



Constructing Knowledge Graph from Unstructured Text

Kundan Kumar

Siddhant Manocha

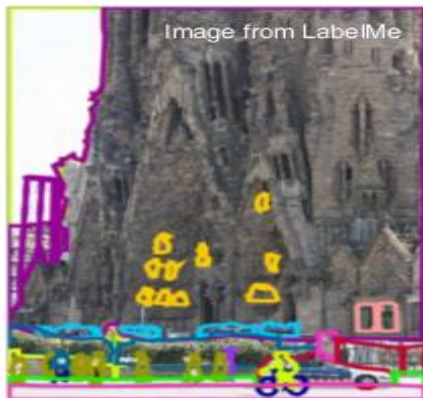
MOTIVATION



Text



Video



Images



Speech/sounds



Artificial worlds?

How to jointly acquire
knowledge from all
these sources?

MOTIVATION

The image shows two overlapping Google search results. The top search is for "san francisco population" and the bottom search is for "san francisco".

Search 1: "san francisco population"

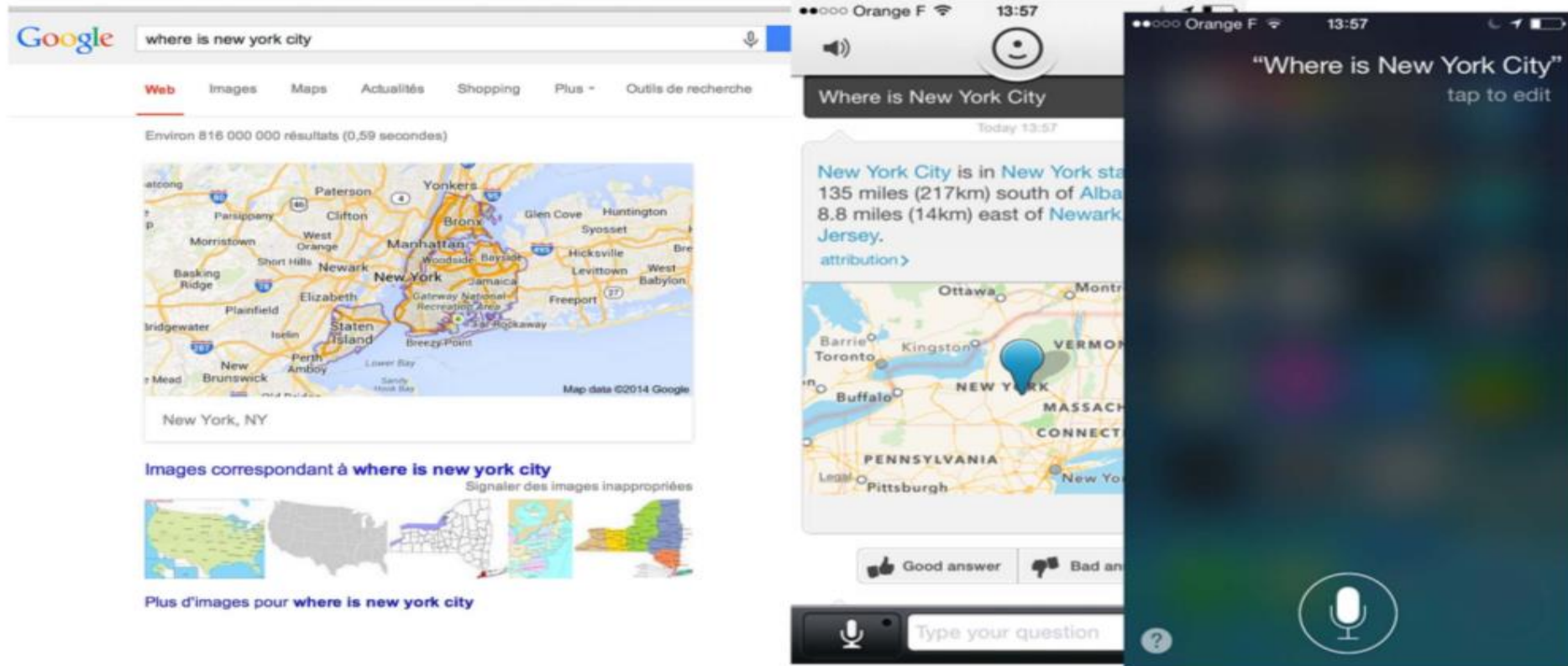
- Search results: About 79,700,000 results (0.38 seconds)
- Result 1: **Population, San Francisco, CA** (www.google.com/publicdata)
 - 812,826 - Jul 2011
 - Source: U.S. Census Bureau
- Result 2: **San Francisco** (Map)

Search 2: "san francisco"

