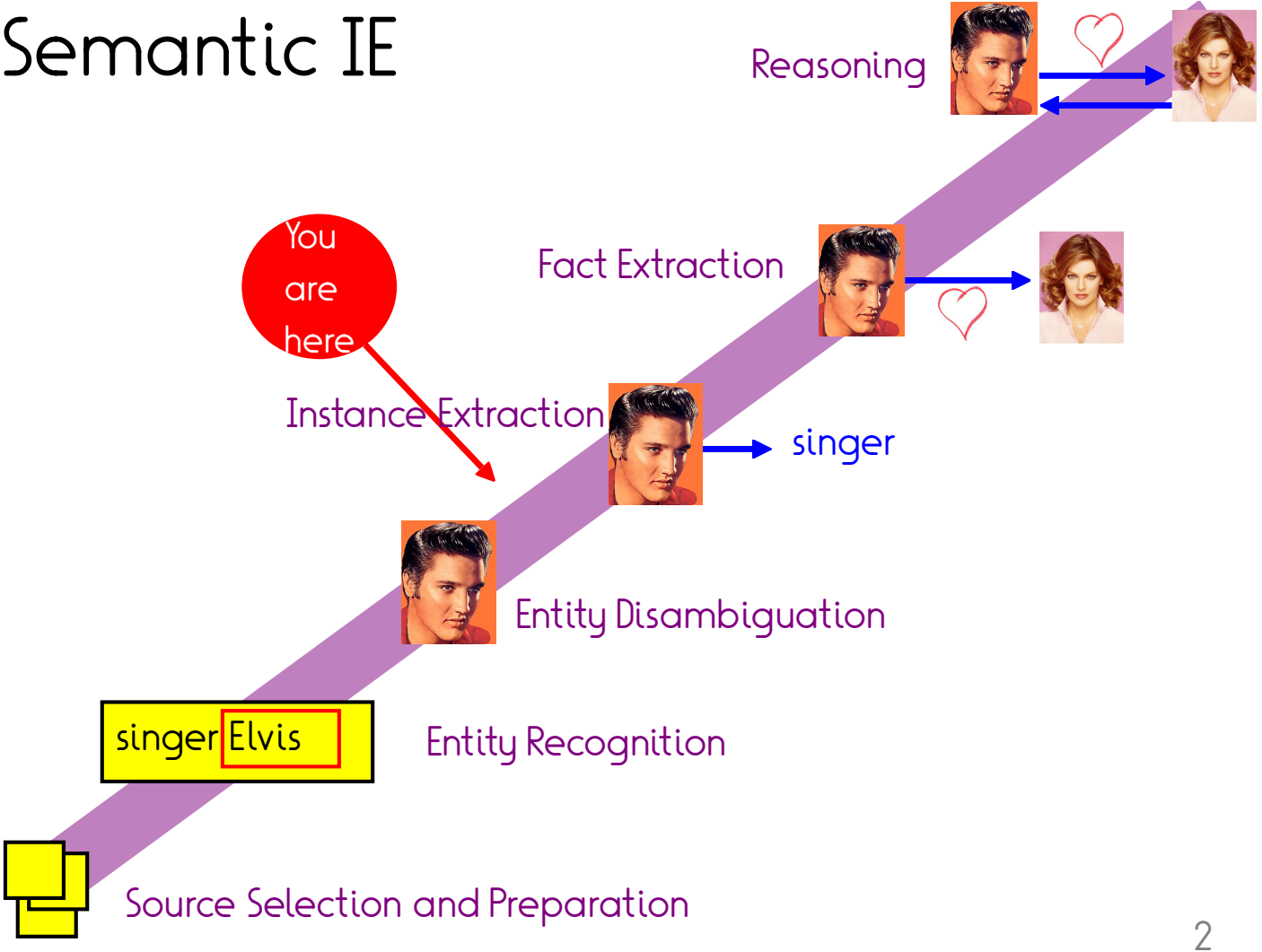
Instance Extraction

Nada Mimouni

Based on slides by:
Fabian M. Suchanek

Semantic IE


Reasoning
You are here
Fact Extraction
Instance Extraction
singer
Entity Disambiguation
singer Elvis
Entity Recognition
Source Selection and Preparation

Def: IsA

Observation: The relations "subclass" and "type" are expressed very similarly in natural language:

Elvis is a singer.	type(Elvis, singer)
A dog is an animal.	subclassof(dog, animal)

For now, let us ignore the distinction between the two.

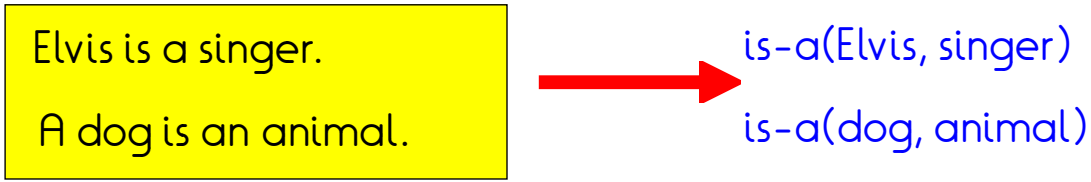
IsA is the relation that holds between x and y if x is an instance of y, or x is a subclass of y.

is-a(Elvis, singer)

is-a(dog, animal)

IsA Extraction

IsA Extraction is the task of extracting "IsA" facts from a corpus.



The problem: Corpora rarely contain sentences that explicitly define the type of an instance.

