Constructing Knowledge Graph from Unstructured Text

Kundan Kumar

Siddhant Manocha

Image Source: www.ibm.com/smarterplanet/us/en/ibmwatson/

MOTIVATION

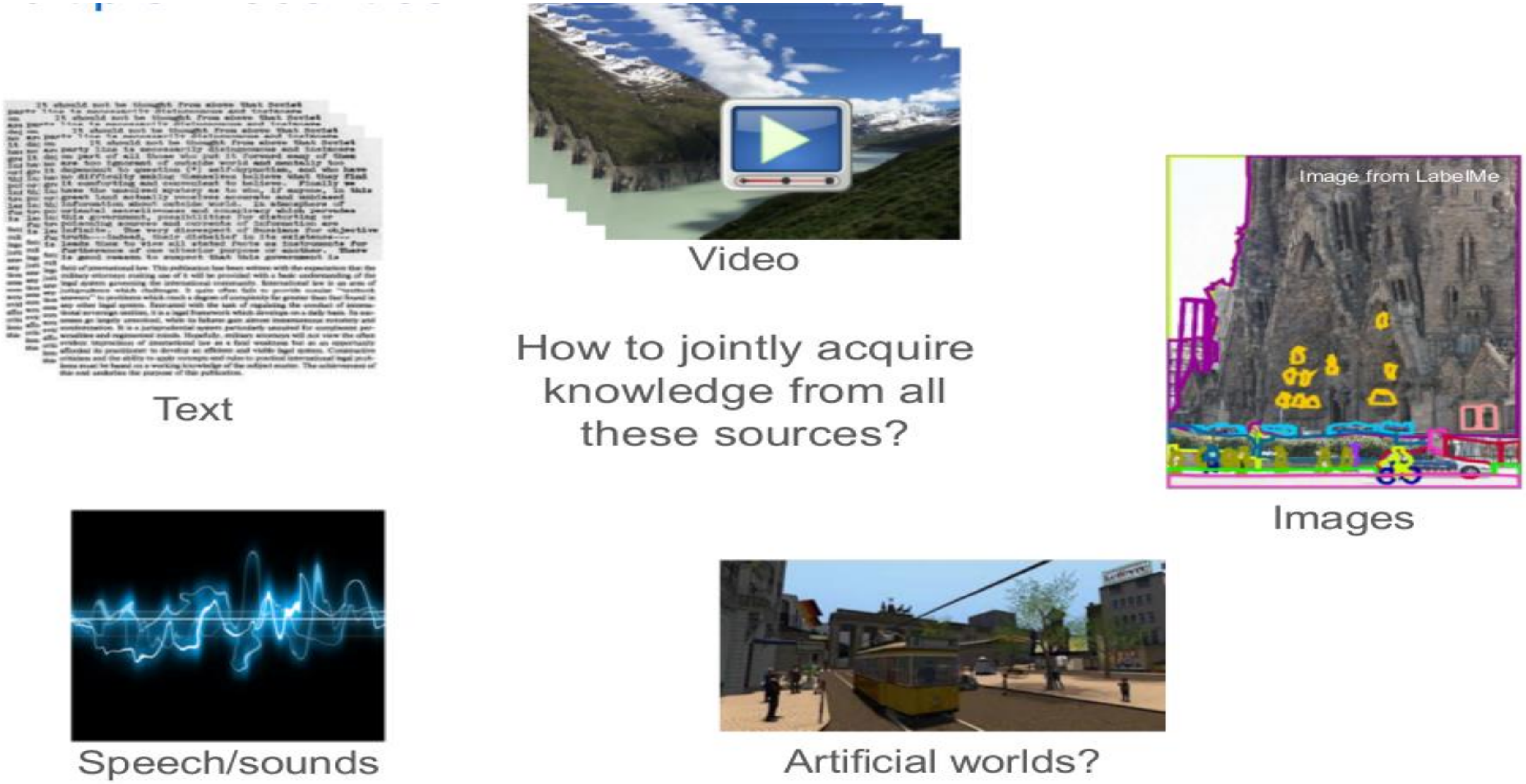


Image Source: KDD 2014 Tutorial on Constructing and Mining Web-scale Knowledge Graphs, New York