- Search results: 50 personal results, 1,510,000,000 other results.
- Result 1: **Welcome to SFGOV, City and County of San Francisco Official site** (www.ci.sf.ca.us/)
 - SFGOV is the official website of the government of the City and County of San Francisco, providing information about all departments, meetings, legislation, ...
- Result 2: **SFO - San Francisco International Airport - Home Page** (www.flysfo.com/)
 - Guides on the airlines, concessions, general services, ground transportation and shopping can be found, including flight information, statistics, future ...
- Result 3: **San Francisco - Wikipedia, the free encyclopedia** (en.wikipedia.org/wiki/San_Francisco)
 - San Francisco officially the City and County of San Francisco, is the leading financial and cultural center of Northern California and the San Francisco Bay Area.
 - San Francisco Bay Area - History of San Francisco - 1906 San Francisco earthquake
- Result 4: **San Francisco Travel Guide: Things to Do, Hotels, Events ...** (www.sanfrancisco.travel/)
 - The official travel and visitors guide for San Francisco. Only in San Francisco can you find San Francisco hotel reservations, tours, flights, maps, popular ...
 - 10 Things Not to Miss in San - Visitor Information Center - San Francisco Events
- Result 5: **San Francisco** (Map)
- Result 6: **San Francisco** (Info)
 - San Francisco, officially the City and County of San Francisco, is the leading financial and cultural center of Northern California and the San Francisco Bay Area. Wikipedia
 - Area: 231.9 sq miles (600.6 km²)
 - Founded: June 29, 1776
 - Weather: 59°F (15°C), Wind NE at 4 mph (6 km/h), 46% Humidity
 - Local time: Sunday 3:55 PM PT
 - Population: 812,826 (2011)

MOTIVATION

“Where is New York City?”



PROBLEM STATEMENT

This is the user sandbox of [Siddhantmanocha](#). A user sandbox is a subpage of the users' user page. It serves as a testing spot.

India

From Wikipedia, the free encyclopedia

This article has multiple issues. Please help improve it or discuss these issues on the talk page.

- This article contains content that is written like an advertisement. (January 2014)
- This article needs additional citations for verification. (January 2014)

The **Indian Institute of Technology Kanpur** (commonly known as **IIT Kanpur** or **IITK**) is a public research college located in Kanpur, Uttar Pradesh. It was declared to be Institute of National Importance by Government of India under IIT Act. It was established in 1959 as one of the first Indian Institutes of Technology, the institute was created with the assistance of a consortium of nine leading US research universities as part of the Kanpur Indo-American Programme (KIAP).^[a]

Contents

- History
- Campus
 - Noida Extension centre
 - Academic Area
 - Students' Activity Centre (SAC)
 - Helicopter Service
- Admissions
- Student life
 - National events
 - Students' Gymkhana
 - Student publications
 - Student voluntary services
 - Convocation
- Rankings
- Academic Bodies and Activities
 - Undergraduate
 - New academic system
 - Postgraduate

Indian Institute of Technology Kanpur

Established 1959

Type Public

Endowment [a]

Director Dr. Indrani Misra

Administrative staff 200^[a]

Undergraduates 3038^[a]

Postgraduates 2540^[a]

Location Kalyanpur, Kanpur, Uttar Pradesh, India

Campus 1055 Acres^[a]

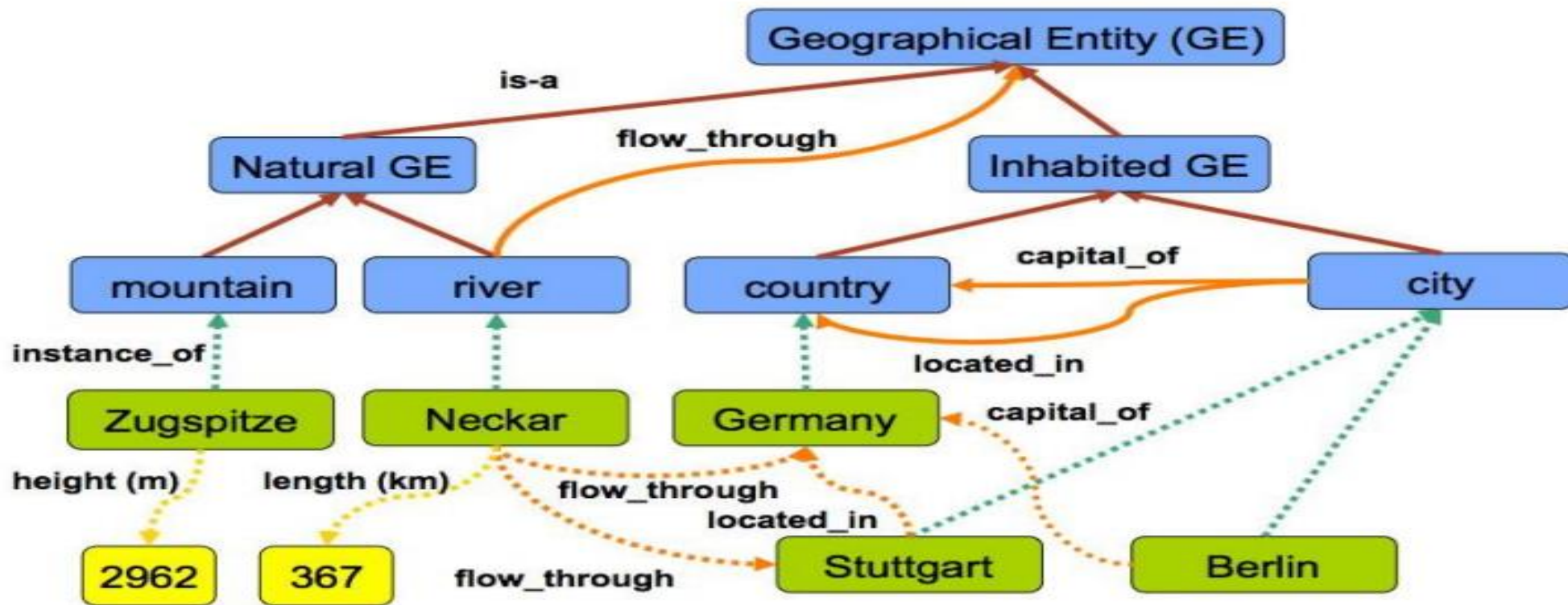
Large Amount Of Unstructured Information out there on the web!