Example: IsA Extraction

In the Simpson episode "HOMR", Doctor Monson discovers a crayon in Homer's brain and removes it. His IQ goes up from 55 to 105, but he feels uncomfortable and wants it back. Moe, who is not only a bartender but also an unlicensed physician, puts the crayon back, returning Homer to the idiot.

Example: IsA Extraction

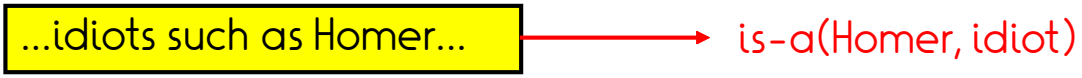
In the Simpson episode "HOMR", Doctor Monson discovers a crayon in Homer's brain and removes it. His IQ goes up from 55 to 105, but he feels uncomfortable and wants it back. Moe, who is not only a bartender but also an unlicensed physician, puts the crayon back, returning Homer to the idiot.

HOMR	isA	Simpson episode
Monson	isA	Doctor
Homer	isA	idiot
Moe	isA	bartender
Moe	isA	unlicensed physician

Def: Hearst Patterns

A Hearst pattern is a simple textual pattern that indicates an IsA fact that is mentioned implicitly.

"Y such as X"

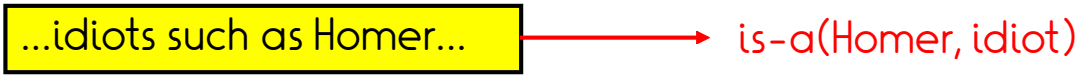


Marti Hearst: Automatic Acquisition of Hyponyms, COLING 1992

Def: Hearst Patterns

A Hearst pattern is a simple textual pattern that indicates an IsA fact that is mentioned implicitly.

"Y such as X"



...many activists, such as Lisa...

...some animals, such as dogs...

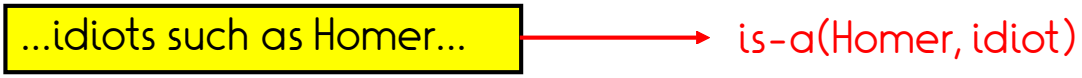
...some scientists, such as computer scientists...

...some plants, such as nuclear power plants....

Def: Hearst Patterns

A Hearst pattern is a simple textual pattern that indicates an IsA fact that is mentioned implicitly.

"Y such as X"



...many activists, such as Lisa...

is-a(Lisa, activist)

...some animals, such as dogs...

is-a(dog, animal)

...some scientists, such as computer scientists...

is-a(computer, scientist) ?

...some plants, such as nuclear power plants....

is-a(nuc.Pow.Plants, plants) ?

Def: Hearst Patterns

A Hearst pattern is a simple textual pattern that indicates an IsA fact that is mentioned implicitly.

"Y such as X"

...idiots such as Homer... → is-a(Homer, idiot)

...many activists, such as Lisa...

is-a(Lisa, activist)

...some animals, such as dogs...

is-a(dog, animal)

...some scientists, such as computer scientists...

is-a(computer, scientist) ?

...some plants, such as nuclear power plants....

is-a(nuc.Pow.Plants, plants) ?

Hearst patterns need

- NER
- disambiguation
- plural removal

Def: Classical Hearst Patterns

The classical Hearst Patterns are

Y such as X+
such Y as X+
X+ and other Y
Y including X+
Y, especially X+

...where X+ is a list of names of the form "X_1,...,X_{n-1} (and|or)? X_n".

(In the original paper, the X_i are noun phrases)

These imply is-a(X_i,Y).

(assuming that the words are noun phrases and disambiguated)