MOTIVATION

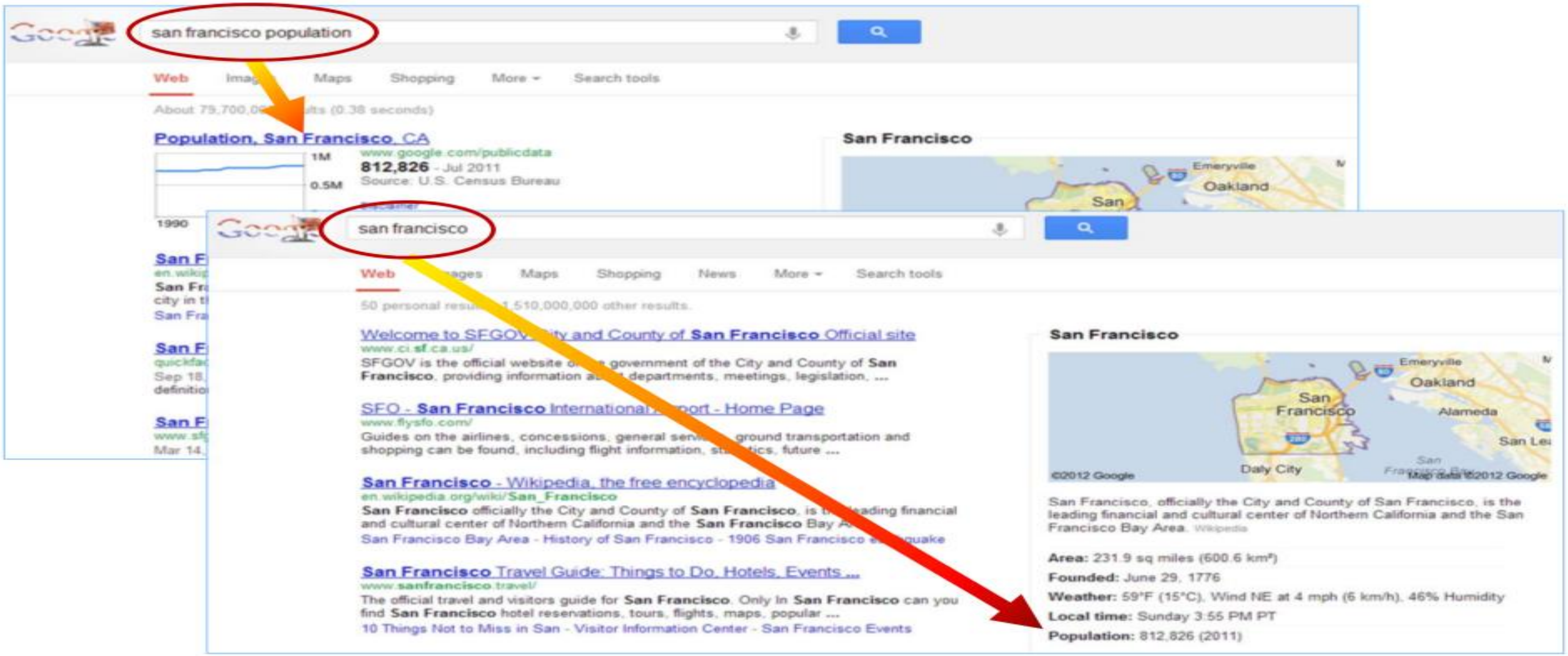


Image Source: KDD 2014 Tutorial on Constructing and Mining Web-scale Knowledge Graphs, New York

MOTIVATION

"Where is New York City?"

