Information Extraction with Wikipedia

馬偉雲
中研院資訊所助研究員
中研院詞庫小組主持人
2019/10/5

Outline

- Information Extraction
  - Distant Supervision for Relation Extraction
  - Hierarchy of Wikipedia Categories

Outline

- Information Extraction
  - Distant Supervision for Relation Extraction
  - Hierarchy of Wikipedia Categories

Distant Supervision for Relation Extraction

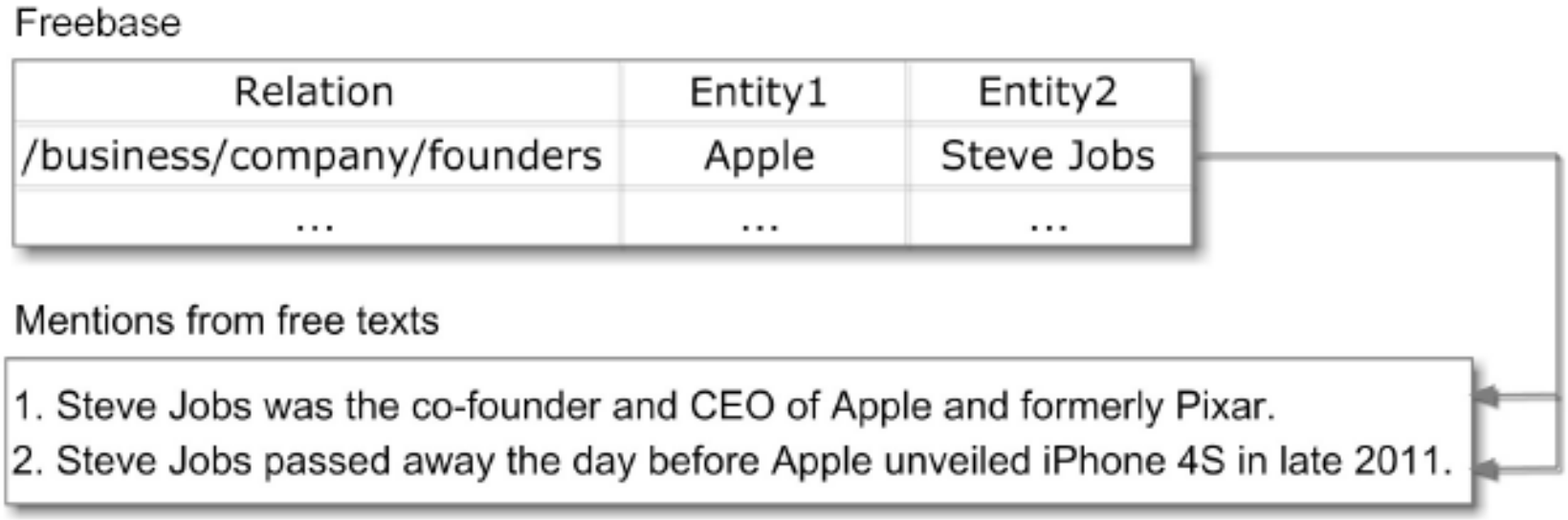
- Mintz et al "Distant supervision for relation extraction without labeled data" ACL 2009
- Zeng et al "Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks" EMNLP 2015
- Ji et al "Distant Supervision for Relation Extraction with Sentence-Level Attention and Entity Descriptions" AAAI 2017

Distant Supervision for Relation Extraction

- Mintz et al "Distant supervision for relation extraction without labeled data" ACL 2009
- Zeng et al "Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks" EMNLP 2015
- Ji et al "Distant Supervision for Relation Extraction with Sentence-Level Attention and Entity Descriptions" AAAI 2017

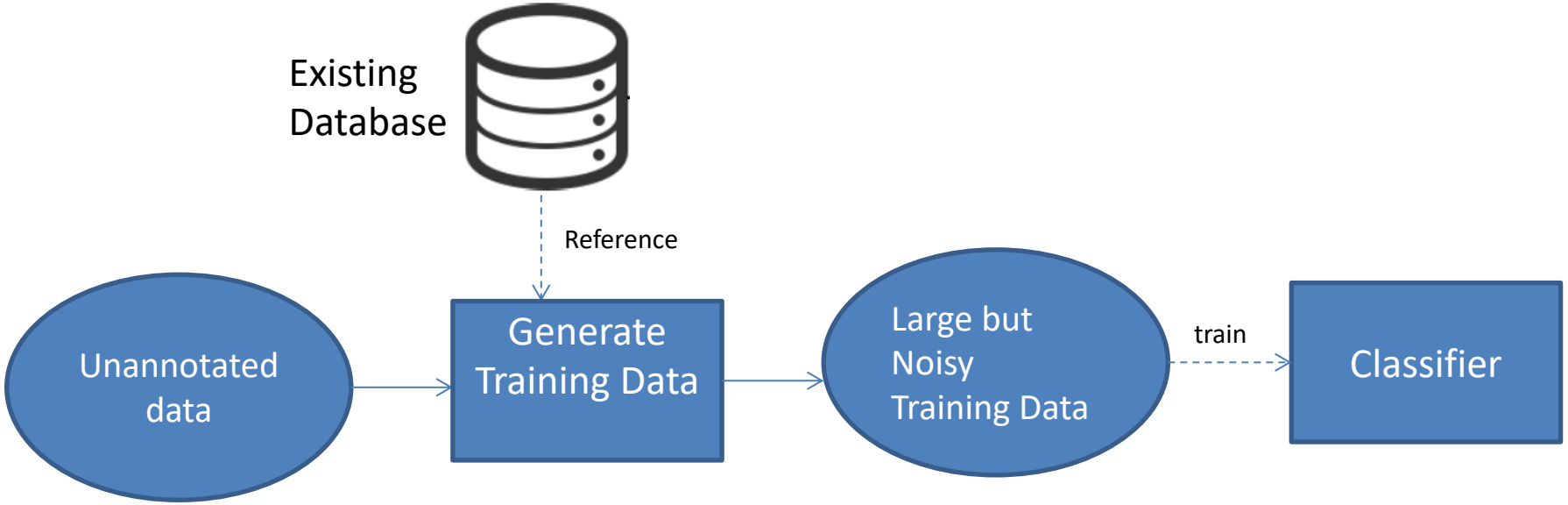
Distant supervision for relation extraction without labeled data

- Assumption
  - If two entities have a relationship in a known knowledge base, then all sentences that mention these two entities will express that relationship.



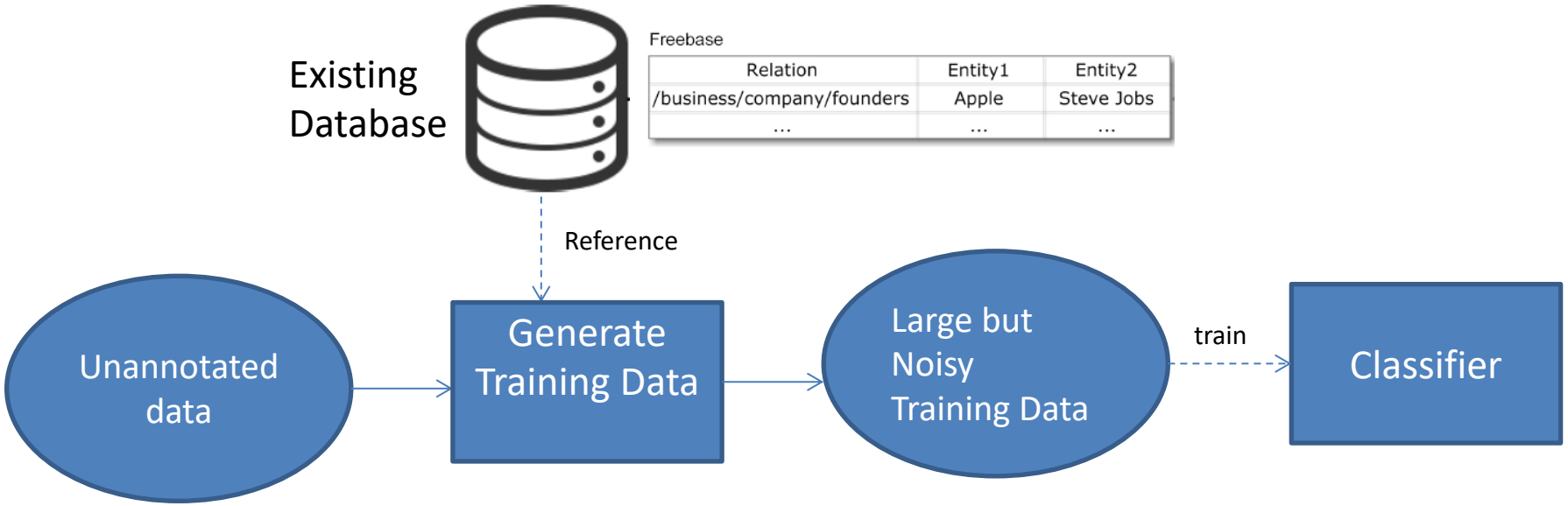
Distant supervision for relation extraction without labeled data

- Obtain Large but Noisy Training Data
  - All sentences that contain the two entities are selected as training instances