>examples

Marti Hearst: Automatic Acquisition of Hyponyms, COLING 1992

Task: Classical Hearst Patterns

Apply

1. Y such as X+
2. such Y as X+
3. X+ and other Y
4. Y including X+
5. Y, especially X+

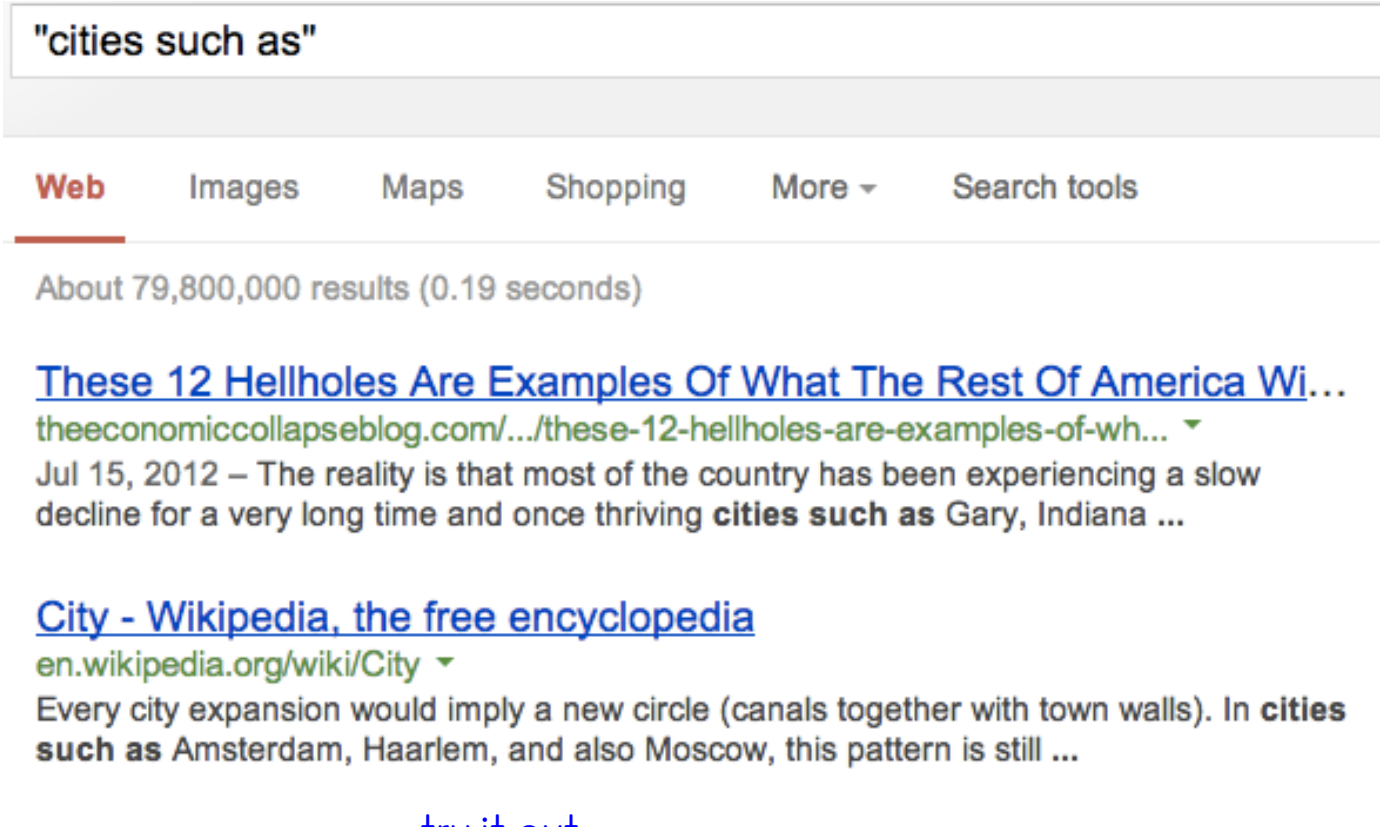
I lived in such countries as Germany, France, and Bavaria.

Trump is a candidate for the Nobel Peace Prize, together with Kim-Jon-Un and other world-class-leaders.

I love people that are not genies, especially Homer.

>examples

Example: Hearst on the Web



try it out

>examples

Problems with Hearst Patterns

Hearst Patterns won't extract the right is-a facts from

- ...domestic animals other than dogs such as cats ...
- ...companies such as IBM, Nokia, Proctor and Gamble ...
- ...classic movies such as Gone with the Wind ...
- ...people in Europe, Russia, Brazil, China, and other countries ...

->taxonomy induction
->set expansion
-> end

Wentao Wu, Hongsong Li, Haixun Wang, Kenny Q. Zhu:
Probase: A Probabilistic Taxonomy for Text Understanding, SIGMOD 2012

Determining the right super-concept

1. Extract all possible super-concepts and all possible sub-concepts

   is-a(cat, dog) ?

   is-a(cat, domestic animal) ?

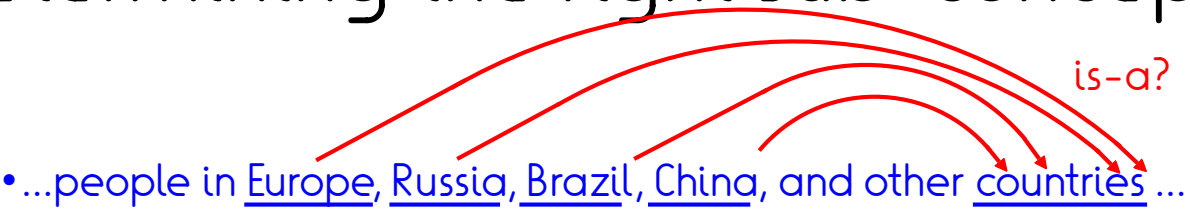
2. Choose the most likely one, given what we have seen before

   We have seen is-a(cat, animal) more often than is-a(cat, dog)

   => is-a(cat, domestic animal)

Wentao Wu, Hongsong Li, Haixun Wang, Kenny Q. Zhu:
Probase: A Probabilistic Taxonomy for Text Understanding, SIGMOD 2012

Determining the right sub-concept



1. The words that are close to the pattern words are most likely correct

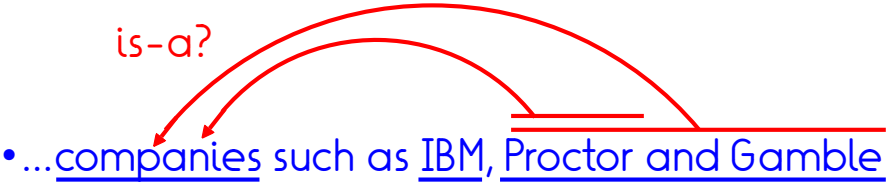
   is-a(China, country)

2. If a word has been seen before, then all words between it and the pattern are most likley correct.

   Assume we have seen is-a(Russia, country) before
   => is-a(Brazil, country)

Wentao Wu, Hongsong Li, Haixun Wang, Kenny Q. Zhu:
Probase: A Probabilistic Taxonomy for Text Understanding, SIGMOD 2012

Determining the right sub-concept



1. Check how often the words co-occur

$P(\text{IBM} \mid \text{Proctor, companies})$

vs $P(\text{IBM} \mid \text{Proctor and Gamble, companies})$

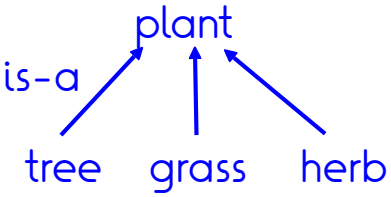
2. Choose more likely one

=> is-a(Proctor and Gamble, company)