Who is the total area of IIT Kanpur?

What is the capital of India?

Who is the director of IIT Kanpur?

KNOWLEDGE GRAPH



KNOWLEDGE GRAPH



textual abstract:
summary for human

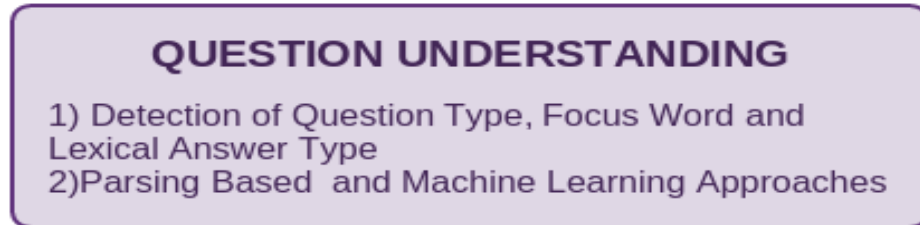


Subject	Relation	Object
p53	is_a	protein
Bax	is_a	protein
p53	has_function	apoptosis
Bax	has_function	induction
apoptosis	involved_in	cell_death
Bax	is_in	mitochondrial outer membrane
Bax	is_in	cytoplasm
apoptosis	related_to	caspase activation
...

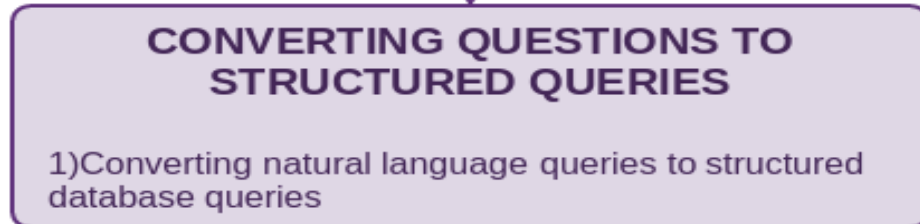
structured knowledge extraction:
summary for machine



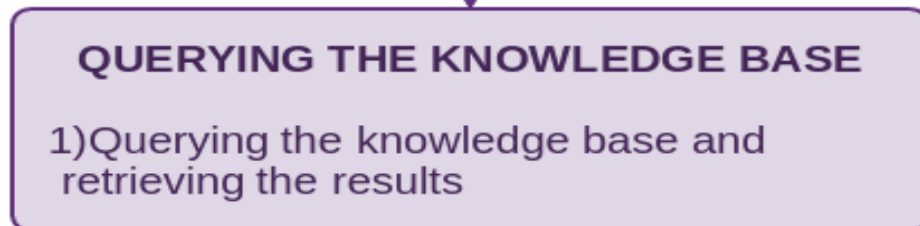
QUESTION ANSWERING



Who is the author of Julius Caesar?
<Who,Person> <Author,Writer>
<Julius Caesar, Book>



Select author from library_db where
book="Julius Caesar"



Book: Julius Caesar
Author: William Shakespeare

EXISTING KNOWLEDGE BASES

Knowledge graphs

 **Freebase**

 **yAGO**
select knowledge

 **DBpedia**

Facebook's
Entity Graph



Microsoft's
Satori



OpenIE
(*Reverb, OLLIE*)

Google's
Knowledge Graph

EXISTING KNOWLEDGE BASES

Supervised Models:

- Learn classifiers from +/- examples, typical features: context words + POS, dependency path between entities, named entity tags
- Require large number of tagged training examples
- Cannot be generalized

Semi-Supervised Models:

- Bootstrap Algorithms: Use seed examples to learn initial set of relations
- Generate +ve/-ve examples to learn a classifier
- Learn more relations using this classifier

Distant Supervision:

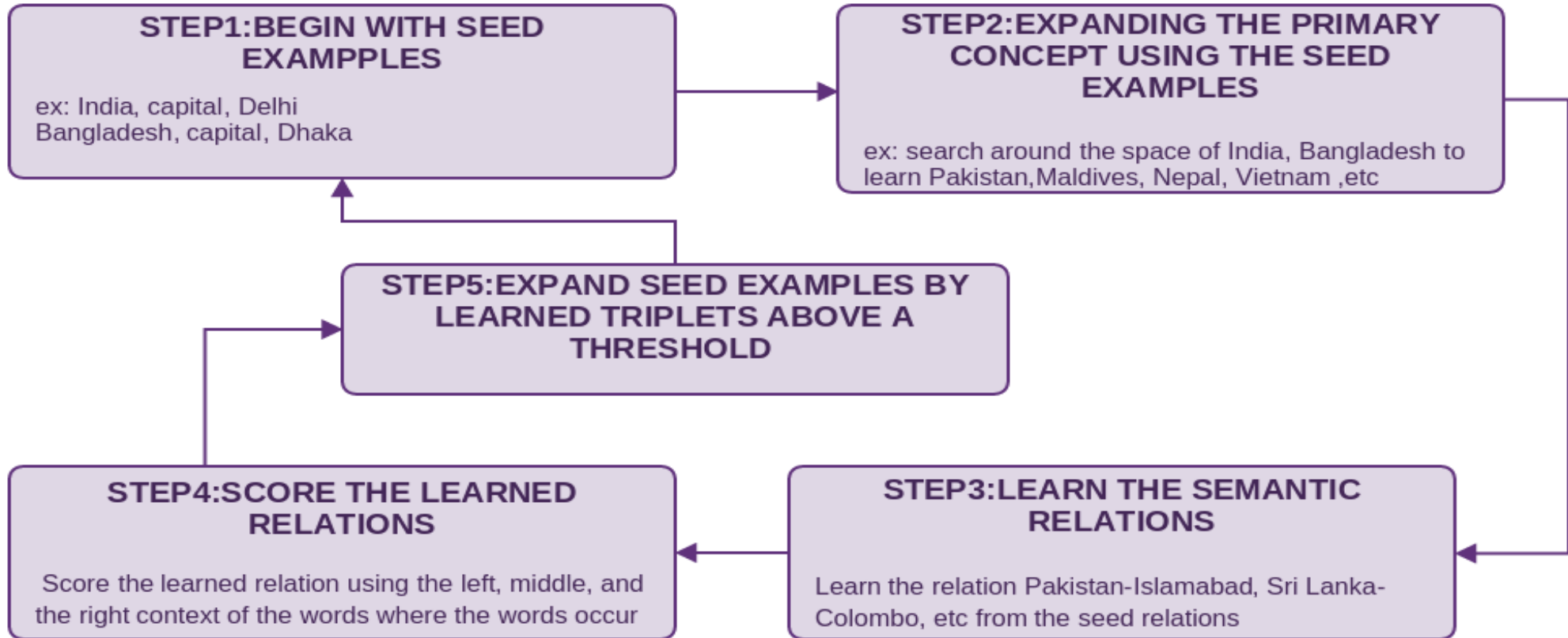
- Existing knowledge base + unlabeled text generate examples
- Learn models using this set of relations

OUR APPROACH

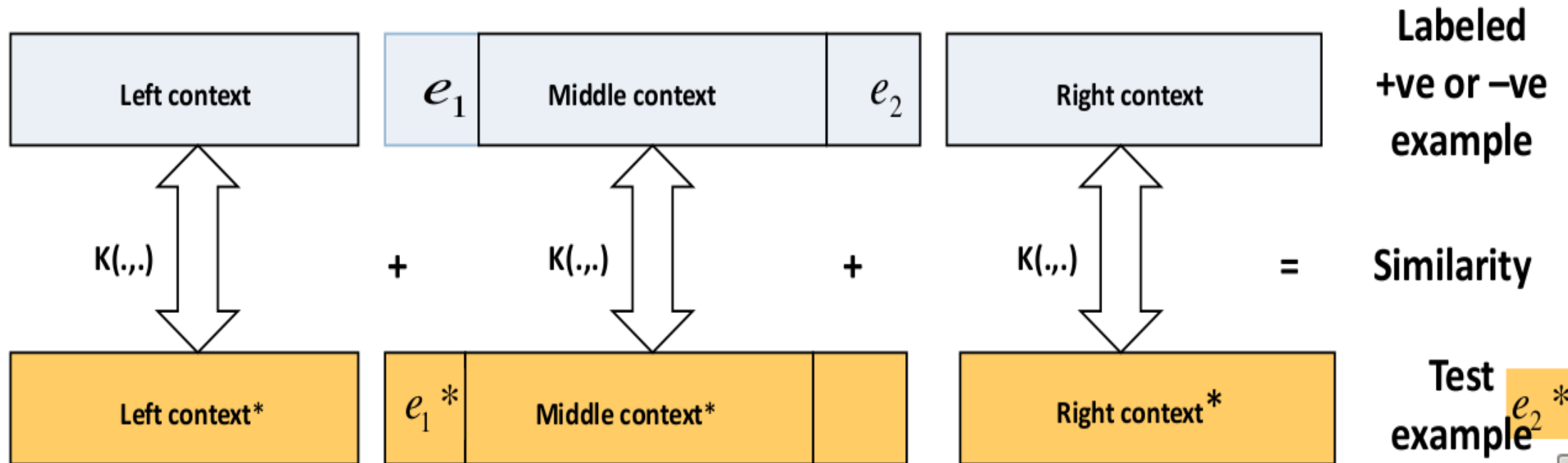
Bootstrapping Relations using Distributed Word Vector Embedding