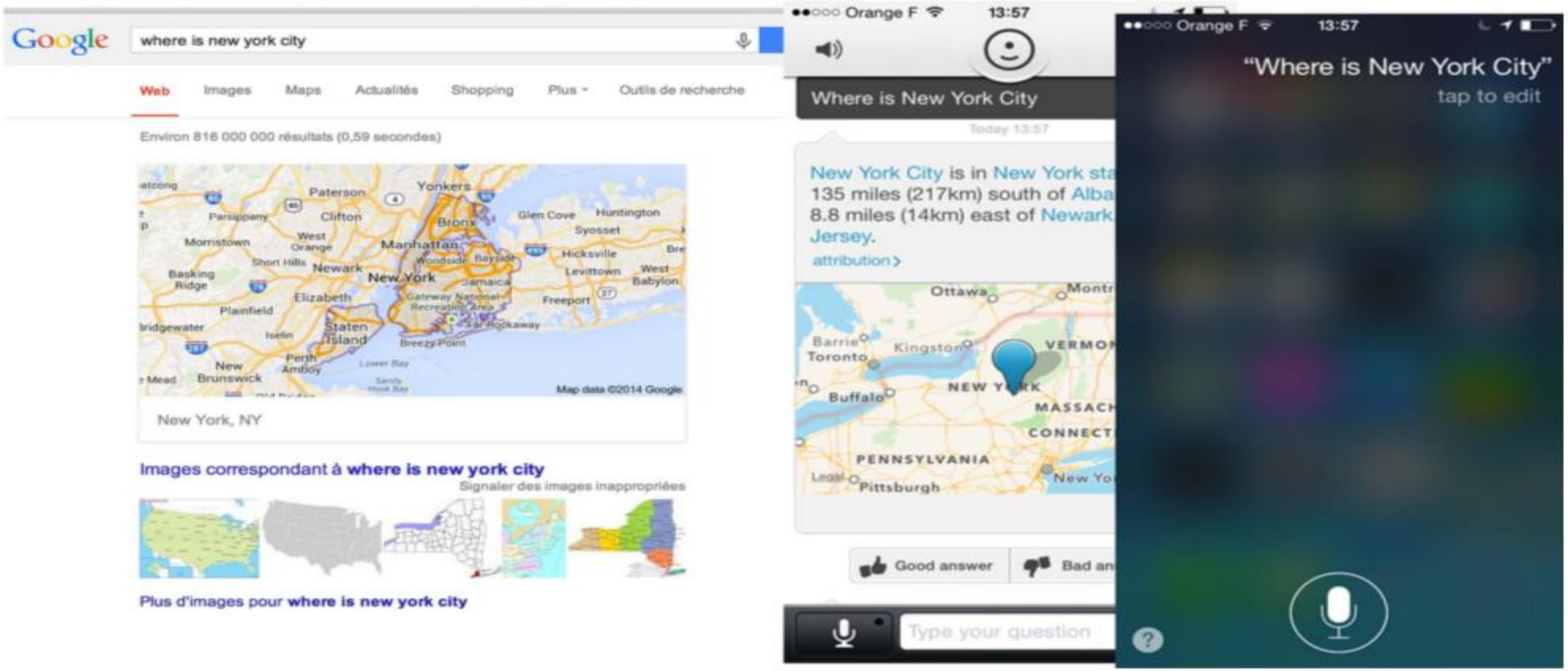
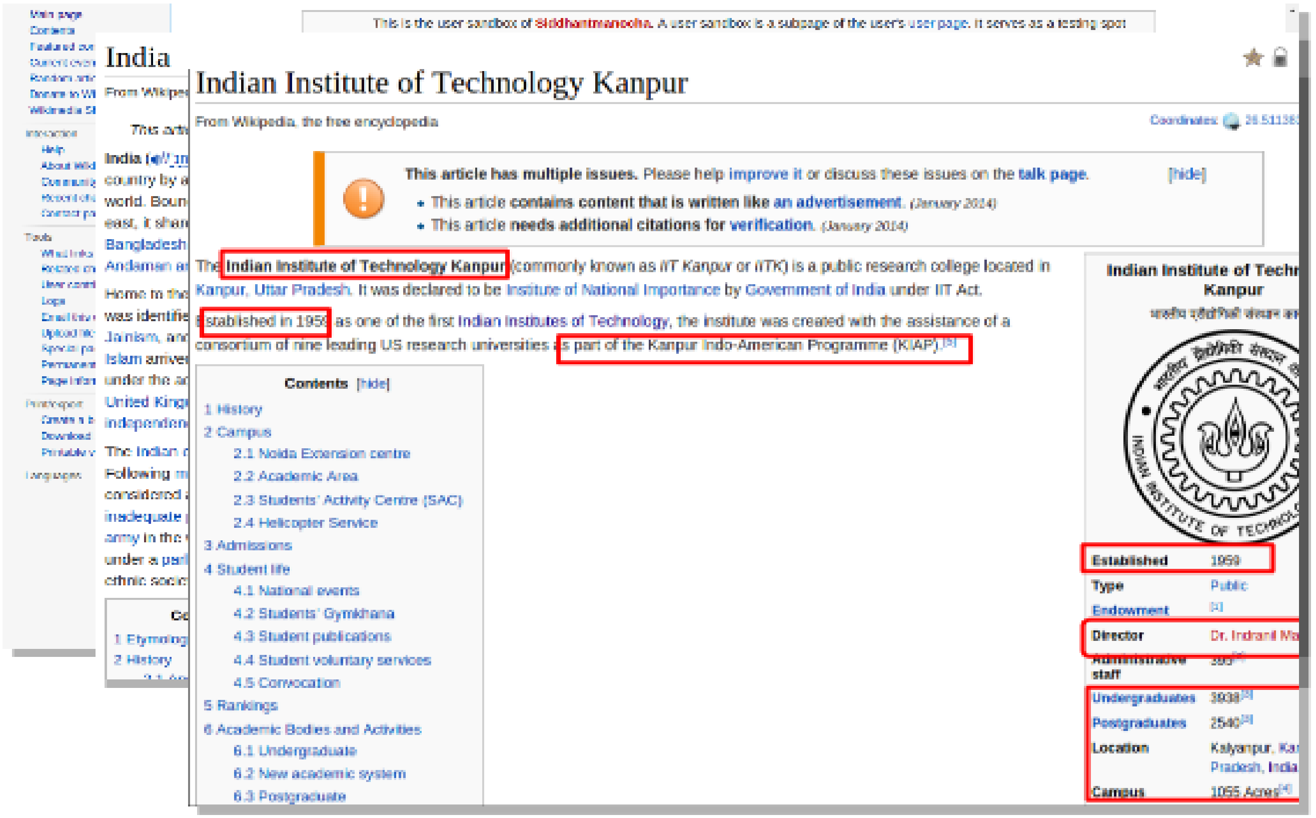