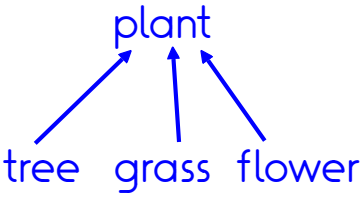
Wentao Wu, Hongsong Li, Haixun Wang, Kenny Q. Zhu:
Probase: A Probabilistic Taxonomy for Text Understanding, SIGMOD 2012

Distinguishing word senses

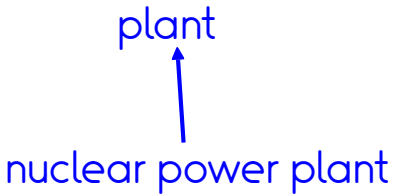
plants such as trees, grass, and herbs



plants such as trees, grass, and flowers



plants such as nuclear power plants

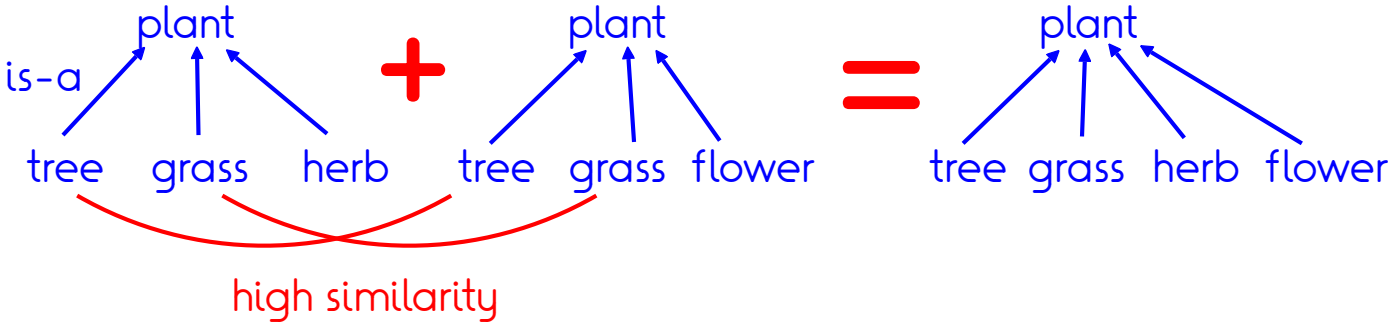


Wentao Wu, Hongsong Li, Haixun Wang, Kenny Q. Zhu:
Probase: A Probabilistic Taxonomy for Text Understanding, SIGMOD 2012

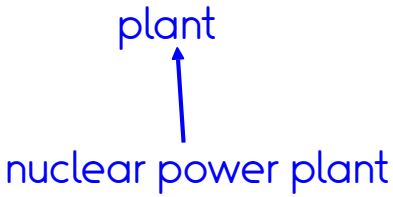
Distinguishing word senses

plants such as trees, grass, and herbs

plants such as trees, grass, and flowers



plants such as nuclear power plants

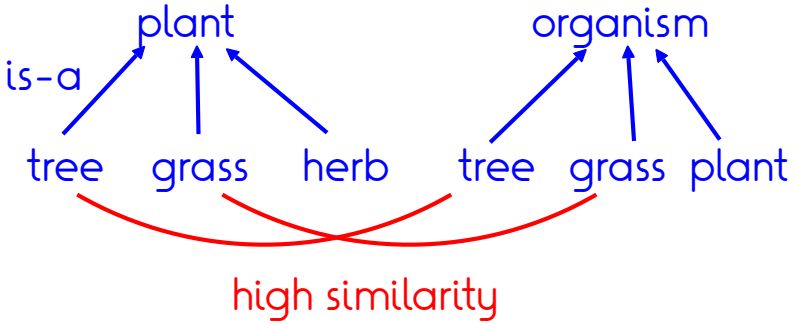


Wentao Wu, Hongsong Li, Haixun Wang, Kenny Q. Zhu:
Probase: A Probabilistic Taxonomy for Text Understanding, SIGMOD 2012

Constructing the taxonomy

plants such as trees, grass, and herbs

organisms such as trees, grass, plants

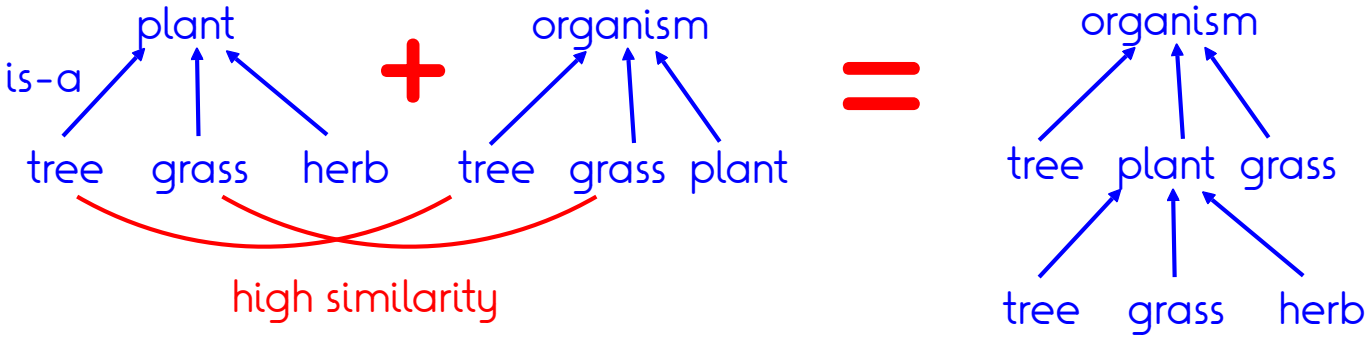


Wentao Wu, Hongsong Li, Haixun Wang, Kenny Q. Zhu:
Probase: A Probabilistic Taxonomy for Text Understanding, SIGMOD 2012