- 1) Word that occur in similar context lie close together in the word embedding space.
- 2) Word Vectors is semantically consistent and capture many linguistic properties (like 'capital city', 'native language', 'plural relations')
- 3) Obtain word vectors from unstructured text (using Google word2vec, Glove, etc)
- 4) Exploit the properties of the manifold to obtain binary relations between entities

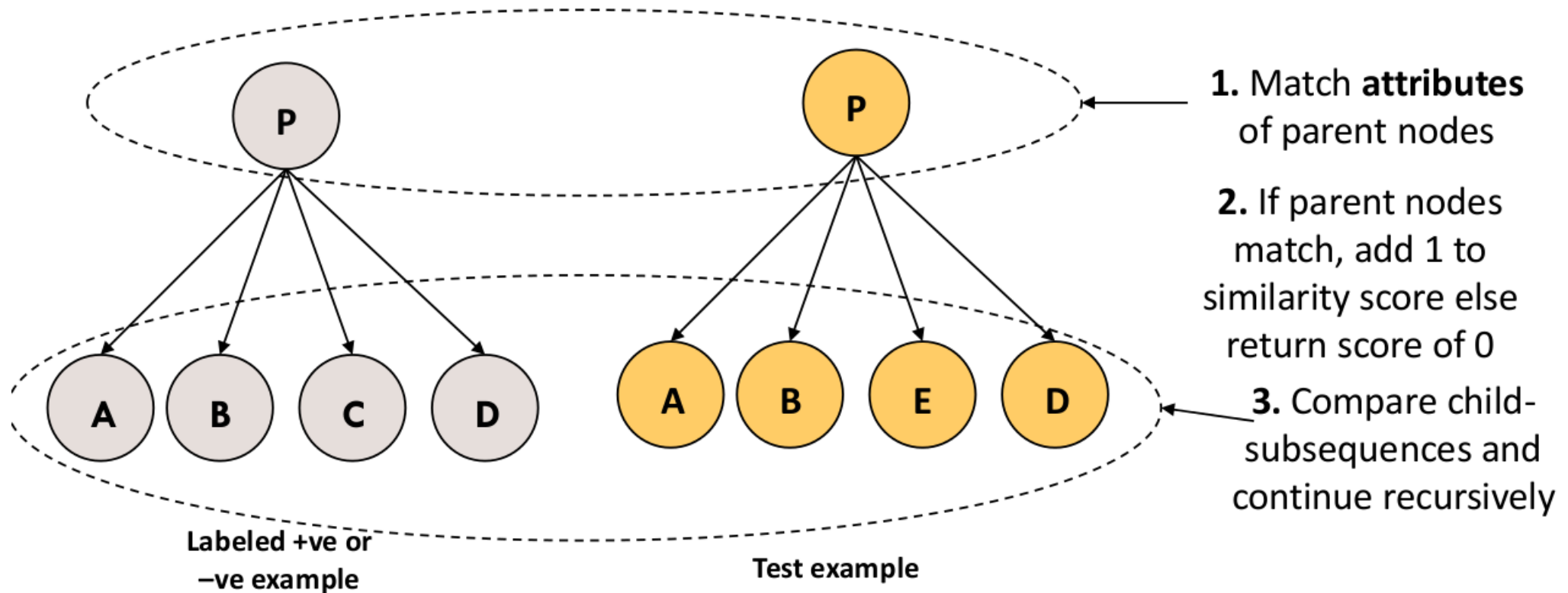
ALGORITHM



SIMILARITY METRIC



KERNEL BASED APPROACHES



DEPENDENCY KERNELS

1. 'his actions in Brcko', and

1. 'his → actions ← in ← Brcko', and

2. 'his arrival in Beijing'.

2. 'his → arrival ← in ← Beijing'.