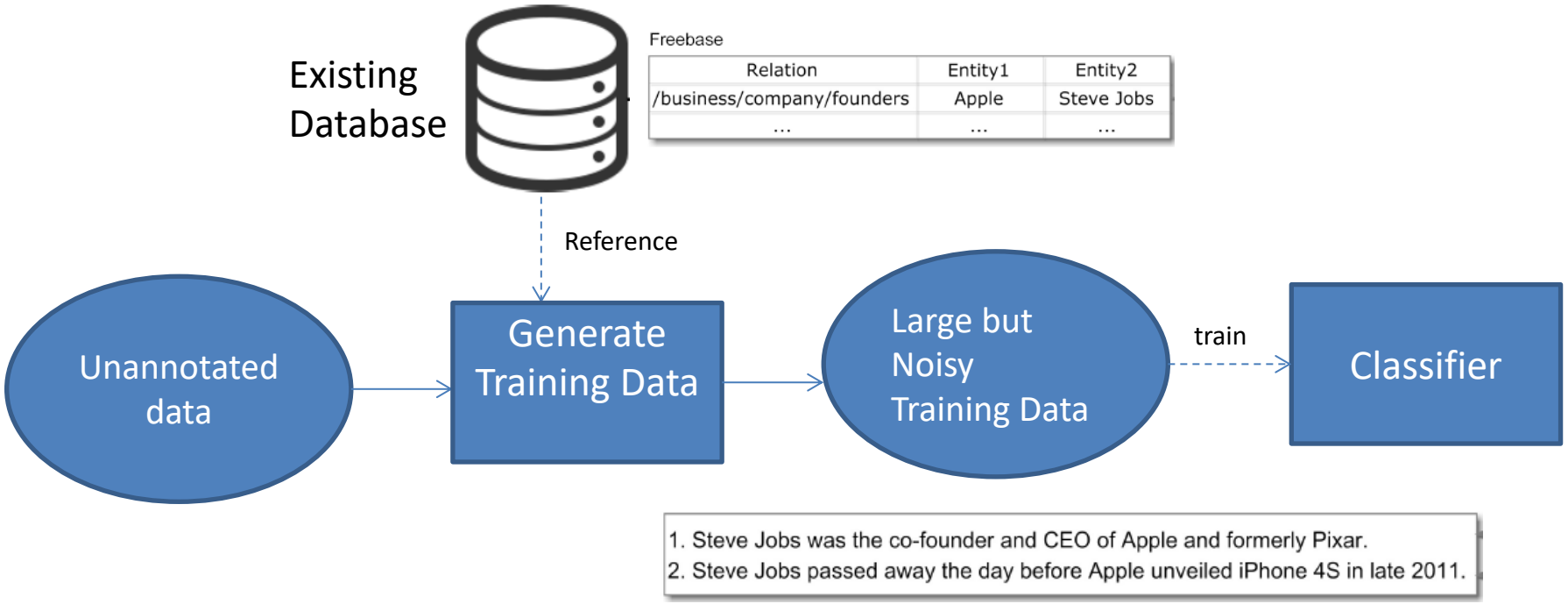
Distant supervision for relation extraction without labeled data

- Obtain Large but Noisy Training Data
  - All sentences that contain the two entities are selected as training instances



Distant supervision for relation extraction without labeled data

- Obtain Large but Noisy Training Data
  - All sentences that contain the two entities are selected as training instances



WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction

Help
About Wikipedia
Community portal
Recent changes
Contact page

Article | Talk | Read | View

Stephen Curry

From Wikipedia, the free encyclopedia

This article is about the American basketball player. For other pe
(disambiguation).

Wardell Stephen "Steph" Curry II (/ˈstɛfən/ *STEF-ən*; born March 14, 1988) is an American professional basketball player for the Golden State Warriors of the National Basketball Association (NBA). A six-time NBA All-Star, he has been named the NBA Most Valuable Player (MVP) twice and won three NBA championships with the Warriors. Many players and analysts have called him the greatest shooter in NBA history.[1] He is credited with revolutionizing the game of basketball by inspiring teams to regularly employ the three-point shot as part of their winning strategy.[2][3][4]

Stephen Curry

Curry with the Warriors in 2017

No. 30 – Golden State Warriors	
Position	Point guard
League	NBA
Personal information	
Born	March 14, 1988 (age 31) Akron, Ohio
Nationality	American
Listed height	6 ft 3 in (1.91 m)
Listed weight	190 lb (86 kg)
Career information	
High school	Charlotte Christian (Charlotte, North Carolina)
College	Davidson (2006–2009)
NBA draft	2009 / Round: 1 / Pick: 7th overall
	Selected by the Golden State Warriors
Playing career	2009–present
Career history	
2009–present	Golden State Warriors

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction

Help
About Wikipedia
Community portal
Recent changes
Contact page

Article | Talk | Read | View

Stephen Curry

From Wikipedia, the free encyclopedia

This article is about the American basketball player. For other pe
(disambiguation).

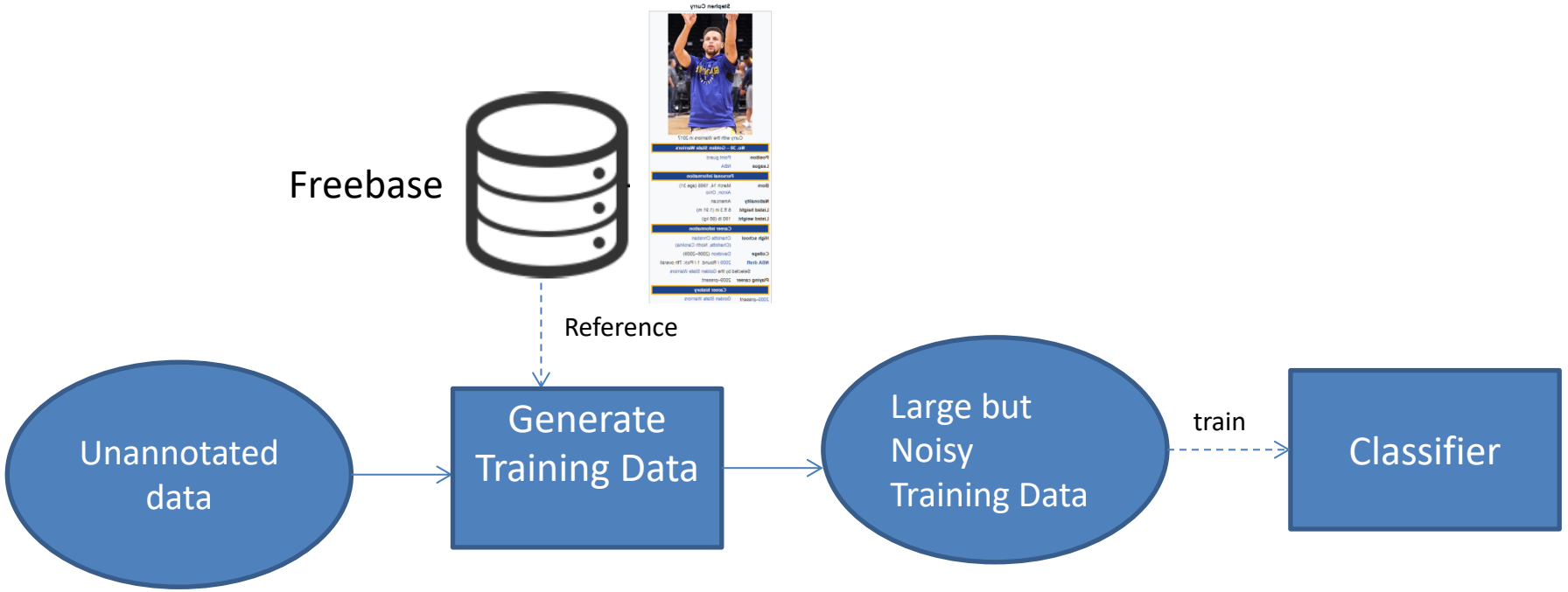
Wardell Stephen "Steph" Curry II (/ˈstɛfən/ *STEF-ən*; born March 14, 1988) is an American professional basketball player for the Golden State Warriors of the National Basketball Association (NBA). A six-time NBA All-Star, he has been named the NBA Most Valuable Player (MVP) twice and won three NBA championships with the Warriors. Many players and analysts have called him the greatest shooter in NBA history.[1] He is credited with revolutionizing the game of basketball by inspiring teams to regularly employ the three-point shot as part of their winning strategy.[2][3][4]

Stephen Curry

Curry with the Warriors in 2017

No. 30 – Golden State Warriors	
Position	Point guard
League	NBA
Personal information	
Born	March 14, 1988 (age 31) Akron, Ohio
Nationality	American
Listed height	6 ft 3 in (1.91 m)
Listed weight	190 lb (86 kg)
Career information	
High school	Charlotte Christian (Charlotte, North Carolina)
College	Davidson (2006–2009)
NBA draft	2009 / Round: 1 / Pick: 7th overall
	Selected by the Golden State Warriors
Playing career	2009–present
Career history	
2009–present	Golden State Warriors

In this paper



<Entity1, Career History, Entity2>

Entity1 is an American professional basketball player for the Entity2 of the National Basketball Association (NBA).

Freebase

- 102 relations, 940k entities,1.8M instances.