Constructing the taxonomy

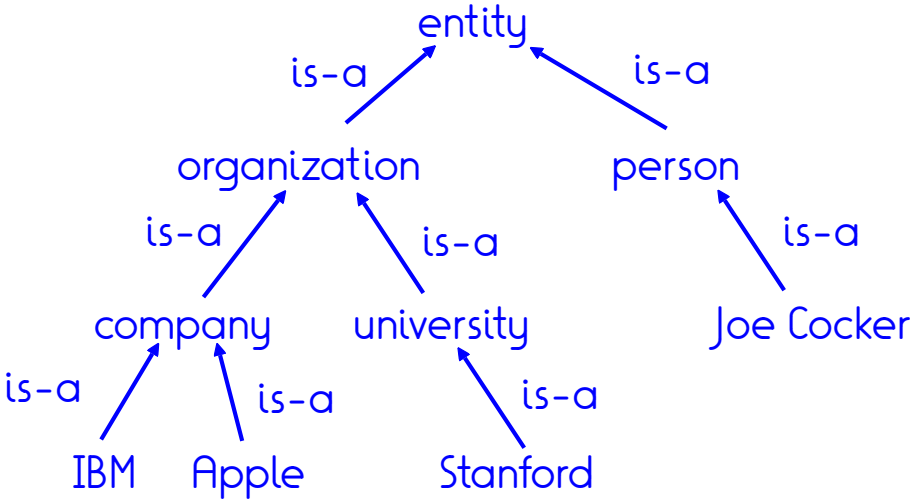
plants such as trees, grass, and herbs

organisms such as trees, grass, plants



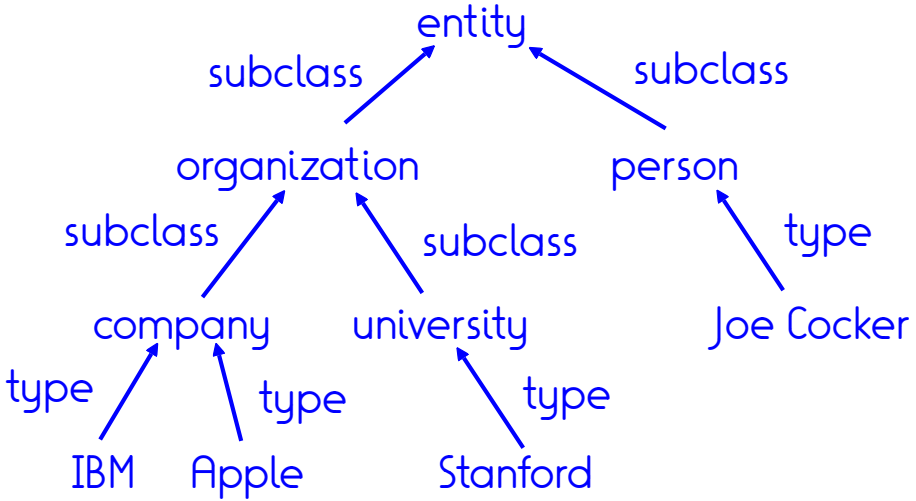
Wentao Wu, Hongsong Li, Haixun Wang, Kenny Q. Zhu:
Probase: A Probabilistic Taxonomy for Text Understanding, SIGMOD 2012

Distinguishing instances & classes



Leaf nodes of the isA taxonomy are instances.