Image Source: KDD 2014 Tutorial on Constructing and Mining Web-scale Knowledge Graphs, New York

PROBLEM STATEMENT



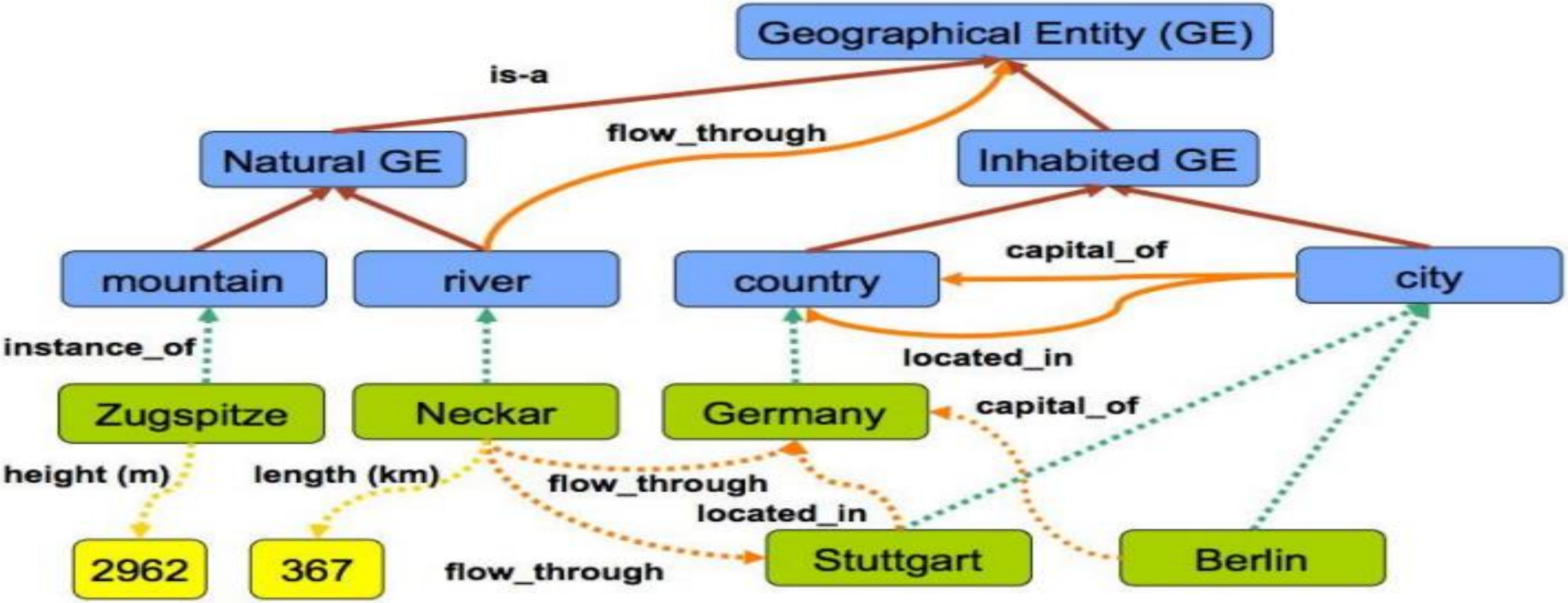
Large Amount Of Unstructured Information out there on the web!