Relation name	Size	Example
/people/person/nationality	281,107	John Dugard, South Africa
/location/location/contains	253,223	Belgium, Nijlen
/people/person/profession	208,888	Dusa McDuff, Mathematician
/people/person/place_of_birth	105,799	Edwin Hubble, Marshfield
/dining/restaurant/cuisine	86,213	MacAyo's Mexican Kitchen, Mexican
/business/business_chain/location	66,529	Apple Inc., Apple Inc., South Park, NC
/biology/organism_classification_rank	42,806	Scorpaeniformes, Order
/film/film/genre	40,658	Where the Sidewalk Ends, Film noir
/film/film/language	31,103	Enter the Phoenix, Cantonese
/biology/organism_higher_classification	30,052	Calopteryx, Calopterygidae
/film/film/country	27,217	Turtle Diary, United States
/film/writer/film	23,856	Irving Shulman, Rebel Without a Cause
/film/director/film	23,539	Michael Mann, Collateral
/film/producer/film	22,079	Diane Eskenazi, Aladdin
/people/deceased_person/place_of_death	18,814	John W. Kern, Asheville
/music/artist/origin	18,619	The Octopus Project, Austin
/people/person/religion	17,582	Joseph Chartrand, Catholicism
/book/author/works_written	17,278	Paul Auster, Travels in the Scriptorium
/soccer/football_position/players	17,244	Midfielder, Chen Tao
/people/deceased_person/cause_of_death	16,709	Richard Daintree, Tuberculosis
/book/book/genre	16,431	Pony Soldiers, Science fiction
/film/film/music	14,070	Stavisky, Stephen Sondheim
/business/company/industry	13,805	ATS Medical, Health care

Training

- Generate training data
  - Find the sentence that contains two entities.
    - This sentence tends to express the relation.
    - Entities are found by a named entity tagger.

< Stephen Curry, Career History, Golden State Warriors>
Wardell Stephen "Steph" Curry II is an American professional basketball player for the Golden State Warriors of the National Basketball Association (NBA).

<Entity1, Career History, Entity2>
Entity1 is an American professional basketball player for the Entity2 of the National Basketball Association (NBA).

- Train classifier
  - Features will be explained in the next slides.

Features for Train classifier

- Lexical features
  - specific words(POSs) between and surrounding the two entities in the sentence.
- Syntactic features
  - dependency path

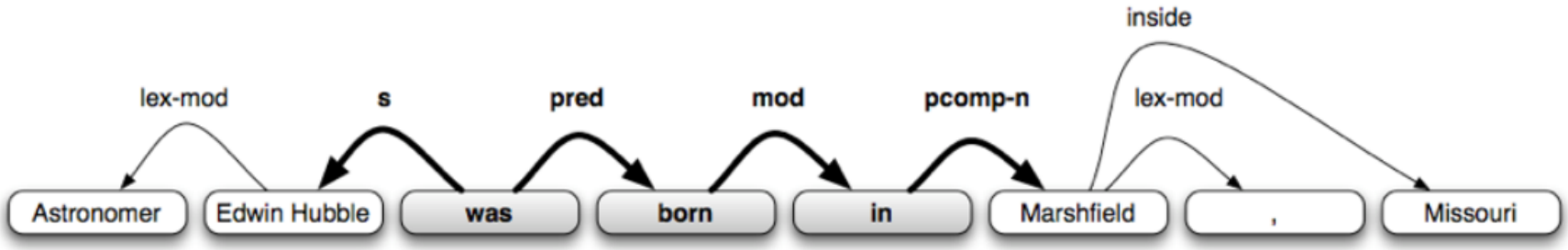


Figure 1: Dependency parse with dependency path from 'Edwin Hubble' to 'Marshfield' highlighted in boldface.

Testing

- Find the sentence that contains two entities.
  - This sentence tends to express the relation.
  - Entities are found by a named entity tagger.
- Using trained classifier, we can know these entities have a relation

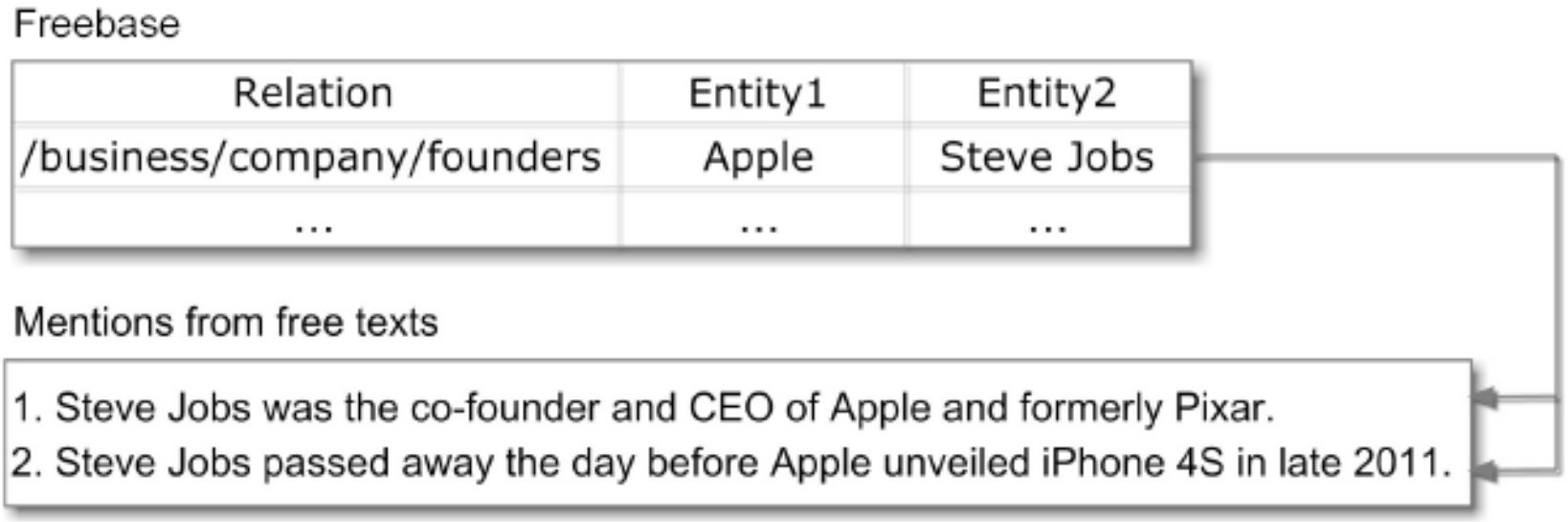
Result

- Manual Evaluation on Wikipedia articles

Relation name	100 instances			1000 instances		
	Syn	Lex	Both	Syn	Lex	Both
/film/director/film	**0.49**	0.43	0.44	**0.49**	0.41	0.46
/film/writer/film	**0.70**	0.60	0.65	**0.71**	0.61	0.69
/geography/river/basin_countries	0.65	0.64	**0.67**	**0.73**	0.71	0.64
/location/country/administrative_divisions	0.68	0.59	**0.70**	**0.72**	0.68	**0.72**
/location/location/contains	0.81	**0.89**	0.84	**0.85**	0.83	0.84
/location/us_county/county_seat	0.51	0.51	**0.53**	0.47	**0.57**	0.42
/music/artist/origin	0.64	0.66	**0.71**	0.61	**0.63**	0.60
/people/deceased_person/place_of_death	0.80	0.79	**0.81**	0.80	**0.81**	0.78
/people/person/nationality	0.61	0.70	**0.72**	0.56	0.61	**0.63**
/people/person/place_of_birth	**0.78**	0.77	**0.78**	0.88	0.85	**0.91**
Average	0.67	0.66	**0.69**	**0.68**	0.67	0.67

Problems of Mintz et al ACL 2009

- Feature representation
  - Traditional features (POS, NER, Parsing…) lead to error propagation
- Wrong label problem



Distant Supervision for Relation Extraction

- Mintz et al "Distant supervision for relation extraction without labeled data" ACL 2009
- Zeng et al "Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks" EMNLP 2015
- Ji et al "Distant Supervision for Relation Extraction with Sentence-Level Attention and Entity Descriptions" AAAI 2017

Zeng et al's Solutions

- Feature representation
  - Traditional features (POS, NER, Parsing…) lead to error propagation
  - Zeng et al's Solution: PCNN
- Wrong label problem
  - Zeng et al's Solution: Multi-instance learning with PCNN

Automatically Learn Features

- Piecewise Convolutional Neural Networks (PCNNs) is proposed to learn features without complicated NLP preprocessing
- An extension of CNN in COLING 2014

CNN in COLING 2014

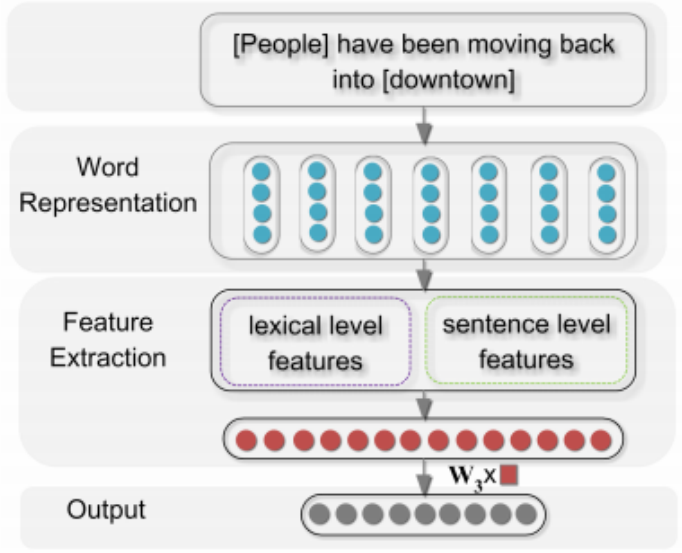


Figure 1: Architecture of the neural network used for relation classification.