Distinguishing instances & classes

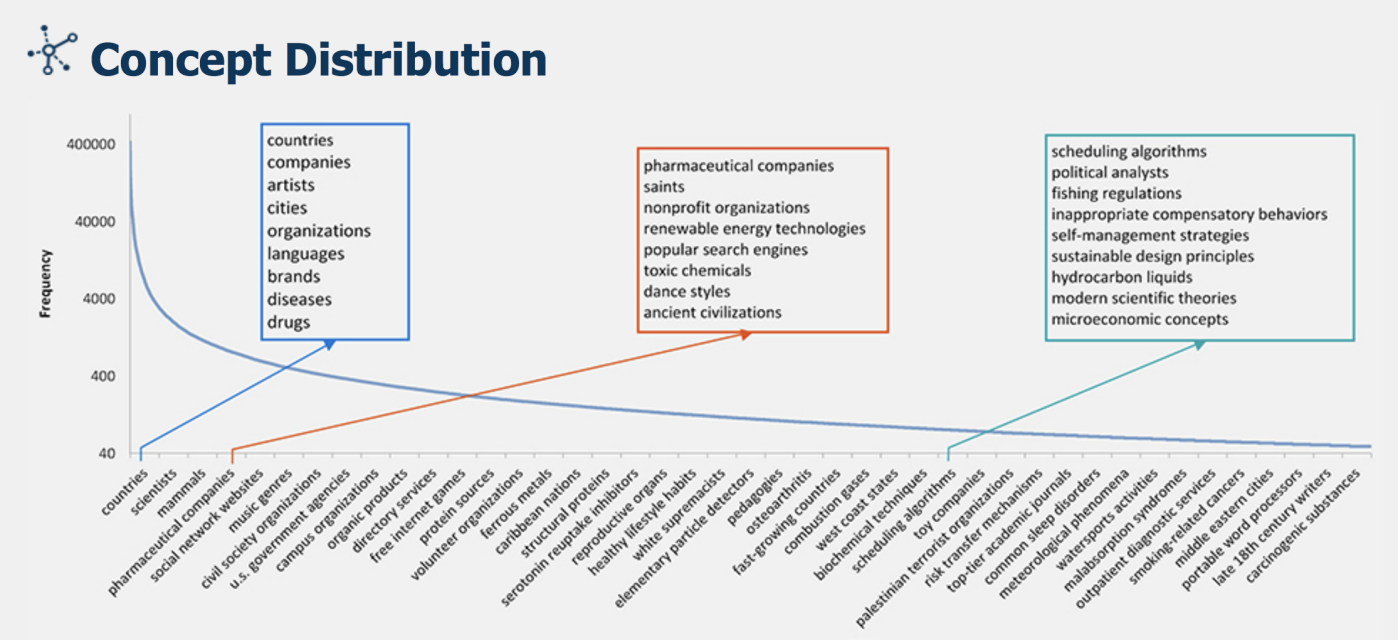


Leaf nodes of the isA taxonomy are instances.

Wentao Wu, Hongsong Li, Haixun Wang, Kenny Q. Zhu:
Probase: A Probabilistic Taxonomy for Text Understanding, SIGMOD 2012

ProBase & Microsoft Concept Graph

Probase and the Microsoft Concept Graphs were projects at Microsoft that extracted a huge is-a taxonomy from the Web.



->set expansion
-> end

Wentao Wu, Hongsong Li, Haixun Wang, Kenny Q. Zhu:
Probase: A Probabilistic Taxonomy for Text Understanding, SIGMOD 2012

Taxonomy induction

Taxonomy induction is the process of creating an entire taxonomy – from the root concept to the leaf concepts.

Usual steps include:

- instance extraction, as seen before

Candidate hypernyms for "apple"

company	5536
fruit	3898
apple	2119
vegetable	928
orange	797
tech company	619
brand	463
hardware company	460
technology company	427
food	370

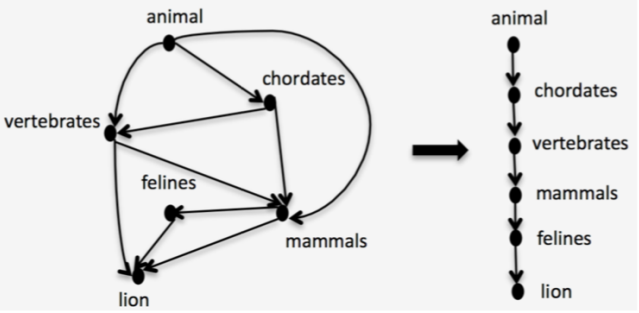
Seitner et al: "A large database of hypernymy relations extracted from the web" ("WebIsA"). LREC 2016

Taxonomy induction

Taxonomy induction is the process of creating an entire taxonomy – from the root concept to the leaf concepts.

Usual steps include:

- instance extraction, as seen before
- Removal of cycles



Zornitsa Kozareva and Eduard H. Hovy:
"A semi-supervised method to learn and construct taxonomies using the web"
EMNLP 2010

Taxonomy induction

Taxonomy induction is the process of creating an entire taxonomy – from the root concept to the leaf concepts.

Usual steps include:

- instance extraction, as seen before
- Removal of cycles
- Classify edges as "is-a" or "non-is-a" with
  - frequency counts (in both directions)
  - substring inclusion
  - difference in generality (distance to the root)

Panchenk et al: "TAXI at SemEval-2016"
10th International Workshop on Semantic Evaluation
http://tudarmstadt-lt.github.io/taxi/

Gupta et al: "Taxonomy Induction using Hypernym Subsequences", CIKM 2017

>set expansion

Def: Set Expansion

Set Expansion is the task of, given names of instances of a class ("seeds"), extracting more such instance names from a corpus.

cities: {Springfield, Seattle}

cities: {Springfield, Seattle, Washington, Chicago, ...}

Def: Recursive Pattern Application

Recursive Pattern Application is the following algorithm for set expansion:

0. Start with the seeds cities: {Austin, Seattle}

Def: Recursive Pattern Application

Recursive Pattern Application is the following algorithm for set expansion:

0. Start with the seeds — cities: {Austin, Seattle}

1. Find the pattern "X, Y, and Z" in the corpus. — Seattle, Chicago, and Austin

Def: Recursive Pattern Application

Recursive Pattern Application is the following algorithm for set expansion:

0. Start with the seeds

cities: {Austin, Seattle}

1. Find the pattern "X, Y, and Z" in the corpus.

Seattle, Chicago, and Austin

2. If 2 variables match known instance names, add the match of the 3rd.

Def: Recursive Pattern Application

Recursive Pattern Application is the following algorithm for set expansion:

0. Start with the seeds

cities: {Austin, Seattle}

1. Find the pattern "X, Y, and Z" in the corpus.

Seattle, Chicago, and Austin

2. If 2 variables match known instance names, add the match of the 3rd.

cities: {Austin, Seattle, Chicago}

3. Go to 1

Task: Recursive Pattern Appl.

cities: {Springfield, Austin, Seattle}

... Austin, Seattle, and Houston...

Task: Recursive Pattern Appl.

cities: {Springfield, Austin, Seattle}

... Austin, Seattle, and Houston...

cities: {Springfield, Austin, Seattle, Houston}

Task: Recursive Pattern Appl.

cities: {Springfield, Austin, Seattle}

... Austin, Seattle, and Houston...

cities: {Springfield, Austin, Seattle, Houston}

... Houston, Chicago, and Springfield...

