxonomies, and ontologies. Specifically, DBPedia, WordNet, and Yago were used. The IBM team provided Watson with millions of documents, including dictionaries, encyclopedias, and other reference material that it could use to build its knowledge. Although Watson was not connected to the Internet during the game, it contained 200 million pages of structured and unstructured content consuming four terabytes of disk storage, including the full text of Wikipedia.


IBM Watson Knowledge Studio
Completed
Alpha...
14pt
1
Mention
Relation
Coreference
2004-02-001.docxml
1 Vehicle 1, a 1995 Honda Civic was traveling north on a two lane undivided roadway, negotiating a curve to the left on an upgrade.
2 V1 went over the right lane line, overcorrected and went over the left lane line into the southbound lane.
3 V1 overcorrected again and went across the northbound lane, over the right lane line.
4 The vhehicle departed the right shoulder and entered a 2-foot ditch, impacting the embankment with its front right.
Entity
Mention
Type
Subtype
Role
o ACCIDENT_OUTCOME
c CONDITION
i INCIDENT
f MANUFACTURER
m MODEL
y MODEL_YEAR
l PART_OF_CAR
p PERSON
s STRUCTURE
H VEHICLE


IBM Watson Knowledge Studio
Ready
Alpha...
14pt
1
Mention
Relation
Coreference
4
occupantOf
hasProperty
hasProperty
hasProp
occupantOf
VEHIC...
MO...
MAN...
MO...
The other vehicle was a 2000 Honda Civic that
occupantOf
PERS...
was being driven by an 18-year-old female; a 16-
PERS...
PART_OF_CAR
Relation Type
causedBy
driveUnder
hasProperty
impactPoint
locatedOn
occupantOf
outcomeOf
sufferedFrom

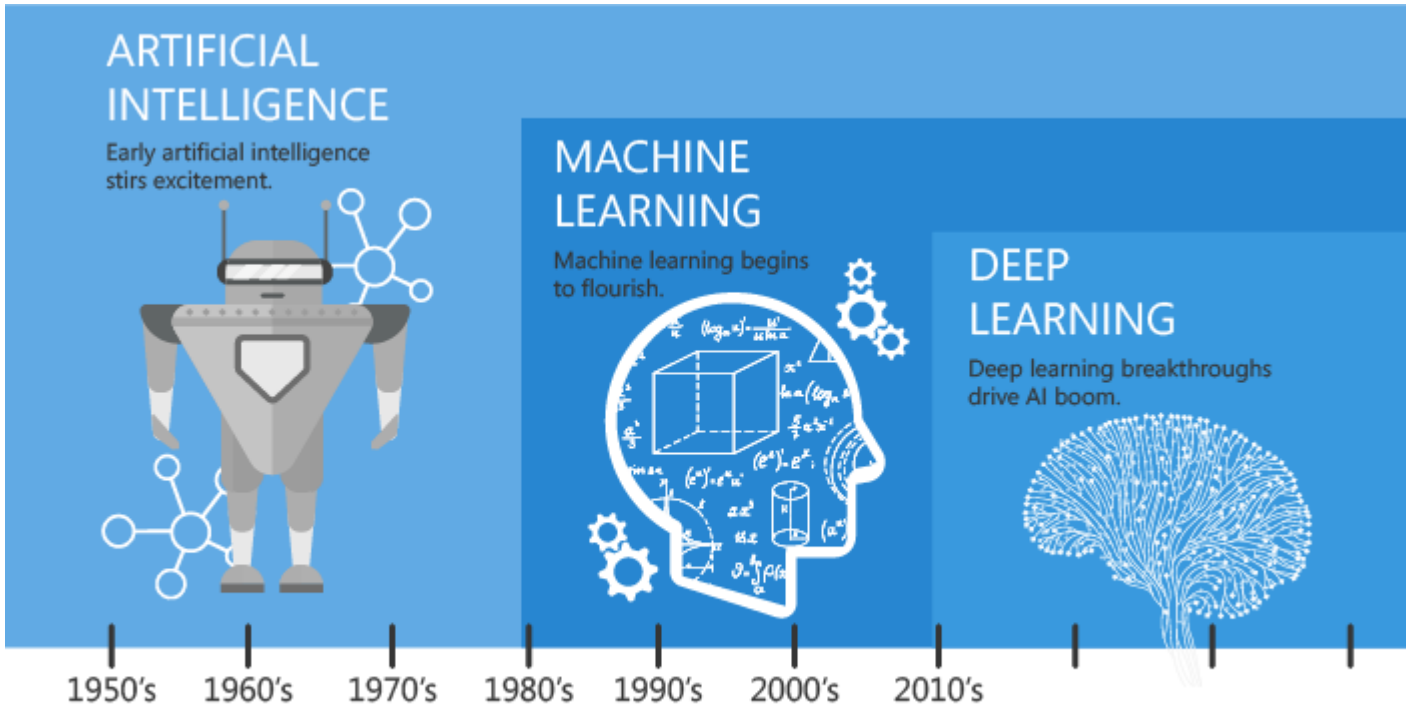

IBM Watson Knowledge Studio
Ready
Alpha...
14pt
1
Mention
Relation
Coreference
683184359.docxml
1 This on-site investigation focused on the post-crash integrity of the high voltage components and the lithium-ion battery of a 2011 Chevrolet Volt.
2 The Chevrolet Volt was operated by an unrestrained 44-year-old male driver on a four-lane roadway.
3 As he traveled within the left lane at a driver-report speed of approximately 89 km/h (55 mph), the right plane of the Chevrolet was struck by the left plane of a 1992 Toyota Corolla.
4 This initial impact induced clockwise (CW) rotation to the Chevrolet, and it entered a CW yaw as it
Coreference Chains
ID
Text
Size
#1 he 2
#2 Chevrolet 10
#3 Volt 2
Single Mention List
post-crash
high voltage components
lithium-ion battery
2011