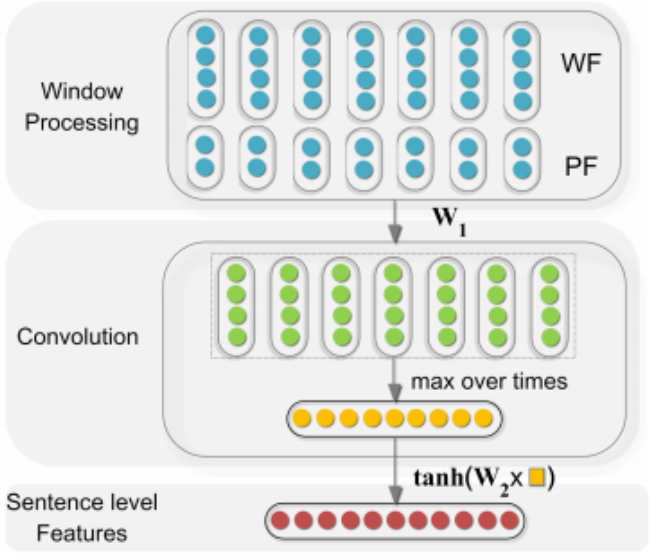
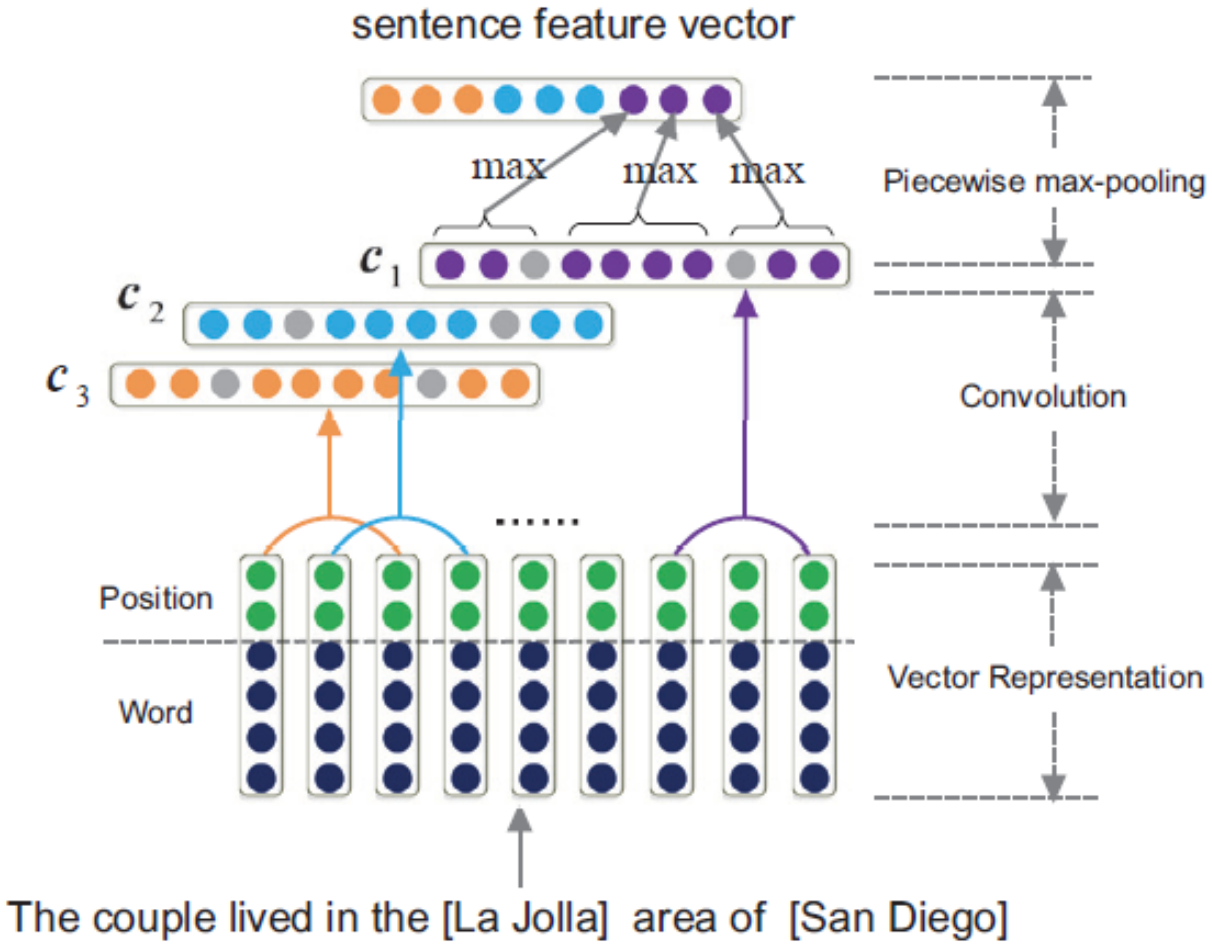


Figure 2: The framework used for extracting sentence level features.

PCNN


sentence feature vector
max
max
max
Piecewise max-pooling
c_1
c_2
c_3
Convolution
.....
Position
Word
Vector Representation
The couple lived in the [La Jolla] area of [San Diego]

Multi-instance learning with PCNN

- The training set consists of many bags

$$\{M_1, M_2, \cdots, M_T\}$$

  – The labels of the bags are known

  Ex: Label(Mi)= <Apple, founder, Steve Jobs>

- Each bag contains many instances

$$M_i = \{m_i^1, m_i^2, \cdots, m_i^{q_i}\}$$

  – The labels of the instances in the bags are unknown

  Ex: Mi1="Steve Jobs passed away the day before Apple unveiled..."

  Label(Mi1) is unknown

Multi-instance learning with PCNN

- The objective of multi-instance learning is to predict the labels of the bags

Given all (T) training bags (M_i, y_i), we can define the objective function using cross-entropy at the bag level as follows:

$$J(\theta) = \sum_{i=1}^{T} \log p(y_i | m_i^j; \theta) \qquad (8)$$

where j is constrained as follows:

$$j^* = \arg\max_j p(y_i | m_i^j; \theta) \;\; 1 \leq j \leq q_i \qquad (9)$$

Result

- Manual Evaluation on NYT articles

Top N	Mintz	MultiR	MIML	PCNNs+MIL
Top 100	0.77	0.83	0.85	**0.86**
Top 200	0.71	0.74	0.75	**0.80**
Top 500	0.55	0.59	0.61	**0.69**
Average	0.676	0.720	0.737	**0.783**

Table 2: Precision values for the top 100, top 200, and top 500 extracted relation instances upon manual evaluation.

Problems of Zeng et al EMNLP 2015

- A bag may contain multiple valid sentences, but Zeng et al only selects one sentence

Freebase /location/location/contains (Nevada, Las Vegas)

S1. **[Nevada] then sanctioned the sport , and the U.F.C. held its first show in [Las Vegas] in September 2001.**

S2. Pinnacle owns casinos in [Nevada], Louisiana , Indiana , Argentina and the Bahamas , but not in the top two American casino cities , Atlantic City and [Las Vegas].

S3. **He has retained two of [Nevada] 's most prominent criminal defense lawyers , Scott Freeman of Reno and David Chesnoff of [Las Vegas].**

S4. The state 's population is growing , but not skyrocketing the way it is in Arizona and [Nevada] , and with no city larger than 100,000 residents , Montana essentially does not have suburbs or exurbs like those spreading around Phoenix, [Las Vegas] and Denver.

Problems of Zeng et al EMNLP 2015

- Zeng et al did not use background knowledge for the entities

Freebase /location/location/contains (Nevada, Las Vegas)

Descriptions

[Nevada]: Nevada is a state in the Western, Mountain West, and Southwestern regions of the United States.

[Las Vegas]: officially the City of Las Vegas and often known as simply Vegas, is a city in the United States, the most populous city in the state of Nevada, the county seat of Clark County, and the city proper of the Las Vegas Valley.