Who is the total area of IIT Kanpur?

What is the capital of India?

Who is the director of IIT Kanpur?

KNOWLEDGE GRAPH



http://courses.cs.washington.edu/courses/cse517/13wi/slides/cse517wi13-RelationExtraction.pdf

KNOWLEDGE GRAPH

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
...	...	...

textual abstract:
summary for human

structured knowledge extraction:
summary for machine

http://courses.cs.washington.edu/courses/cse517/13wi/slides/cse517wi13-RelationExtraction.pdf

QUESTION ANSWERING

QUESTION UNDERSTANDING

1) Detection of Question Type, Focus Word and Lexical Answer Type
2)Parsing Based and Machine Learning Approaches

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

CONVERTING QUESTIONS TO STRUCTURED QUERIES

1)Converting natural language queries to structured database queries

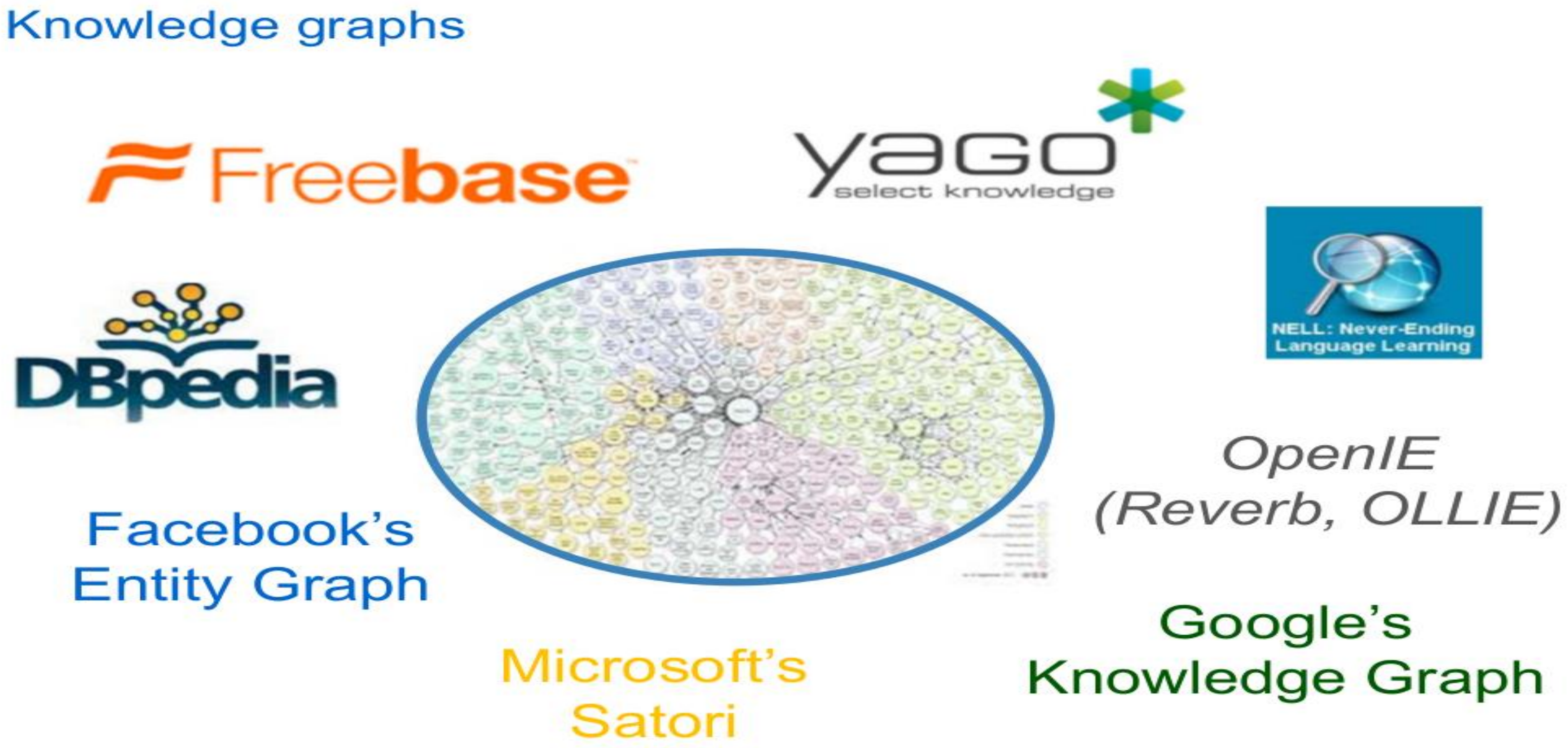
Select author from library_db where book="Julius Caesar"