Task: Recursive Pattern Appl.

cities: {Springfield, Austin, Seattle}

... Austin, Seattle, and Houston...

cities: {Springfield, Austin, Seattle, Houston}

... Houston, Chicago, and Springfield...

cities: {Springfield, Austin, Seattle, Houston, Chicago}

Task: Recursive Pattern Appl.

cities: {Springfield, Austin, Seattle}

... Austin, Seattle, and Houston...

cities: {Springfield, Austin, Seattle, Houston}

... Houston, Chicago, and Springfield...

cities: {Springfield, Austin, Seattle, Houston, Chicago}

... Austin, Texas, and Seattle, Washington...

Task: Recursive Pattern Appl.

cities: {Springfield, Austin, Seattle}

... Austin, Seattle, and Houston...

cities: {Springfield, Austin, Seattle, Houston}

... Houston, Chicago, and Springfield...

cities: {Springfield, Austin, Seattle, Houston, Chicago}

... Austin, Texas, and Seattle, Washington...

Precision may suffer over time

Def: Semantic Drift

Semantic Drift is the problem in Set Expansion that names of instances of other classes get into the set.

cities: {Springfield, Austin, Seattle, Houston}

... Houston, Chicago, and Springfield...

cities: {Springfield, Austin, Seattle, Houston, Chicago}

... Austin, Texas, and Seattle, Washington...

cities: {Chicago, Seattle, ..., Texas}

>tables

Def: Table Set Expansion

Table Set Expansion is the following algorithm for set expansion:

0. Start with the seeds

countries: {Russia, China}

1. Find HTML tables where one column contains 2 known instance names

Largest Countries in the World

view as: list / slideshow / map

▲	Country	Total Area (sq km)
1.	Russia	17,098,242
2.	Canada	9,984,670
3.	United States	9,826,675
4.	China	9,596,961

2. Add all column entries to the set

countries: {Russia, China, Canada, United States}

3. Go to 1

>tables

Example: Table Set Expansion

countries: {Russia, China, Brazil}

Example: Table Set Expansion

countries: {Russia, China, Brazil}

Richest Countries in the World

view as: list / slideshow / map

▲	Country	GDP
1.	United States	$15,290,000,000,000
2.	China	$11,440,000,000,000
3.	India	$4,515,000,000,000
4.	Japan	$4,497,000,000,000
5.	Germany	$3,139,000,000,000
6.	Russia	$2,414,000,000,000

Example: Table Set Expansion

countries: {Russia, China, Brazil}

Richest Countries in the World

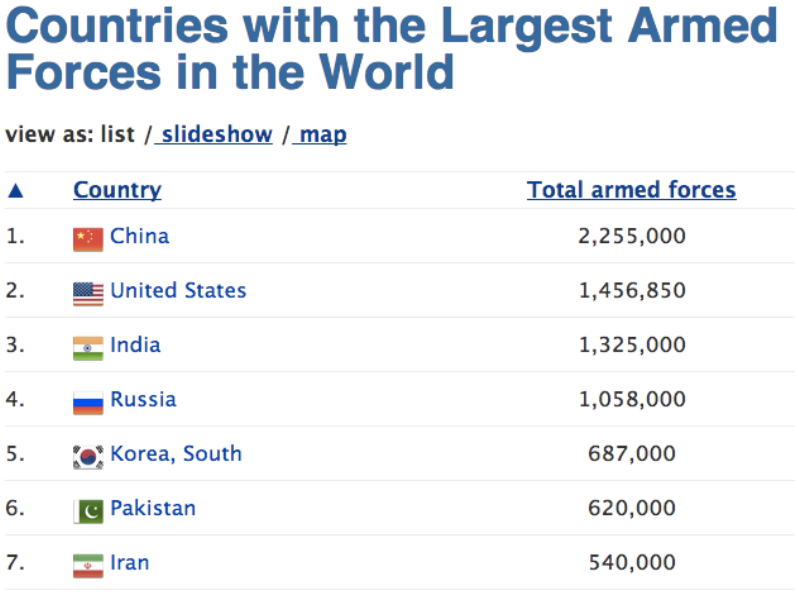
view as: list / slideshow / map

▲	Country	GDP
1.	United States	$15,290,000,000,000
2.	China	$11,440,000,000,000
3.	India	$4,515,000,000,000
4.	Japan	$4,497,000,000,000
5.	Germany	$3,139,000,000,000
6.	Russia	$2,414,000,000,000

countries: {Russia, China, Brazil, United States, Japan, India, Germ.}

Example: Table Set Expansion

countries: {Russia, China, Brazil, United States, Japan, India, Germ.}



Countries with the Largest Armed Forces in the World

view as: list / slideshow / map

▲	Country	Total armed forces
1.	China	2,255,000
2.	United States	1,456,850
3.	India	1,325,000
4.	Russia	1,058,000
5.	Korea, South	687,000
6.	Pakistan	620,000
7.	Iran	540,000

countries: {Russia, …, Germany, Korea, South, Pakistan, Iran}

Example: Table Set Expansion

countries: {Russia, …, Germany, Korea, South, Pakistan, Iran}

Countries and dependencies

Rank	Country	To km
—	*World*	51 (196
1	Russia	1 (6
—	*Antarctica*	1 (5
2	Canada	(3
3	China	(3
4	America	(3

countries: {
Russia,…,
World,
Antarctica,
America}

Example: Table Set Expansion

countries: {
Russia,...,
World,
Antarctica,
America}

All continents:
Antarctica
Africa
Asia
America
Australia
Europe

Example: Table Set Expansion

countries: {
Russia,...,
World,
Antarctica,
America}

All continents:
Antarctica
Africa
Asia
America
Australia
Europe

Semantic Drift may occur

Summary: Set Expansion

Set Expansion extends a set of instance names. We saw 2 methods:

1. Recursively applied patterns

X, Y, and Z

2. Table Set Expansion

Richest Countries in the World

view as: list / slideshow / map

▲	Country	GDP
1.	United States	$15,290,000,000,000
2.	China	$11,440,000,000,000
3.	India	$4,515,000,000,000
4.	Japan	$4,497,000,000,000
5.	Germany	$3,139,000,000,000
6.	Russia	$2,414,000,000,000

Summary: IsA Extraction

IsA extraction finds instances with their class in corpora.