IBM Watson Knowledge Studio
Menu
HOME / Traffic / Rules
Rules
Dictionaries
Regex
Rules
Enter text to filter rules
Rule name:
MyRuleDate
Save
Cancel
Edit a rule
Open Properties
Each cell represents a token in the text that you highlighte
RuleDate
Click a word or annotation to adjust the conditions by whic
DictMonth
April
19
Feature Setting
Text : 19
Length : 2
Class
Enter text to filter classes
Check the class to display occurrenc
of it in the document.
Uncheck All
DictMonth
RegExpTime
RuleDate

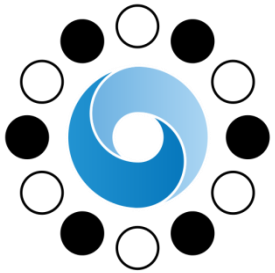
기계학습을 활용한 인공지능

- 컴퓨터가 학습할 수 있도록 하는 알고리즘과 기술을 개발하는 분야
- 컴퓨터에 명시적으로 프로그래밍을 하지 않고 통계 기술을 사용하여 학습을 수행하는 인공지능

- 알고리즘
  - Artificial neural networks
  - Support vector machines
  - Clustering
  - Bayesian networks
  - Genetic algorithms
  - etc.



Since an early flush of optimism in the 1950's, smaller subsets of artificial intelligence - first machine learning, then deep learning, a subset of machine learning - have created ever larger disruptions.



AlphaGo

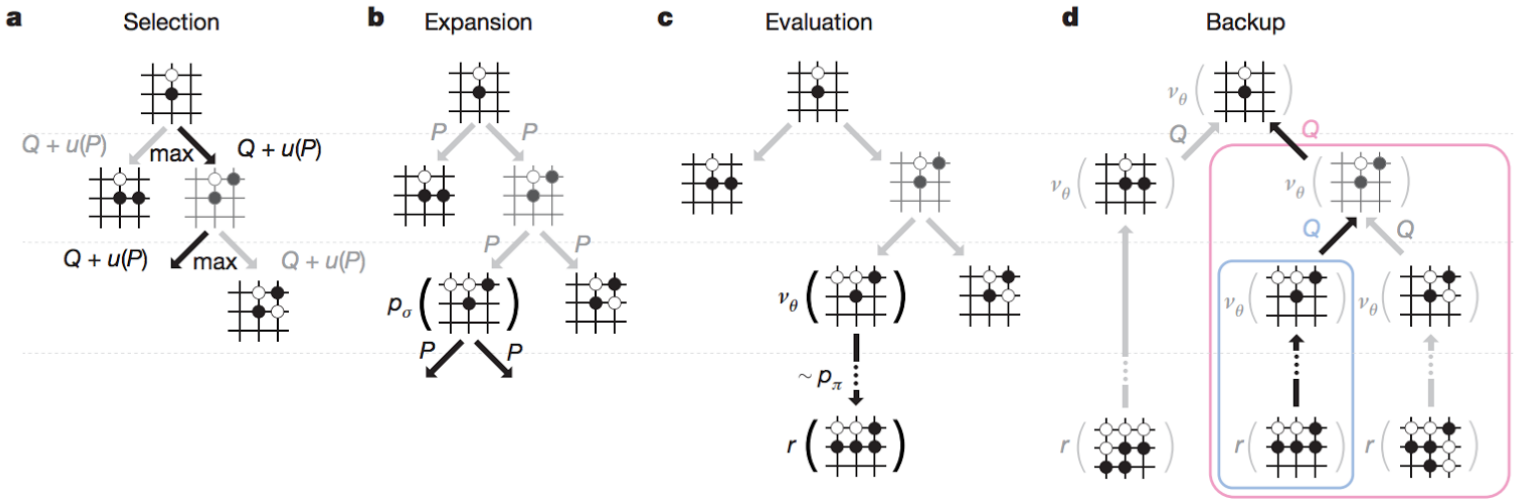
- 구글 딥마인드(Google DeepMind)가 개발한 인공지능(AI, Artificial Intelligence) 바둑 프로그램
- 핸디캡 없이 사람이 이긴 최초의 컴퓨터 바둑 프로그램

Versions	Hardware	Elo rating	Matches
AlphaGo Fan	176 GPUs, distributed	3,144	5:0 against Fan Hui
AlphaGo Lee	48 TPUs, distributed	3,739	4:1 against Lee Sedol
AlphaGo Master	4 TPUs, single machine	4,858	60:0 against professional players; Future of Go Summit
AlphaGo Zero	4 TPUs, single machine	5,185	100:0 against AlphaGo Lee 89:11 against AlphaGo Master
AlphaZero	4 TPUs, single machine	N/A	60:40 against AlphaGo Zero