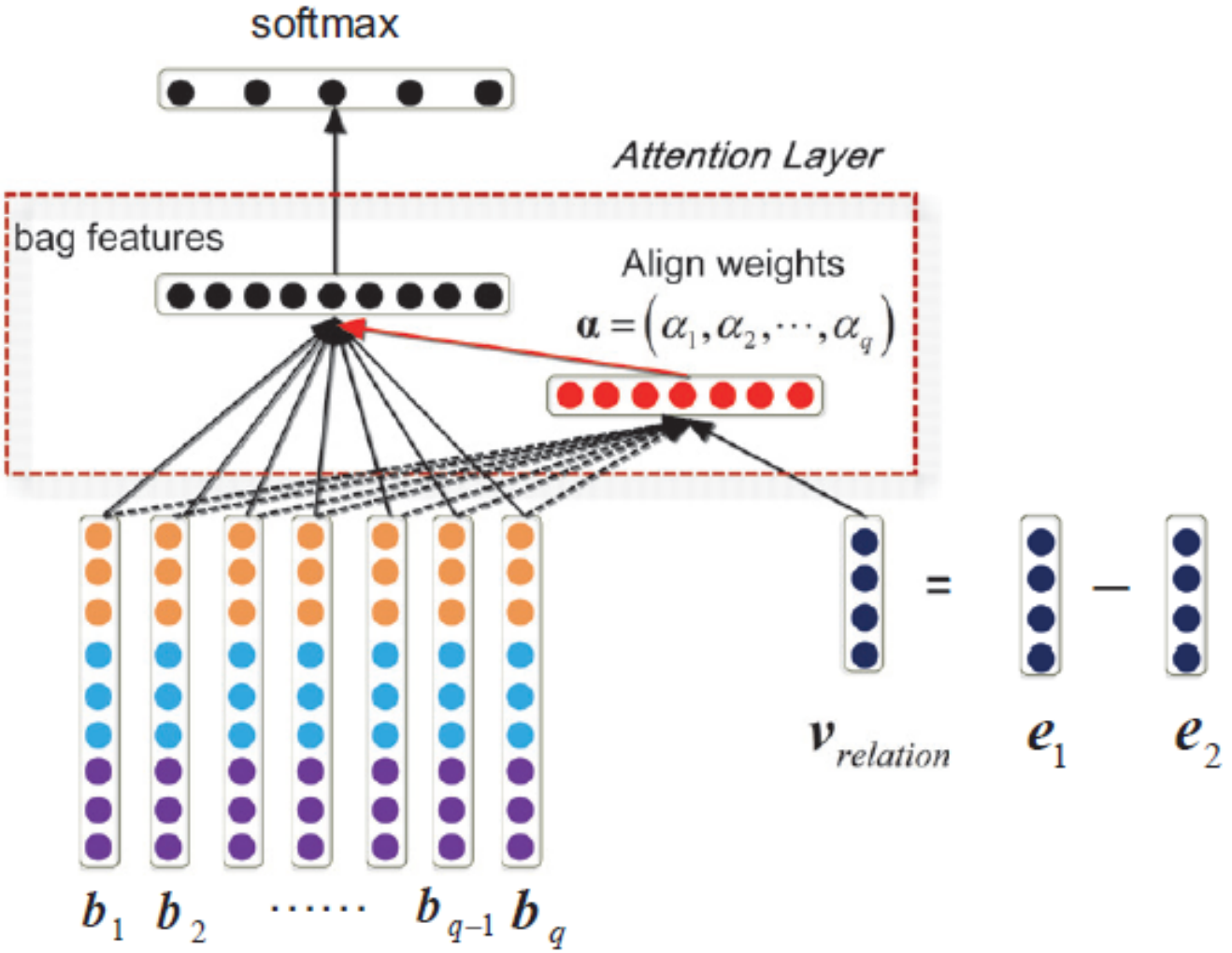
Distant Supervision for Relation Extraction

- Mintz et al "Distant supervision for relation extraction without labeled data" ACL 2009
- Zeng et al "Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks" EMNLP 2015
- Ji et al "Distant Supervision for Relation Extraction with Sentence-Level Attention and Entity Descriptions" AAAI 2017

Ji et al's Solutions

- A bag may contain multiple valid sentences, but Zeng et al only selects one sentence
  - Ji et al's Solution: Sentence-level Attention Module
- Zeng et al did not use background knowledge for the entities
  - Ji et al's Solution: Use descriptions for entities from Freebase and Wikipedia pages

Sentence-level Attention Module



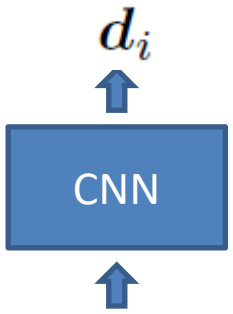
Use descriptions for entities from Freebase and Wikipedia pages

The objective function is

$$\min \mathcal{L} = \mathcal{L}_A + \lambda \mathcal{L}_e$$

$$\mathcal{L}_A = \sum_{i=1}^{N} \log p(r_i | B_i, \theta)$$

$$\mathcal{L}_e = \sum_{i=1}^{|\mathcal{D}|} \| \boldsymbol{e}_i - \boldsymbol{d}_i \|_2^2$$



Descriptions

[Nevada]: Nevada is a state in the Western, Mountain West, and Southwestern regions of the United States.

Outline

- Information Extraction
  - Distant Supervision for Relation Extraction
  - Hierarchy of Wikipedia Categories

Hierarchy of Wikipedia Categories

- Ponzetto and Strube. “Deriving a Large Scale Taxonomy from Wikipedia.” AAAI 2007
- Ponzetto and Navigli. “Large-Scale Taxonomy Mapping for Restructuring and Integrating Wikipedia”. IJCAI 2009
- Nastase and Strube. “Decoding Wikipedia Categories for Knowledge Acquisition”. AAAI 2008

Hierarchy of Wikipedia Categories are not always IsA relations

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction

Help
About Wikipedia
Community portal
Recent changes

Not l

Article | Talk | Read | View source | Vi

iPhone

From Wikipedia, the free encyclopedia
(Redirected from Iphone)

This article is about the line of smartphones by Apple. For the original iPhone, see iPho of iPhones and other uses, see iPhone (disambiguation).

This article **may be too long to read and navigate comfortabl** 91 kilobytes. Please consider splitting content into sub-articles, subheadings. *(February 2019)*

The **iPhone** is a line of smartphones designed and marketed by Apple Inc. All generations of the iPhone use Apple's iOS mobile operating system software. The first-generation iPhone was released on June 29, 2007, and multiple new hardware iterations with new iOS releases have been released since.

Categories: IPhone | Apple Inc. mobile phones | Computer-related introductions in 2007 | Digital audio players | IOS | ITunes | Mobile phones introduced in 2007 | Smartphones

Hierarchy of Wikipedia Categories/Pages are not always IsA relations

Categories: IPhone | Apple Inc. mobile phones | Computer-related introductions in 2007 | Digital audio players | IOS | ITunes | Mobile phones introduced in 2007 | Smartphones

<page, R, category>

<iPhone, IsA, IPhone>
<iPhone, IsA, Apple Inc. mobile phones>
<iPhone, IsA, Computer-related introductions in 2007>
<iPhone, IsA, Digital audio players>
<iPhone, OperationSystem, IOS>
<iPhone, App, ITunes>
<iPhone, IsA, Mobile phones introduced in 2007>
<iPhone, IsA, Smartphones>

Category:iPhone

From Wikipedia, the free encyclopedia

*The main article for this category is **iPhone**.*

Subcategories

This category has the following 3 subcategories, out of 3 total.

I

- IPhone accessories (6 P)
- IPhone video game engines (15 P)

S

- IOS software (9 C, 767 P)

Pages in category "iPhone"

The following 45 pages are in this category, out of 45 total. This list may not reflect recent changes (learn more).

- IPhone

*

- IPhone (1st generation)
- IPhone 3G
- IPhone 3GS
- IPhone 4
- IPhone 4S
- IPhone 5
- IPhone 5C
- IPhone XS
- *IPhone XS Max*

0–9

- 300-page iPhone bill

C

- Imran Chaudhri
- Cocoa Touch

D

- Dock connector

Categories: Apple Inc. mobile phones | Touchscreen mobile phones | Smartphones | IOS

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item

In other projects
Wikimedia Commons

Print/export
Create a book
Download as PDF
Printable version

Category | Talk

Category:iPhone

From Wikipedia, the free encyclopedia

*The main article for this category is **iPhone**.*

Subcategories

This category has the following 3 subcategories, out of 3 total.

I

- IPhone accessories (6 P)
- IPhone video game engines (15 P)

S

- IOS software (9 C, 767 P)

Pages in category "iPhone"

The following 45 pages are in this category, out of 45 total. This list may not reflect recent changes (learn more).

- IPhone

*

- IPhone (1st generation)
- IPhone 3G
- IPhone 3GS
- IPhone 4
- IPhone 4S
- IPhone 5
- IPhone 5C
- IPhone XS
- *IPhone XS Max*

0–9

- 300-page iPhone bill

C

- Imran Chaudhri
- Cocoa Touch

D

- Dock connector

Categories: Apple Inc. mobile phones | Touchscreen mobile phones | Smartphones | IOS

Subcategory

Page

(Hyper)category

Hierarchy of Wikipedia Categories

- Ponzetto and Strube. “Deriving a Large Scale Taxonomy from Wikipedia.” AAAI 2007
- Ponzetto and Navigli. “Large-Scale Taxonomy Mapping for Restructuring and Integrating Wikipedia”. IJCAI 2009
- Nastase and Strube. “Decoding Wikipedia Categories for Knowledge Acquisition”. AAAI 2008

Generate Hierarchy of Wikipedia Categories

- Goal: Identify IsA relations from wikipedia categoties
- Approach
  - Syntax-based Methods
  - Connectivity-based methods
  - Lexico-syntactic based methods
  - Inference-based methods

Generate Hierarchy of Wikipedia Categories

- Goal: Identify IsA relations from wikipedia categoties
- Approach
  - Syntax-based Methods
  - Connectivity-based methods
  - Lexico-syntactic based methods
  - Inference-based methods

Syntax-based Methods

- Head matching
  - Step1: Parse the category sentence
  - Step2: Find the head using heuristic rules
  - Step3: if head of subcategory is a modifier of category then we have <subcategory, NotIsA, category>