1. Actual Sentences

2. Dependency Graph

1. $\mathbf{x} = [x_1 \ x_2 \ x_3 \ x_4 \ x_5 \ x_6 \ x_7]$, where $x_1 = \{\text{his, PRP, PERSON}\}$, $x_2 = \{\rightarrow\}$, $x_3 = \{\text{actions, NNS, Noun}\}$, $x_4 = \{\leftarrow\}$, $x_5 = \{\text{in, IN}\}$, $x_6 = \{\leftarrow\}$, $x_7 = \{\text{Brcko, NNP, Noun, LOCATION}\}$

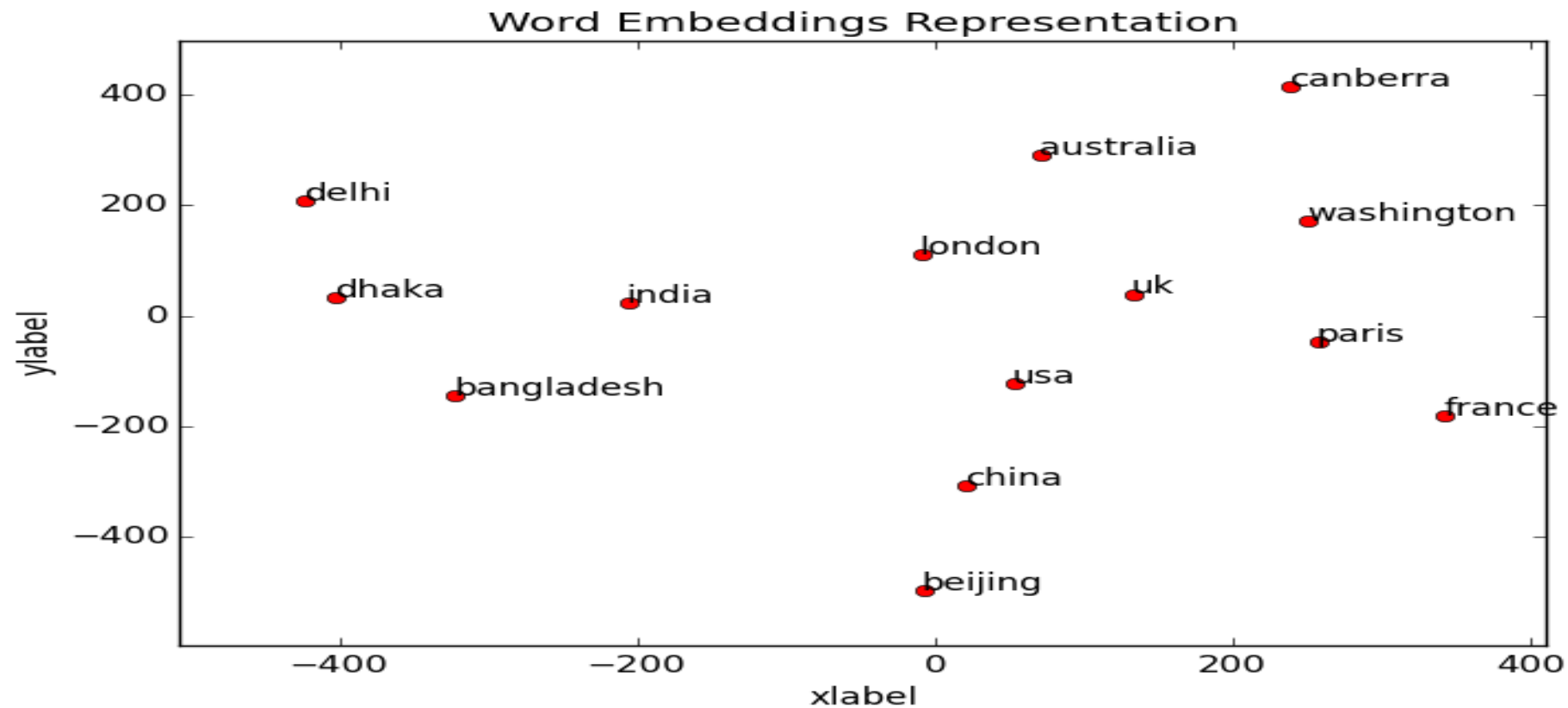
2. $\mathbf{y} = [y_1 \ y_2 \ y_3 \ y_4 \ y_5 \ y_6 \ y_7]$, where $y_1 = \{\text{his, PRP, PERSON}\}$, $y_2 = \{\rightarrow\}$, $y_3 = \{\text{arrival, NN, Noun}\}$, $y_4 = \{\leftarrow\}$, $y_5 = \{\text{in, IN}\}$, $y_6 = \{\leftarrow\}$, $y_7 = \{\text{Beijing, NNP, Noun, LOCATION}\}$

Kernel:

$$K(\mathbf{x}, \mathbf{y}) = 3 \times 1 \times 1 \times 1 \times 2 \times 1 \times 3 = 18$$