구글 딥마인드의 알파고

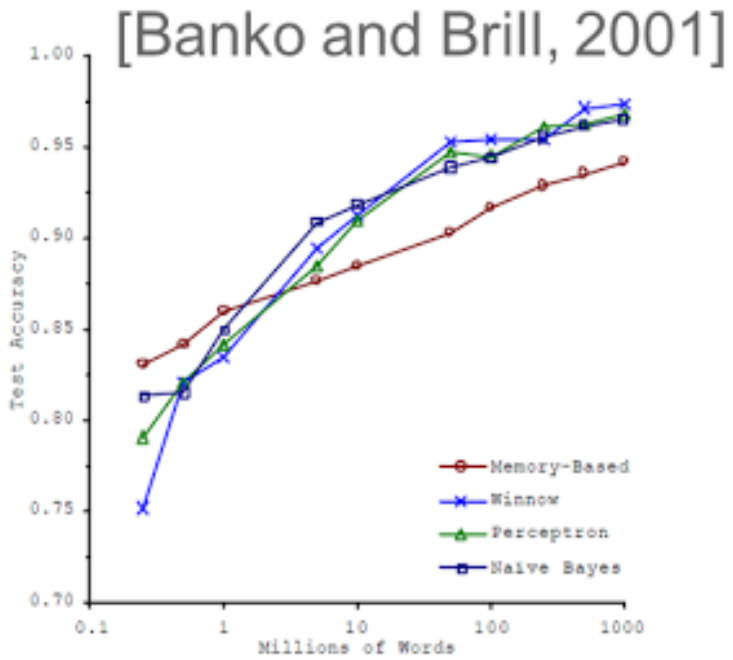
- 알고리즘
  - 훈련된 심층신경망(DNN, Deep Neural Network)이 몬테카를로 트리 탐색(MCTS, Monte Carlo Tree Search) 통해 가장 유리한 선택을 하도록 설계
  - 반복적인 자가 대국으로 성능을 개선하는 강화학습 수행

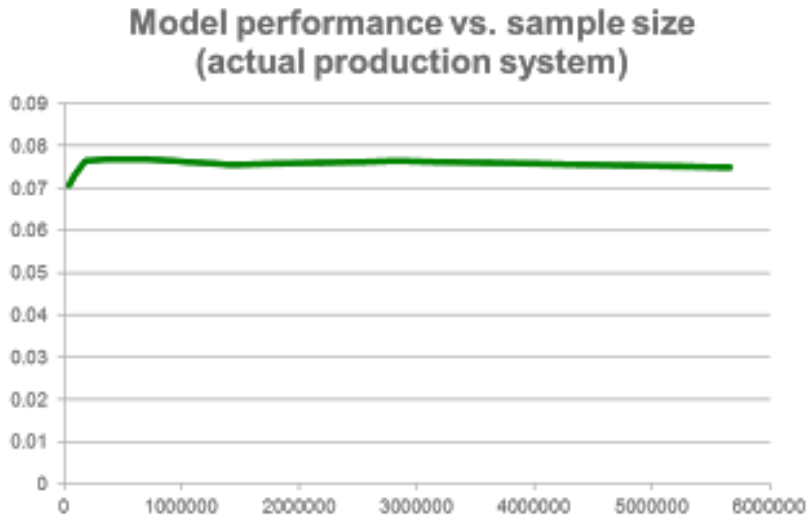
- 학습 데이터
  - KGS 바둑 서버에 등록된 16만 개의 기보
  - 3천 만개 정도의 착점 위치 정보와 패턴



데이터의 중요성

Is Data More Important Than Algorithms In AI?





> We don't have better algorithms than anyone else;
> we just have more data
>
> - Google's Zeitgeist 2011

Google's Chief Scientist Peter Norvig

제한된 데이터 환경에서의 문제점

- 인터넷에서 이미 우위를 선점하고 있는 대형 플랫폼에 의해 지배될 위험이 존재
- 경쟁과 혁신의 감소
- 신생 기업이나 중소 기업이 새로운 아이디어를 통해 문제를 해결할 수 있는 기회의 제한
- 전체적인 AI 시장 및 AI로 인해 얻을 수 있는 혜택의 감소

SQuAD

The Stanford Question Answering Dataset

- 위키피디아 문서에서 참여자들에 의해 만들어진 질의응답 관련 데이터 셋
- 500개 이상의 문서로부터 100,000개 이상의 질문-답변으로 구성


```
{    "data": [ {
                "title": "University_of_Notre_Dame",
                "paragraphs": [ {
                        "context": "Architecturally, the school has a Catholic character. Atop the Main Building's gold dome is a golden statue of the Virgin Mary. Immediately in front of the Main Building and facing it, is a copper statue of Christ with arms upraised with the legend ₩"Venite Ad Me Omnes₩". Next to the Main Building is the Basilica of the Sacred Heart. Immediately behind the basilica is the Grotto, a Marian place of prayer and reflection. It is a replica of the grotto at Lourdes, France where the Virgin Mary reputedly appeared to Saint Bernadette Soubirous in 1858. At the end of the main drive (and in a direct line that connects through 3 statues and the Gold Dome), is a simple, modern stone statue of Mary.",
                        "qas": [ {
                                "answers": [ {
                                        "answer_start": 515,
                                        "text": "Saint Bernadette Soubirous"
                                    }
                                ],
                                "question": "To whom did the Virgin Mary allegedly appear in 1858 in Lourdes France?",
                                "id": "5733be284776f41900661182"
                            },
```