We saw 2 methods:

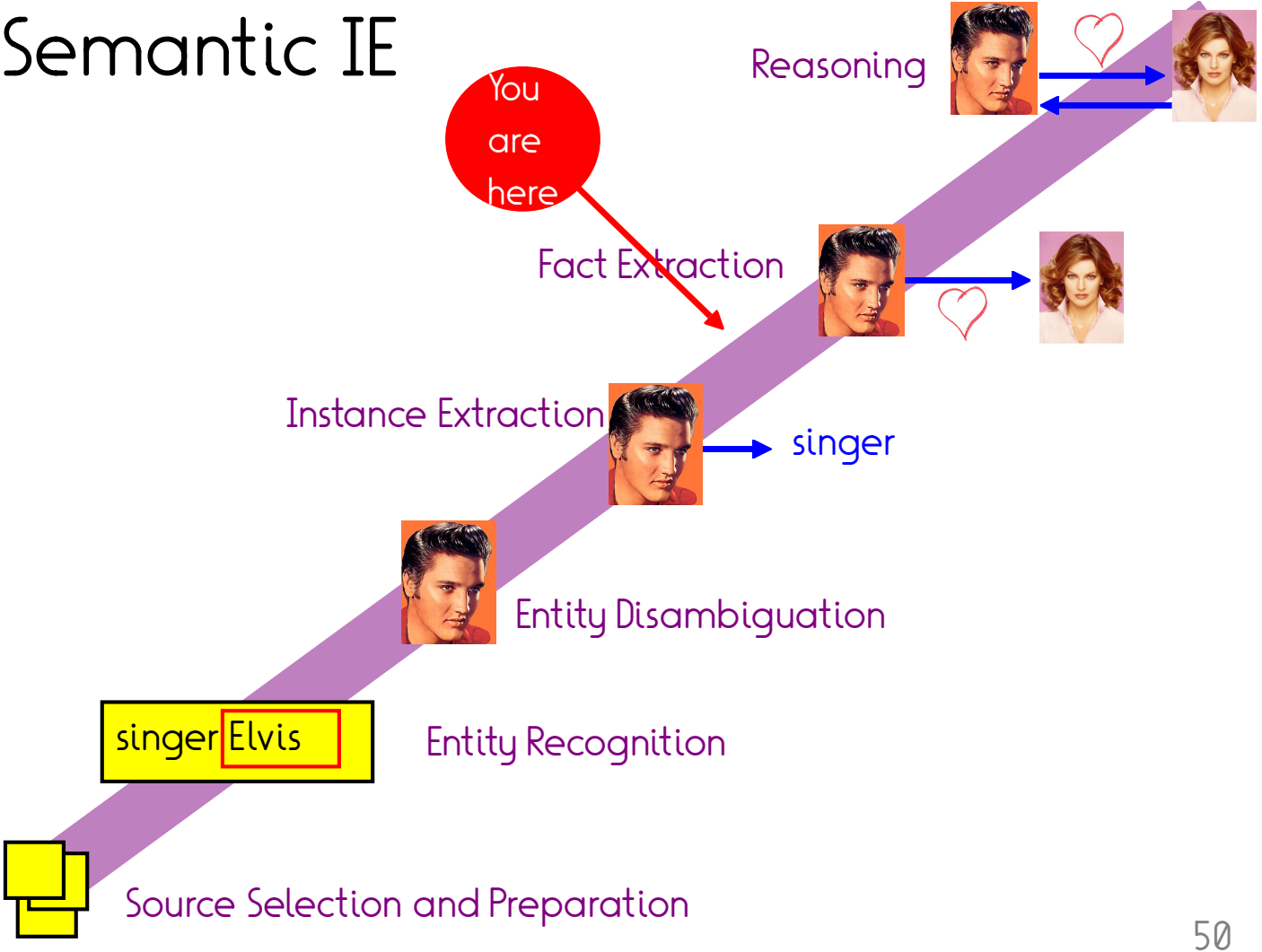
1. Hearst Patterns

vegetarians such as Lisa

2. Set Expansion

cities: {Chicago, Springfield}

Semantic IE


You are here
Reasoning
Fact Extraction
Instance Extraction
singer
Entity Disambiguation
singer Elvis
Entity Recognition
Source Selection and Preparation

References

Fabian Suchanek and Gerhard Weikum:
Knowledge Harvesting from Text and Web Sources, ICDE 2010 tutorial

Marti Hearst:
Automatic Acquisition of Hyponyms, COLING 1992

Zornitsa Kozareva and Eduard Hovy:
Learning Arguments and Supertypes of Semantic Relations, ACL 2010

Wentao Wu, Hongsong Li, Haixun Wang, Kenny Q. Zhu:
Probase: A Probabilistic Taxonomy for Text Understanding, SIGMOD 2012

->wrapper-induction

->dipre