3. Kernel Computation

PRELIMINARY RESULTS



PRELIMINARY RESULTS (wikipedia corpus)

Seed Examples for capital relationship

Country	Capital
India	Delhi
Bangladesh	Dhaka

Positive relations learnt

Country	Capital
Nepal	Kathmandu
Afghanistan	Kabul
Thailand	Bangkok
Russia	Moscow

Negative Relations learnt

Country	Capital
Bhutan	Sikkim
Algeria	Tunisia
Burma	Jalpaiguri
Kuwait	Cairo

PRELIMINARY RESULTS_(google news corpus)

Seed Examples

Country	Capital
India	Delhi
Bangladesh	Dhaka

Positive Relations Learned

Country	Capital
Nepal	Kathmandu
Pakistan	Islamabad

Negative Relations Learned

Country	Capital
Srilanka	Tamil
Bhutan	Sikkim
Burma	Jalpaiguri
LTTE	tamil

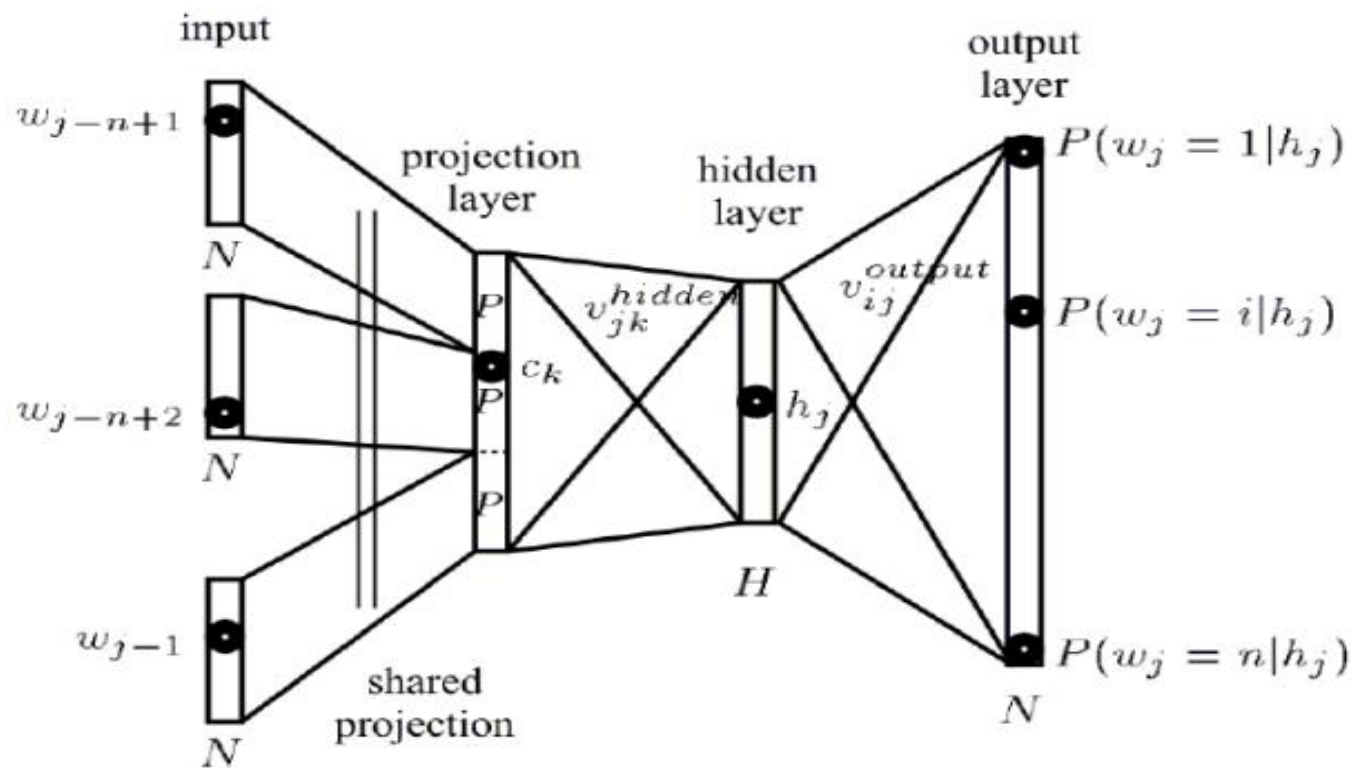
References

- 1) Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic Regularities in Continuous Space Word Representations. In Proceedings of NAACL HLT, 2013.
- 2) Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. [Distributed Representations of Words and Phrases and their Compositionality](#). In Proceedings of NIPS, 2013.
- 3) Eugene Agichtein Luis Gravano. Snowball: Extracting Relations from Large Plain-Text Collections. In *Proceedings of the fifth ACM conference on Digital libraries*, June 2000

Questions!