Leaderboard

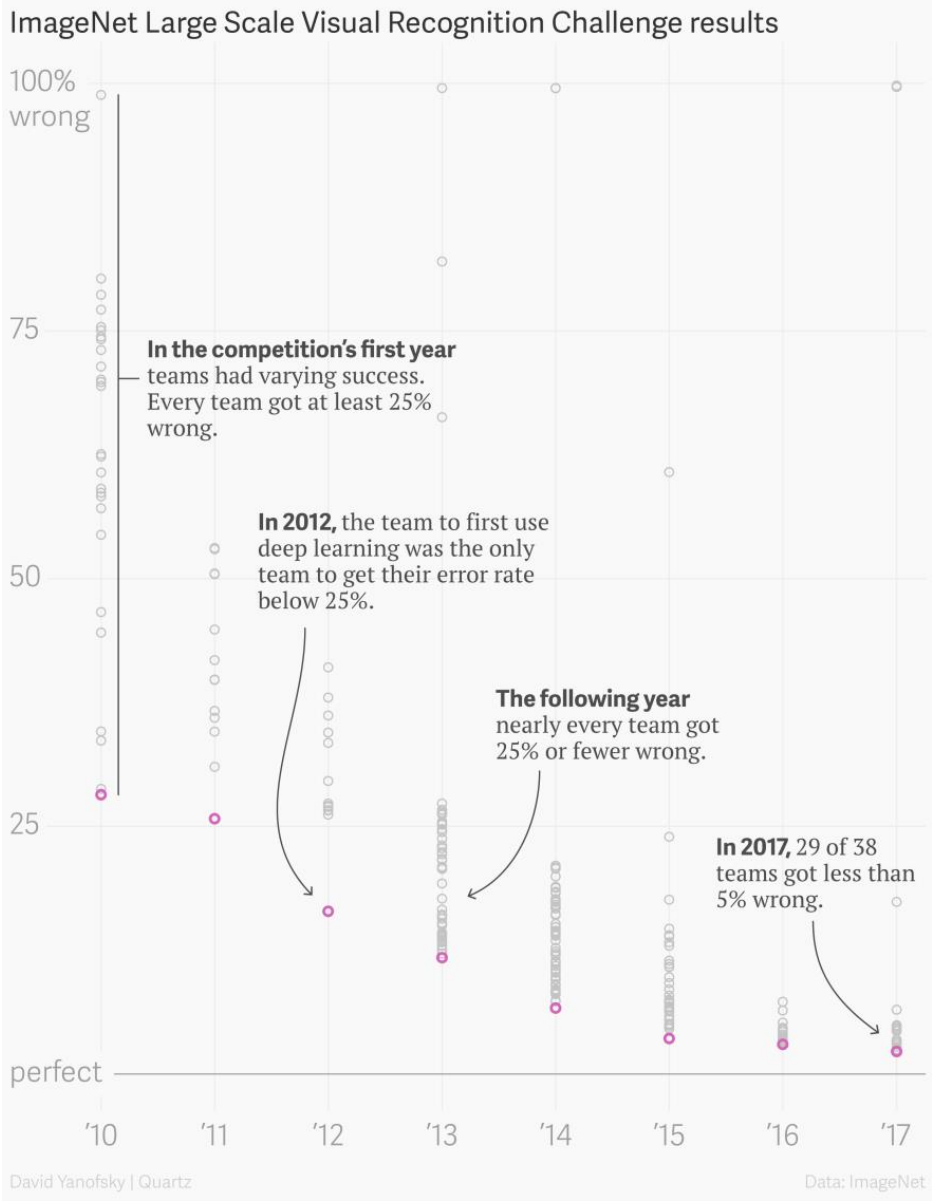
Since the release of our dataset, the community has made rapid progress! Here are the ExactMatch (EM) and F1 scores of the best models evaluated on the test set of v1.1. Will your model outperform humans on the QA task?

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Mar 19, 2018	QANet (ensemble) *Google Brain & CMU*	**83.877**	**89.737**
2 Jan 22, 2018	Hybrid AoA Reader (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	82.482	89.281
2 May 09, 2018	MARS (single model) *YUANFUDAO research NLP*	82.587	88.880
2 Feb 19, 2018	Reinforced Mnemonic Reader + A2D (ensemble model) *Microsoft Research Asia & NUDT*	82.849	88.764
2 Mar 06, 2018	QANet (ensemble) *Google Brain & CMU*	82.744	89.045
3 Jan 03, 2018	r-net+ (ensemble) *Microsoft Research Asia*	82.650	88.493
3 Jan 05, 2018	SLQA+ (ensemble) *Alibaba iDST NLP*	82.440	88.607
4 Feb 02, 2018	Reinforced Mnemonic Reader (ensemble model) *NUDT and Fudan University*	82.283	88.533
63 Oct 26, 2016	Match-LSTM with Bi-Ans-Ptr (Boundary) *Singapore Management University* https://arxiv.org/abs/1608.07905	64.744	73.743
64 Feb 19, 2017	Attentive CNN context with LSTM *NLPR, CASIA*	63.306	73.463
65 Nov 02, 2016	Fine-Grained Gating *Carnegie Mellon University* https://arxiv.org/abs/1611.01724	62.446	73.327
65 Sep 21, 2017	OTF dict+spelling (single) *University of Montreal* https://arxiv.org/abs/1706.00286	64.083	73.056
66 Sep 21, 2017	OTF spelling (single) *University of Montreal* https://arxiv.org/abs/1706.00286	62.897	72.016
67 Sep 21, 2017	OTF spelling+lemma (single) *University of Montreal* https://arxiv.org/abs/1706.00286	62.604	71.968
68 Sep 28, 2016	Dynamic Chunk Reader *IBM* https://arxiv.org/abs/1610.09996	62.499	70.956
69 Aug 27, 2016	Match-LSTM with Ans-Ptr (Boundary) *Singapore Management University* https://arxiv.org/abs/1608.07905	60.474	70.695
70 Jan 05, 2018	PivRet (single model) *anonymous*	58.764	69.276
71 Aug 27, 2016	Match-LSTM with Ans-Ptr (Sentence) *Singapore Management University* https://arxiv.org/abs/1608.07905	54.505	67.748