Ex: <Apple Inc. mobile phones, IsA, phones >

Syntax-based Methods

- Modifier matching
  - Step1: Parse the category sentence
  - Step2: Find the head using heuristic rules
  - Step3: if head of subcategory/category is a modifier of its counterpart, then we have <subcategory, NotIsA, category>

Ex: <Basketball equipment, NotIsA, Basketball >

Connectivity-based methods

- For a page and its category, if head(category) is plural, then we have <page, IsA, category>

Ex: <Stephen Curry, IsA, African-American basketball players >

Lexico-syntactic based methods

- Identify <X, IsA, Y> if either the following patterns occur frequently in corpus
  - Y's X
  - Y with X
  - Y such as X
  - …

Inference-based methods

- If <X, IsA, Y> and <Y, IsA, Z> are identified, then we have <X, IsA, Z>

Hierarchy of Wikipedia Categories

- Ponzetto and Strube. “Deriving a Large Scale Taxonomy from Wikipedia.” AAAI 2007
- Ponzetto and Navigli. “Large-Scale Taxonomy Mapping for Restructuring and Integrating Wikipedia”. IJCAI 2009
- Nastase and Strube. “Decoding Wikipedia Categories for Knowledge Acquisition”. AAAI 2008

Use WordNet Hierarchy to help improve Wikipedia Hierarchy

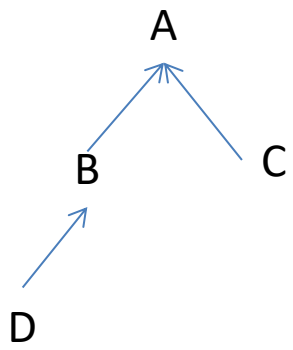
- Approach
  - Step1: Wikipedia Taxonomy is automatically mapped to WordNet (Category disambiguation)

    Ex: The meaning of Wikipedia category - "PLANTS" is the WordNet synset – "$plant^{2}_{n}$"
  - Step2: Identify links in Wikipedia hierarchy whose corresponding links in WordNet hierarchy are the most inconsistent, and find the alternative categories.

Category Disambiguation

- The disambiguation is based on maximizing the structural overlap between Wikipedia and WordNet Hierarchy

Input



Wikipedia Hierarchy

Output

A refers to a2 out of {a1, a2}
B refers to b2 out of {b1, b2, b3}
C refers to c1 out of {c1}
D refers to d1 out of {d1, d2}

Category Disambiguation

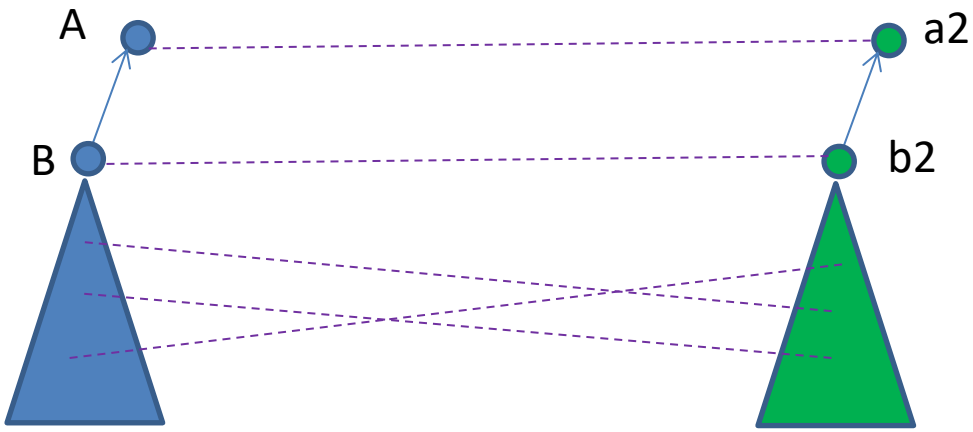
- The disambiguation is based on maximizing the structural overlap between Wikipedia and WordNet Hierarchy

Wikipedia Hierarchy

WordNet Hierarchy

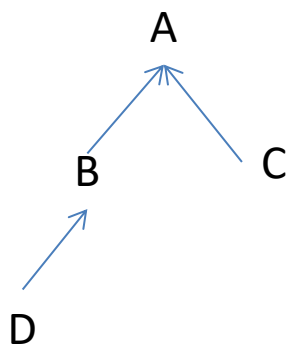
Category Disambiguation

- The disambiguation is based on maximizing the structural overlap between Wikipedia and WordNet Hierarchy

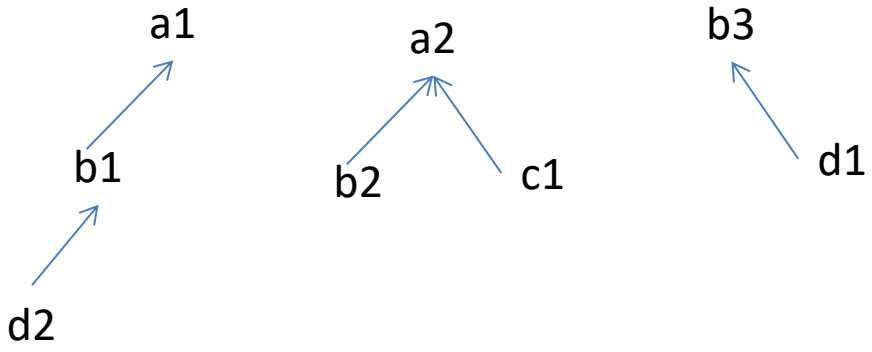


Category Disambiguation

- The disambiguation is based on maximizing the structural overlap between Wikipedia and WordNet Hierarchy



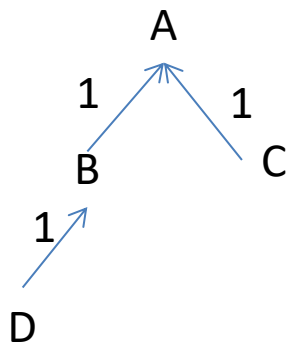
Wikipedia Hierarchy



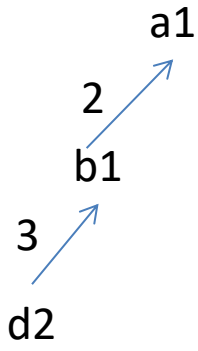
WordNet Hierarchy

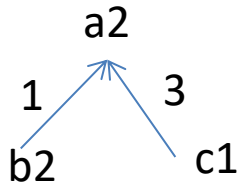
Category Disambiguation

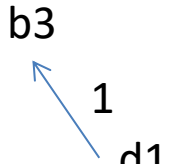
- The disambiguation is based on maximizing the structural overlap between Wikipedia and WordNet Hierarchy



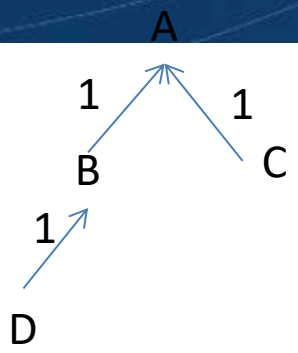
Wikipedia Hierarchy with distance



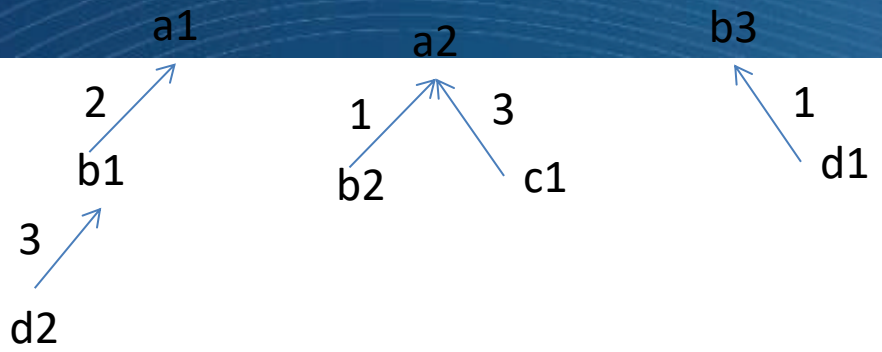




WordNet Hierarchy distance



Wikipedia Hierarchy with distance



WordNet Hierarchy distance

For each v0

$$w(v, v') = w(v, v') + \frac{1}{2^{d_{WN}(v_0, v')-1} \cdot 2^{d_{Wiki}(c_0, c')-1}}$$