CBOW MODEL



- input vector represented as 1-of- V encoding
- Linear sum of input vectors are projected onto the projection layer
- Hierarchical Softmax layer is used to ensure that the weights in the output layer are between $0 \leq p \leq 1$
- Weights learnt using back-propagation
- The projection matrix from the projection layer to the hidden layer give the word vector embeddings

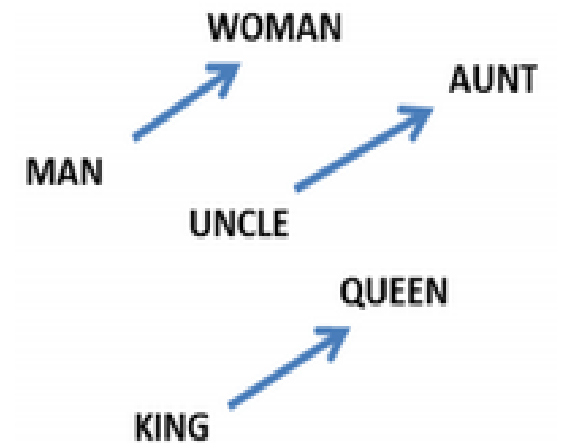
WORD VECTOR MODEL

FRANCE	JESUS	XBOX	REDDISH	SCRATCHED	MEGABITS
AUSTRIA	GOD	AMIGA	GREENISH	NAILED	OCTETS
BELGIUM	SATI	PLAYSTATION	BLUISH	SMASHED	MB/S
GERMANY	CHRIST	MSX	PINKISH	PUNCHED	BIT/S
ITALY	SATAN	IPOD	PURPLISH	POPPED	BAUD
GREECE	KALI	SEGA	BROWNISH	CRIMPED	CARATS
SWEDEN	INDRA	PSNUMBER	GREYISH	SCRAPED	KBIT/S
NORWAY	VISHNU	HD	GRAYISH	SCREWED	MEGAHERTZ
EUROPE	ANANDA	DREAMCAST	WHITISH	SECTIONED	MEGAPIXELS
HUNGARY	PARVATI	GEFORCE	SILVERY	SLASHED	GBIT/S
SWITZERLAND	GRACE	CAPCOM	YELLOWISH	RIPPED	AMPERES

What words have embeddings closest to a given word? From Collobert
et al. (2011)

WORD VECTOR MODEL

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza



From Mikolov *et al.*
(2013a)

Relationship pairs in a word embedding. From Mikolov *et al.* (2013b).

KERNEL BASED APPROACHES

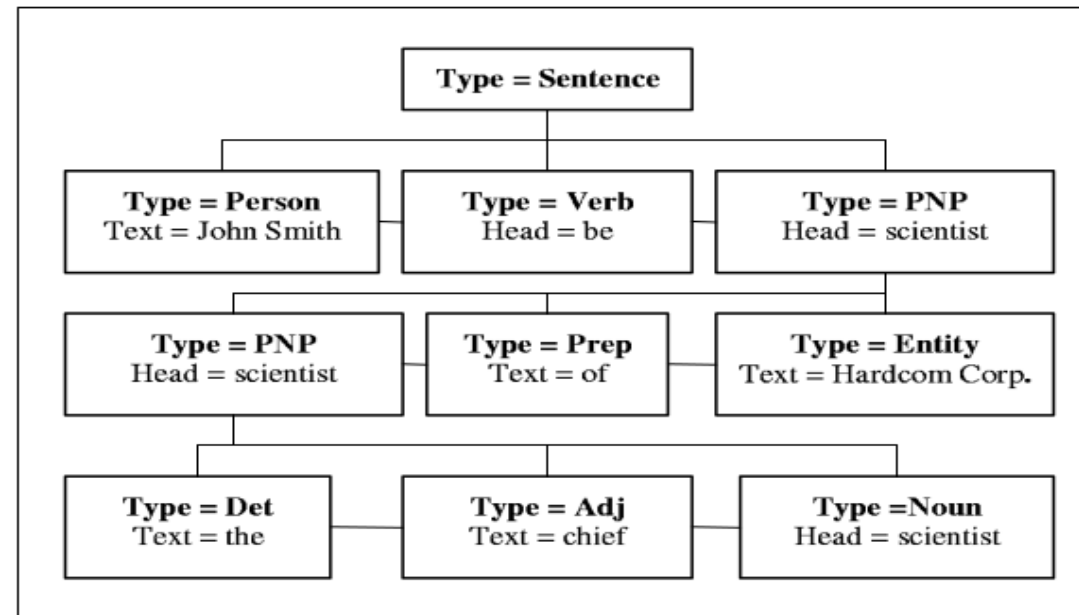


Figure 1: The shallow parse representation of the the sentence “John Smith is the chief scientist of the Hardcom Corporation”.The types “PNP”, “Det”, “Adj”, and “Prep” denote “Personal Noun Phrase”, “Determiner”, “Adjective”, and “Preposition”, respectively.