IMAGENET

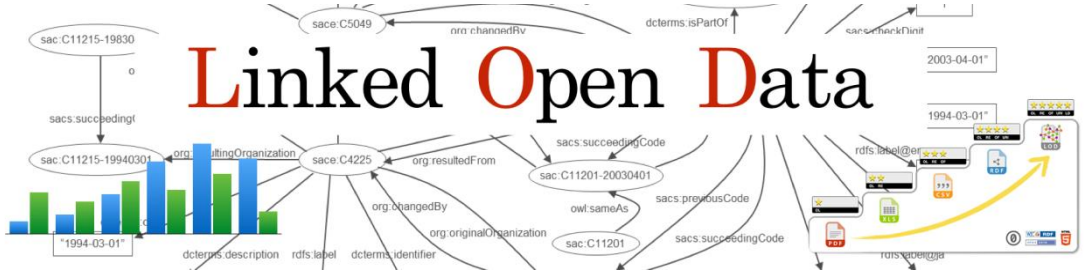
- ImageNet
  - 시각적 개체 인식 소프트웨어 연구를 위한 대규모 이미지 데이터베이스
  - 2천개 이상의 분류로 1400만 개 이상의 이미지를 포함

- ILSVRC (ImageNet Large Scale Visual Recognition Challenge)
  - 2010년 정확도 71.8%
  - 2017년 정확도 97.3%

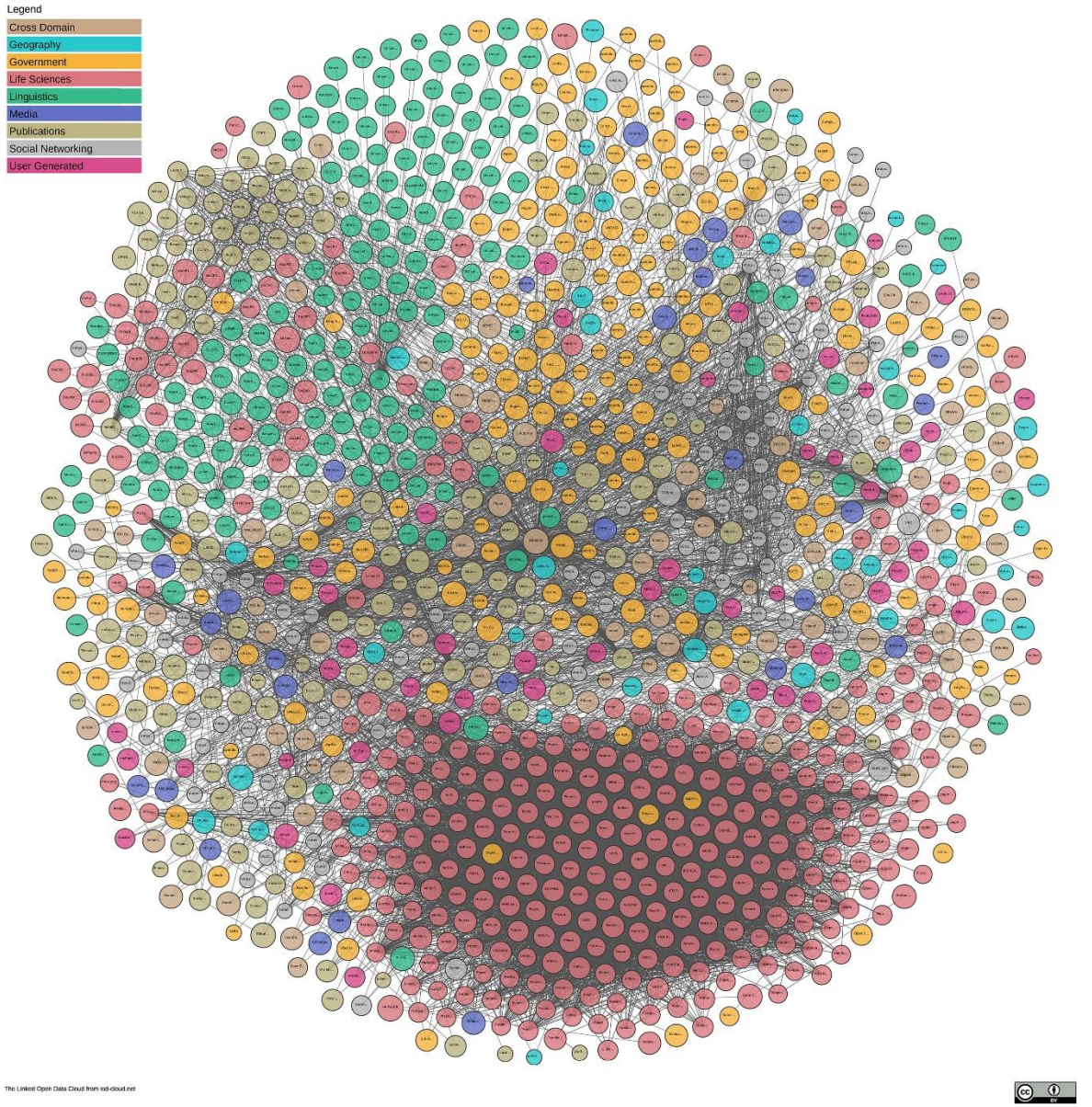


인공지능을 위한 데이터 개방의 중요성

- 인공지능을 위한 데이터의 접근성을 높임으로써,
- 중소기업의 경쟁력을 올리고
- 공정한 AI 시장에서의 경쟁 체계 구축



일자	데이터 셋
2018-04-30	1,184
2017-08-22	1,163
2017-02-20	1,139
2017-01-26	1,146
2014-08-30	570
2011-09-19	295
2010-09-22	203
2009-07-14	95
2009-03-27	93
2009-03-05	89
2008-09-18	45
2008-03-31	34
2008-02-28	32
2007-11-10	28
2007-11-07	28
2007-10-08	25
2007-05-01	12