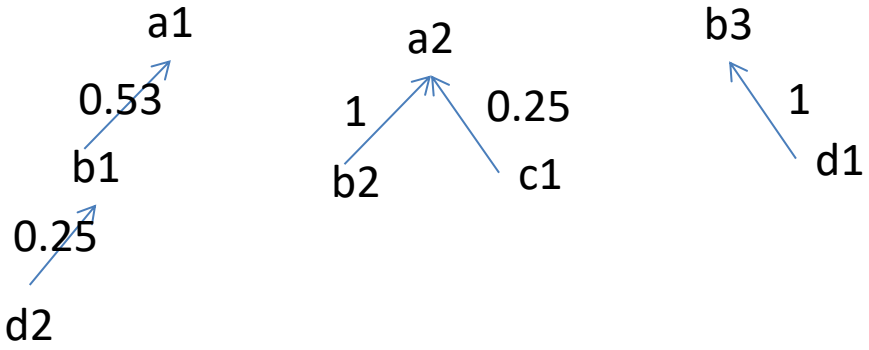
w(b1,d2)=1/4=0.25

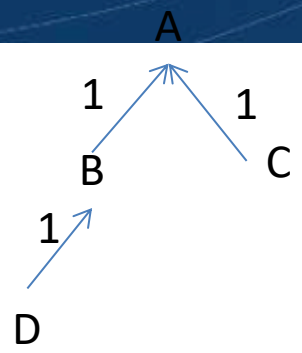
w(a1,b1)=1/2+1/32=0.53

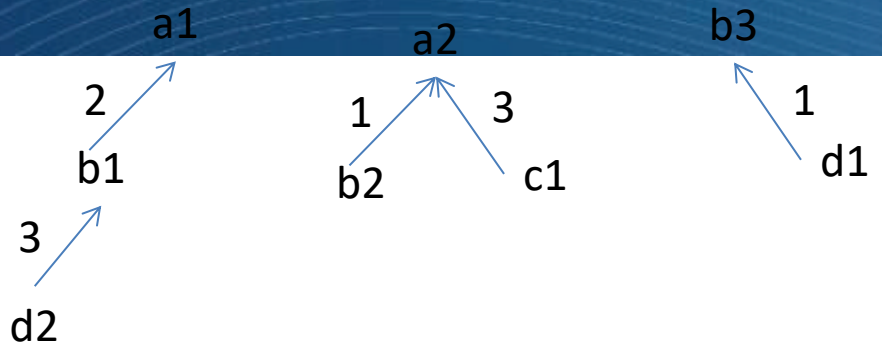
w(a2,b2)=1

w(a2,c1)=1/4=0.25

w(b3,d1)=1







Wikipedia Hierarchy with distance

WordNet Hierarchy distance

For each v0

$$w(v, v') = w(v, v') + \frac{1}{2^{d_{WN}(v_0, v')-1} \cdot 2^{d_{Wiki}(c_0, c')-1}}$$

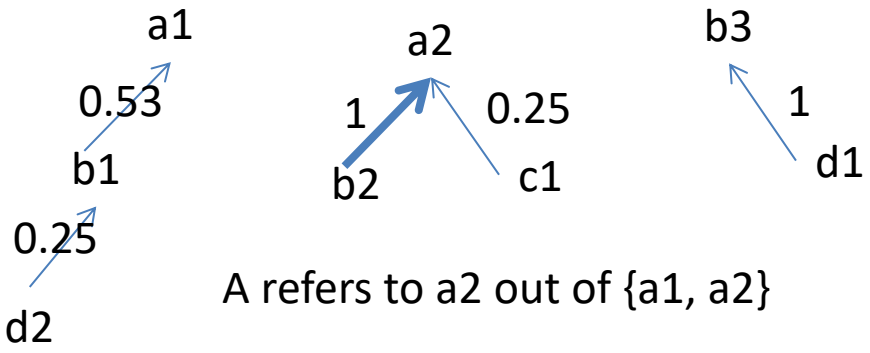
w(b1,d2)=1/4=0.25

w(a1,b1)=1/2+1/32=0.53

w(a2,b2)=1

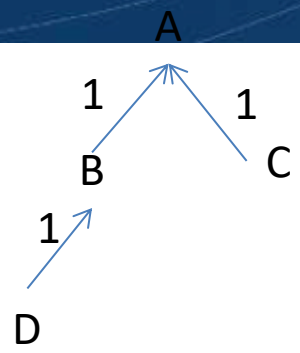
w(a2,c1)=1/4=0.25

w(b3,d1)=1

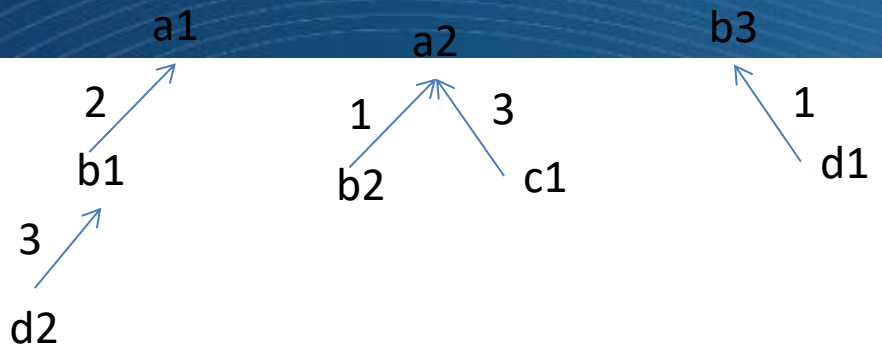


A refers to a2 out of {a1, a2}

B refers to b2 out of {b1, b2, b3}



Wikipedia Hierarchy with distance



WordNet Hierarchy distance

For each v0

$$w(v, v') = w(v, v') + \frac{1}{2^{d_{WN}(v_0, v')-1} \cdot 2^{d_{Wiki}(c_0, c')-1}}$$

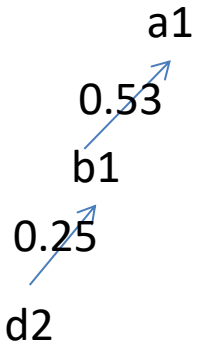
w(b1,d2)=1/4=0.25

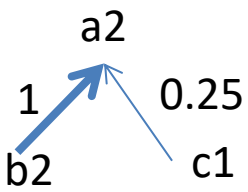
w(a1,b1)=1/2+1/32=0.53

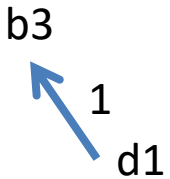
w(a2,b2)=1

w(a2,c1)=1/4=0.25

w(b3,d1)=1



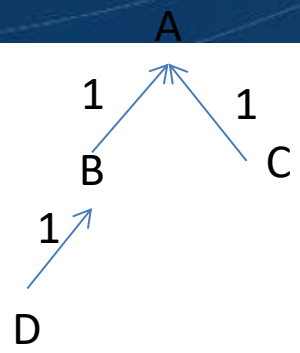




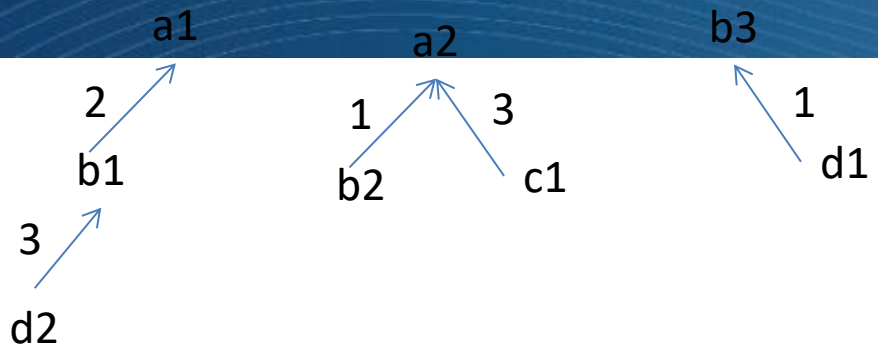
A refers to a2 out of {a1, a2}

B refers to b2 out of {b1, b2, b3}

D refers to d1 out of {d1, d2}



Wikipedia Hierarchy with distance



WordNet Hierarchy distance

For each v0

$$w(v, v') = w(v, v') + \frac{1}{2^{d_{WN}(v_0, v')-1} \cdot 2^{d_{Wiki}(c_0, c')-1}}$$

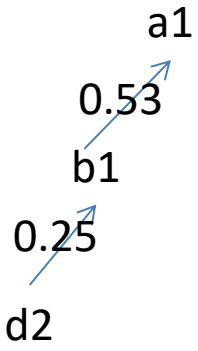
w(b1,d2)=1/4=0.25

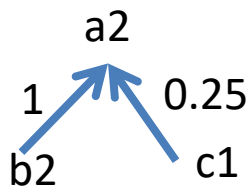
w(a1,b1)=1/2+1/32=0.53

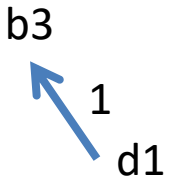
w(a2,b2)=1

w(a2,c1)=1/4=0.25

w(b3,d1)=1







A refers to a2 out of {a1, a2}

B refers to b2 out of {b1, b2, b3}

D refers to d1 out of {d1, d2}

C refers to c1 out of {c1}

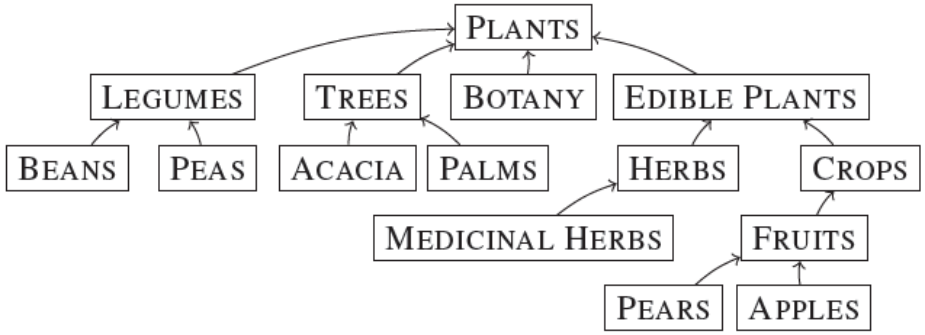


Figure 1: An excerpt of the Wikipedia category tree rooted at PLANTS.