기계학습을 위한 개방 데이터 셋

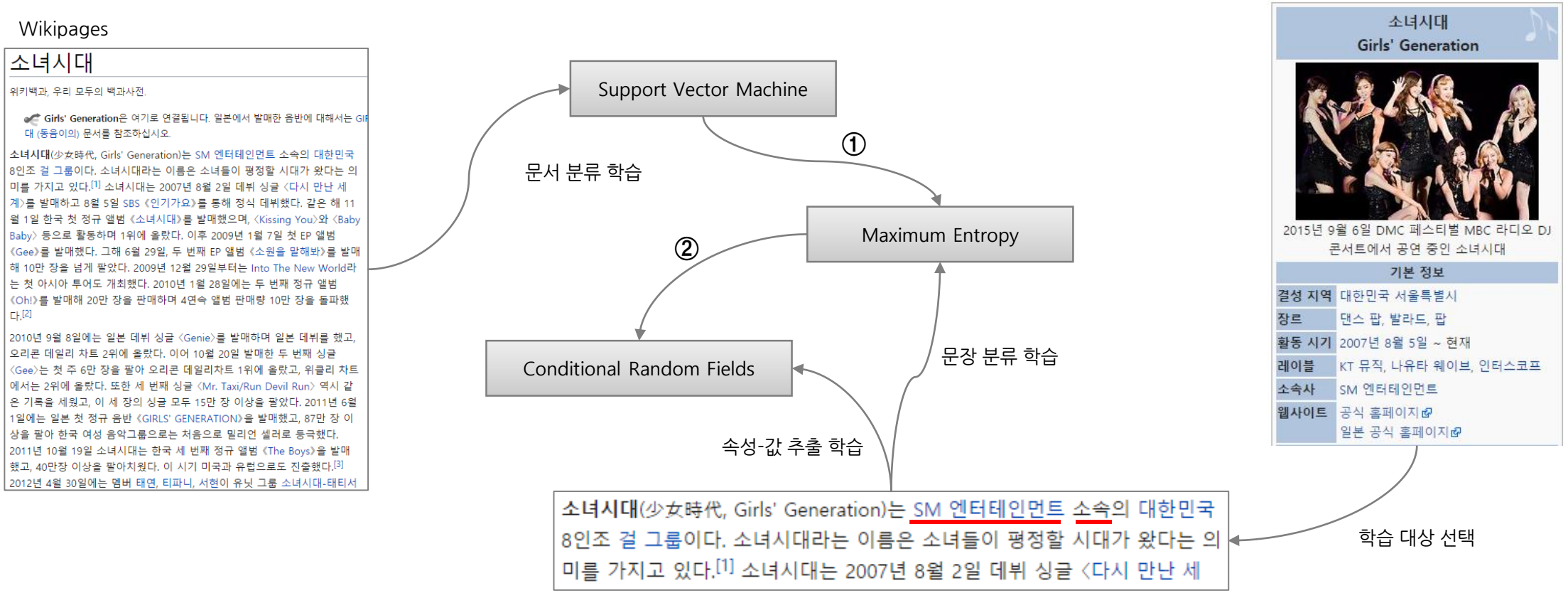
- List of datasets for machine learning research
  - https://en.wikipedia.org/wiki/List_of_datasets_for_machine_learning_research
- Open Data for Deep Learning & Machine Learning
  - https://deeplearning4j.org/opendata
- Fueling the Gold Rush: The Greatest Public Datasets for AI
  - https://medium.com/startup-grind/fueling-the-ai-gold-rush-7ae438505bc2
- UCI Machine Learning Repository
  - http://archive.ics.uci.edu/ml/datasets.html

기계학습을 위한 개방 데이터 셋

- OpenML - https://www.openml.org/

우리의 경험 #1

- KEF (Knowledge Extraction Framework)
  - 지식베이스 구축을 위해 비정형 데이터로부터 지식을 추출하기 위한 기계학습 기반의 플랫폼



분석 정보

아이디
a

이름
김재헌

분석 요청 건수
266

분석 완료 건수
123

분석 미완료 건수
143

분석하기

분석요청 목록

- 로그인을 이용해서 기록을 남겨보세요.

문서 제목 문서 제목은 트리플에 주어(S)로 사용

문재인

위키피디아 URL 문서를 위키피디아 URL 을 통해서 가져 올 수 있습니다.

Wikipedia URL을 넣어주세요 위키 본문 가져오기

문서 내용

문재인(文在寅, 1953년 1월 24일 ~)은 대한민국의 제19대 대통령이다. 본관은 남평(南平)이다.[2]

경희대학교 재학 시절 학생운동을 이끌며 박정희 유신독재에 항거하다가 1975년 서대문구치소에 투옥됐고 대학에서 제적당했다. 출소 후에는 바로 군에 강제징집 됐다. 특전사 제대 후 복학해 다시 학생운동을 이끌며 전두환 군부독재에 항거하다가 1980년 청량리구치소에 투옥됐으나, 옥중에서 사법시험에 합격하면서 극적으로 풀려났다. 1982년 사법연수원을 최우수 성적[3]으로 수료했으나 학

143

재에 항거하다가 1980년 청량리구치소에 투옥됐으나, 옥중에서 사법시험에 합격하면서 극적으로 풀려났다. 1982년 사법연수원을 최우수 성적[3]으로 수료했으나 학생운동 전력으로 판사 임용이 거부되자 부산으로 내려가 노무현 변호사와 합동법률사무소를 운영하며 인권변호사로 활동했다.

분석 대상의 주제 정보

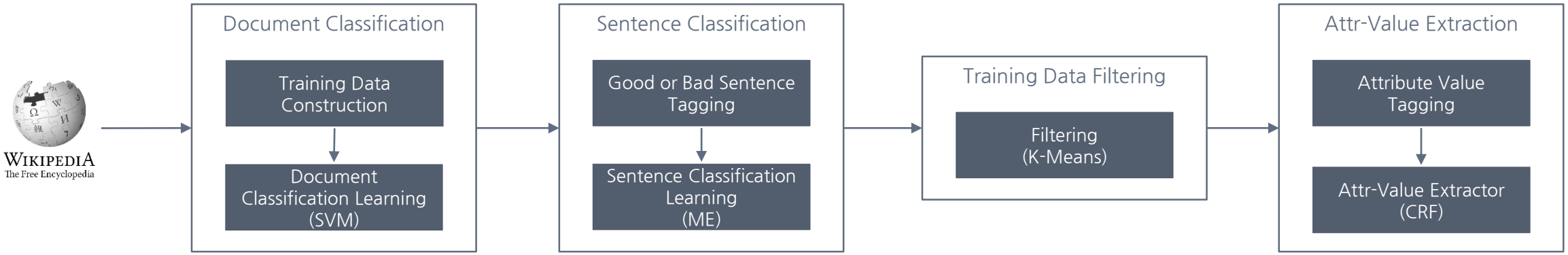
주제	인물 정보
점수	0.321299

문서로부터 추출한 속성 - 값

속성	순위	값	점수	근거문장
원어이름	1순위	文在寅	0.780807	문장 확인하기
출생일	1순위	1953년 1월 24일	0.983353	문장 확인하기
출생지	1순위	대한민국	0.579422	문장 확인하기

우리의 경험 #1

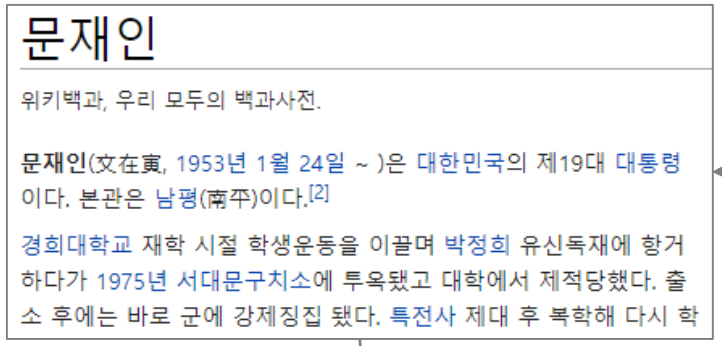
- 학습 데이터
  - Wikipedia의 Infobox를 이용하여 학습 수행
  - 120,000 개의 Wikipedia 문서를 학습하여 200개 이상의 분류와 3,000개 이상 다양한 유형의 지식추출



우리의 경험 #1

- 학습 데이터 생성

원문데이터



학습 대상

이것이 문제!!!

문장 태깅

문장	구분
문재인(文在寅, 1953년 1월 24일 ~)은 대한민국의 제19대 대통령이다.	좋은 문장
본관은 남평(南平)이다.	나쁜 문장
경희대학교 재학 시절 학생운동을 이끌며 박정희 유신독재에 항거하다가 1975년 서대문구치소에 투옥됐고 대학에서 제적당했다.	나쁜 문장
출소 후에는 바로 군에 강제징집 됐다.	나쁜 문장

단어 태깅

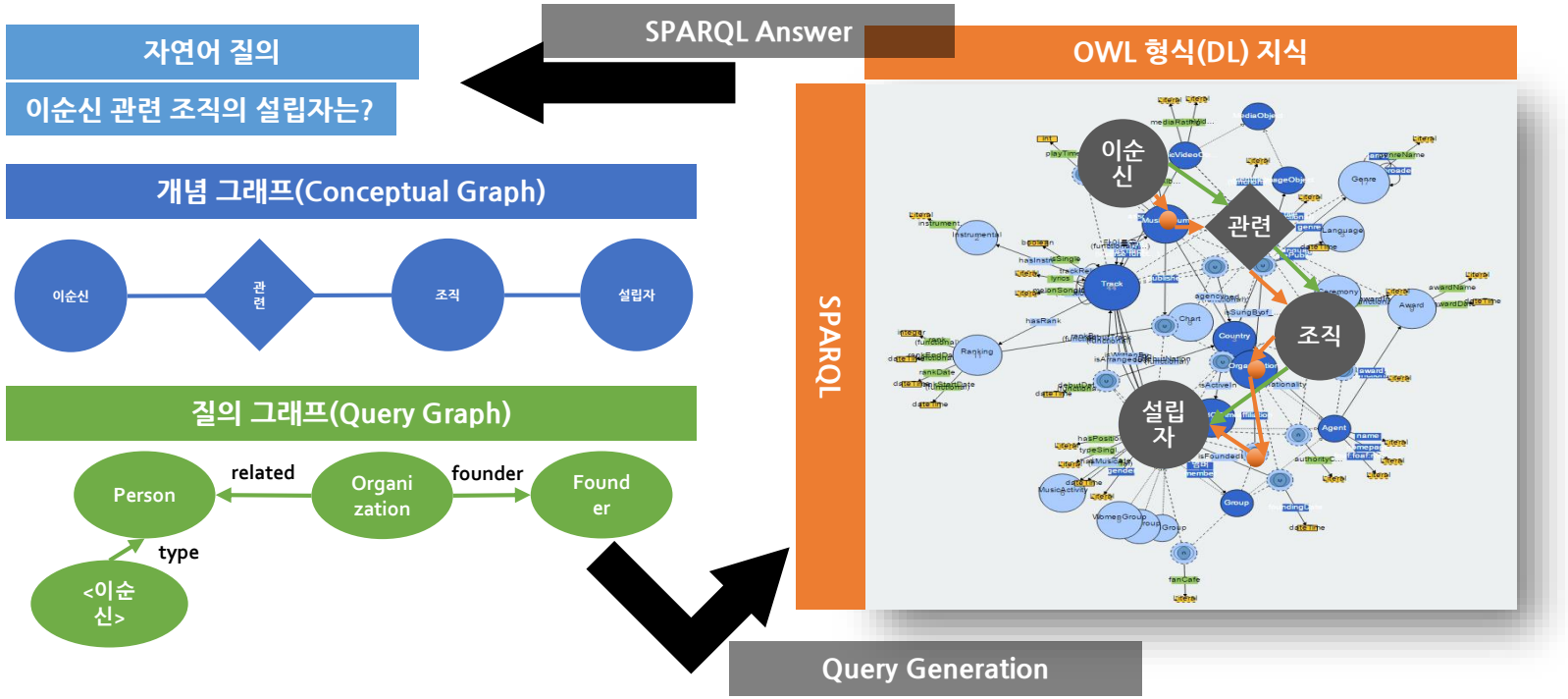
문재인	은	대한민국	의	제19대	대통령	이다.
Pre	Pre	V	Post	Post	N	N