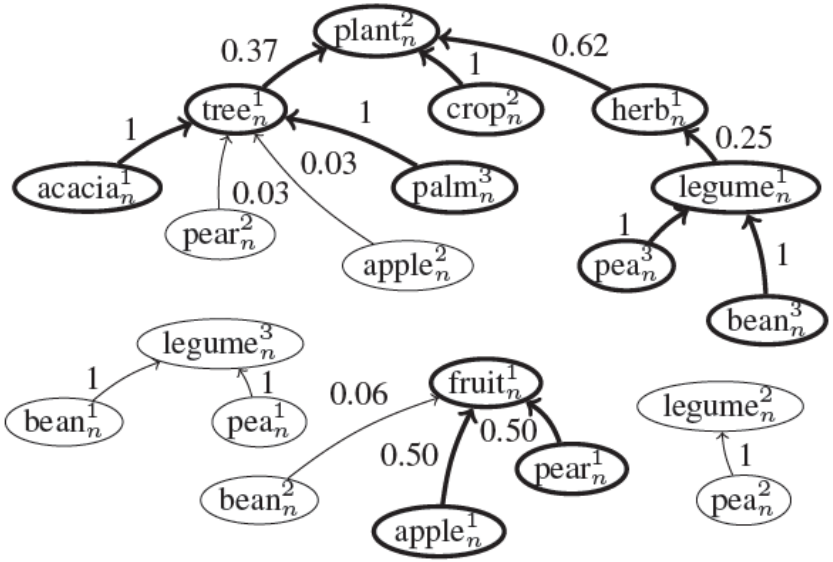


Figure 2: An excerpt of the WordNet graph associated with the category tree rooted at PLANTS. Thick lines correspond to highest-ranking edges and their incident vertices selected as sense interpretations for the corresponding categories. Singleton vertices are not shown.

Hierarchy of Wikipedia Categories

- Ponzetto and Strube. “Deriving a Large Scale Taxonomy from Wikipedia.” AAAI 2007
- Ponzetto and Navigli. “Large-Scale Taxonomy Mapping for Restructuring and Integrating Wikipedia”. IJCAI 2009
- Nastase and Strube. “Decoding Wikipedia Categories for Knowledge Acquisition”. AAAI 2008

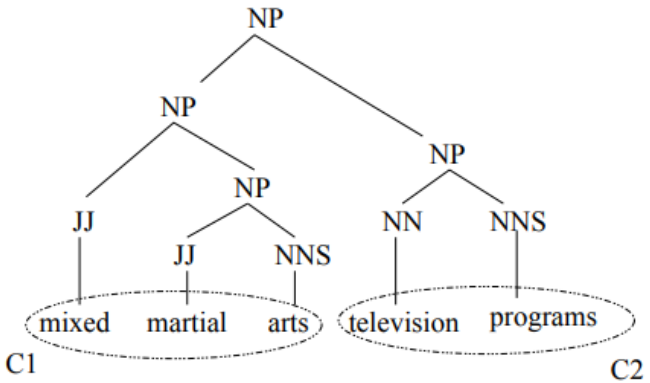
Extraction from Category Names

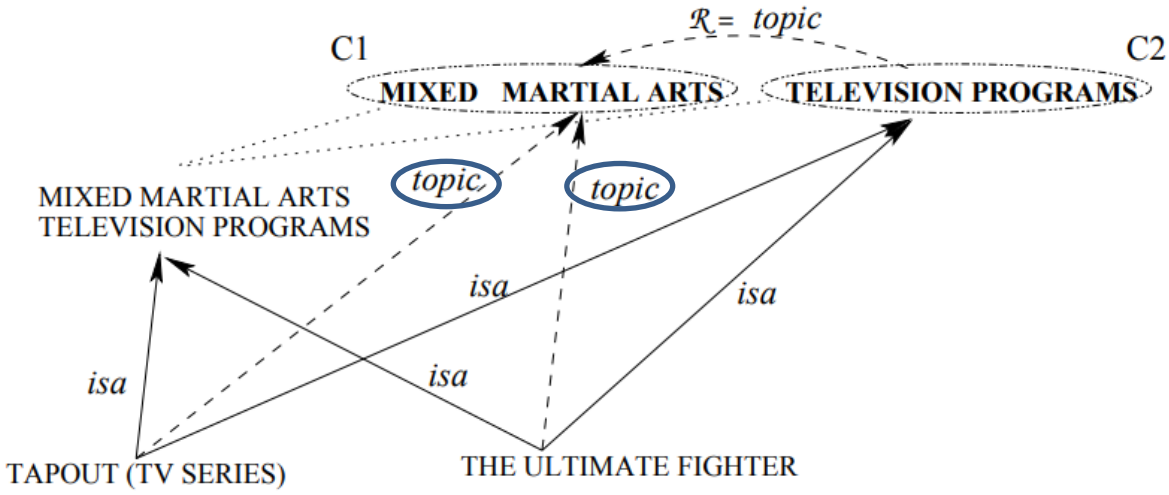
Category type	Category name	Pattern	Relations
explicit relation	QUEEN (BAND) MEMBERS	X members members of X	FREDDY MERCURY *member_of* QUEEN (BAND) BRIAN MAY *member_of* QUEEN (BAND) ...
explicit relation	MOVIES DIRECTED BY WOODY ALLEN	X [VBN IN] Y	ANNIE HALL *directed_by* WOODY ALLEN ANNIE HALL *isa* MOVIE DECONSTRUCTING HARRY *directed_by* WOODY ALLEN DECONSTRUCTING HARRY *isa* MOVIE ...
partly explicit relation	VILLAGES IN BRANDENBURG	X [IN] Y	SIETHEN *located_in* BRANDENBURG SIETHEN *isa* VILLAGE ...
implicit relation	MIXED MARTIAL ARTS TELEVISION PROGRAMS	X Y	MIXED MARTIAL ARTS $\mathcal{R}$ TELEVISION PROGRAMS TAPOUT (TV SERIES) $\mathcal{R}$ MIXED MARTIAL ARTS TAPOUT (TV SERIES) *isa* TELEVISION PROGRAM ...
class attribute	ALBUMS BY ARTIST	X by Y	ARTIST *attribute_of* ALBUM MILES DAVIS *isa* ARTIST BIG FUN *isa* ALBUM ...

Table 1: Examples of information encoded in category names and the knowledge we extract

Extraction from Category Names and Network

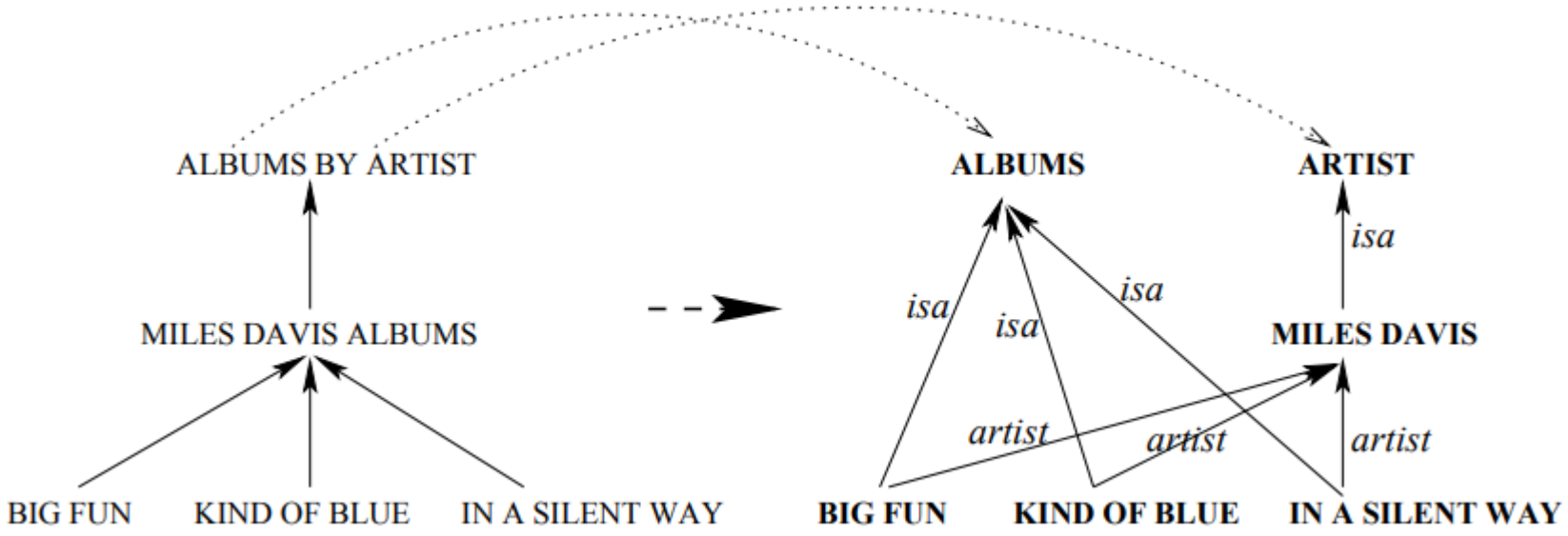
- Implicit Relation Categories





Extraction from Category Names and Network

- Extract class attribute and attribute values



Thank You!

Q/A