우리의 경험 #1

- 경험을 통한 교훈
  - 좋은 데이터를 확보하는 것이 쉽지 않다.
  - 좋은 데이터가 있다면 더 좋은 성능을 발휘할 수 있다.
  - 일반 문서를 위한 데이터는 어떻게?

우리의 경험 #2

- ALVIS
  - 지식베이스를 기반으로 한 지능형 질의응답 플랫폼
  - 지식베이스를 기반으로 사용자의 자연어 질의를 해석하고 그에 적합한 응답을 지식베이스로부터 찾아 반환하는 질의응답 플랫폼



우리의 경험 #2

- 한국사 관련 질의응답 플랫폼 구현을 위한 한국사 지식베이스
  - 어디서 데이터를 얻을 것인가?
    - 국사편찬위원회의 한국사 LOD
    - 한국학중앙연구원의 한국민족문화대백과사전
    - 이뮤지엄
  - 지식베이스 구축 시 어려웠던 점
    - 간단한 수집 방법의 부재(LOD, Open API 등)
    - 정규화 되어 있지 않은 값들
    - 통일되어 있지 않은 데이터 형식
    - 동일 개체에 대한 서로 다른 아이디
    - 의미 없이 구성된 데이터 구조
    - 등등

우리의 경험 #2

- 경험을 통한 교훈
  - 좋은 데이터 구축은 인간의 노동을 요한다.
  - 웹 페이지 등과 같이 가공된 데이터보다는 날 것 그대로의 데이터(raw data)가 중요하다.
  - 데이터 잘 만들면 부족할지 언정 공유 해야겠다.

더 나은 인공지능 시대를 위하여

- Better data is always better.

- 인공지능에서 알고리즘의 성능을 높이기 위해 대량의 고품질 데이터가 중요

- 가능한 데이터를 공개적으로 사용할 수 있고 기계가 읽을 수 있는 형식의 데이터를 만드는 것이 중요

- 정부의 역할
  - 정부가 보유 및 생산(직접적이든 간접적이든)한 데이터는 엄청난 자산임을 인식
  - 인공지능을 위해 데이터를 원천 데이터 형식으로 수집 및 제공
  - 사용자가 보다 쉽게 데이터를 발견하고 이용할 수 있도록 다양한 형태의 노력이 필요
  - 데이터의 유통이 일어날 수록 개인정보에 대한 철저한 관리와 규제 필요

References

- https://www.quora.com/In-machine-learning-is-more-data-always-better-than-better-algorithms
- https://rajpurkar.github.io/SQuAD-explorer/
- https://www.techuk.org/insights/opinions/item/10708-guest-blog-peter-wells-odi-access-to-data-is-key-to-a-competitive-ai-market
- https://www.forbes.com/sites/quora/2017/01/26/is-data-more-important-than-algorithms-in-ai/#353c87f842c1
- https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world/
- https://www.forbes.com/sites/scottcleland/2011/10/03/googles-infringenovation-secrets/#67b70b9b30a6
- https://theodi.org/article/using-artificial-intelligence-and-open-data-for-innovation-and-accountability/
- https://www.squiz.net/learn/blog/why-data-is-so-important-when-it-comes-to-ai
- https://www.govloop.com/refining-value-government-data-machine-learning/
- https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html
- https://searchenterpriseai.techtarget.com/definition/Turing-test
- http://heidloff.net/article/understanding-natural-language-text
- https://en.wikipedia.org/wiki/Open_science
- http://cckorea.org/xe/news/632741

Q&A

Myungjin Lee

LiST, Linked Data and Semantic Web Technology
Ph.D. / CTO

e-Mail : mjlee@LiSTInc.kr
Twitter : http://twitter.com/MyungjinLee
Facebook : http://www.facebook.com/mjinlee
SlideShare : http://www.slideshare.net/onlyjiny/