QUERYING THE KNOWLEDGE BASE

1)Querying the knowledge base and retrieving the results

Book: Julius Caesar
Author: William Shakespeare

EXISTING KNOWLEDGE BASES



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

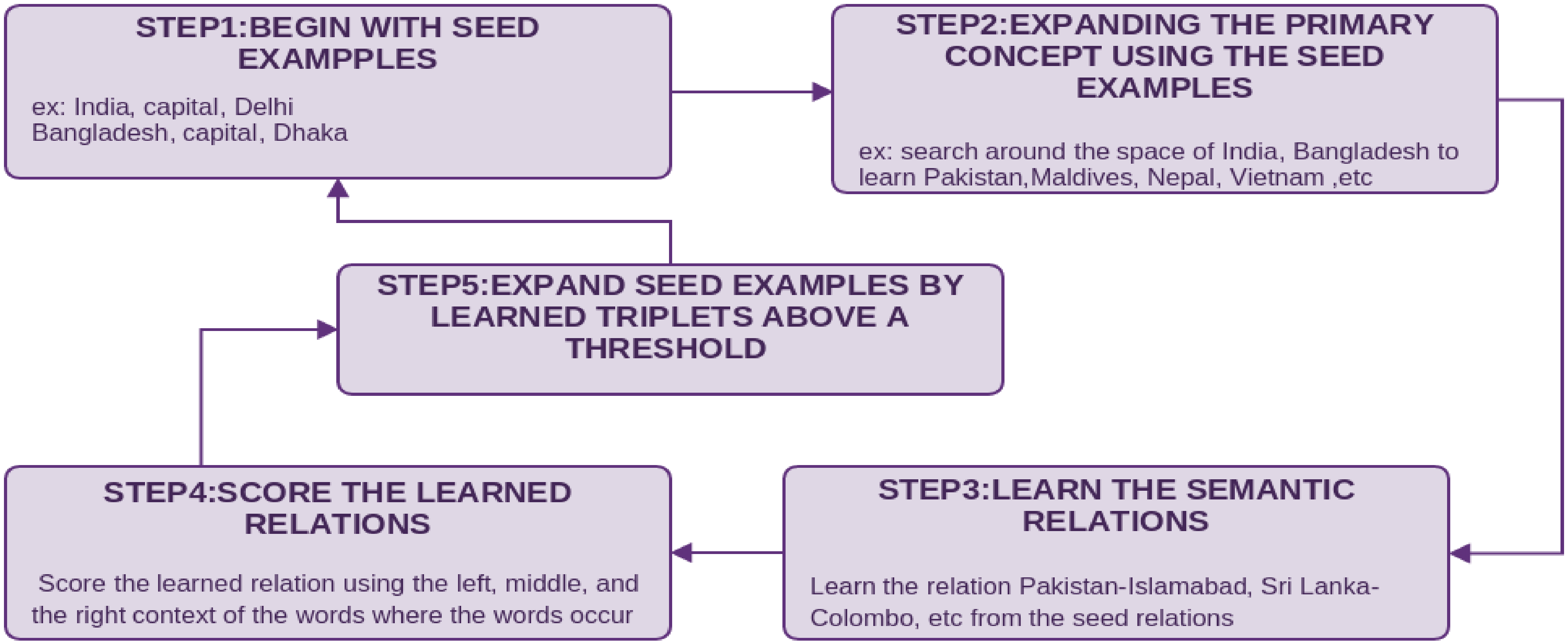
1) Word that occur in similar context lie close together in the word embedding space.

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


STEP1:BEGIN WITH SEED EXAMPPLES
ex: India, capital, Delhi
Bangladesh, capital, Dhaka
STEP2:EXPANDING THE PRIMARY CONCEPT USING THE SEED EXAMPLES
ex: search around the space of India, Bangladesh to learn Pakistan,Maldives, Nepal, Vietnam ,etc
STEP3:LEARN THE SEMANTIC RELATIONS
Learn the relation Pakistan-Islamabad, Sri Lanka-Colombo, etc from the seed relations
STEP4:SCORE THE LEARNED RELATIONS
Score the learned relation using the left, middle, and the right context of the words where the words occur
STEP5:EXPAND SEED EXAMPLES BY LEARNED TRIPLETS ABOVE A THRESHOLD

SIMILARITY METRIC

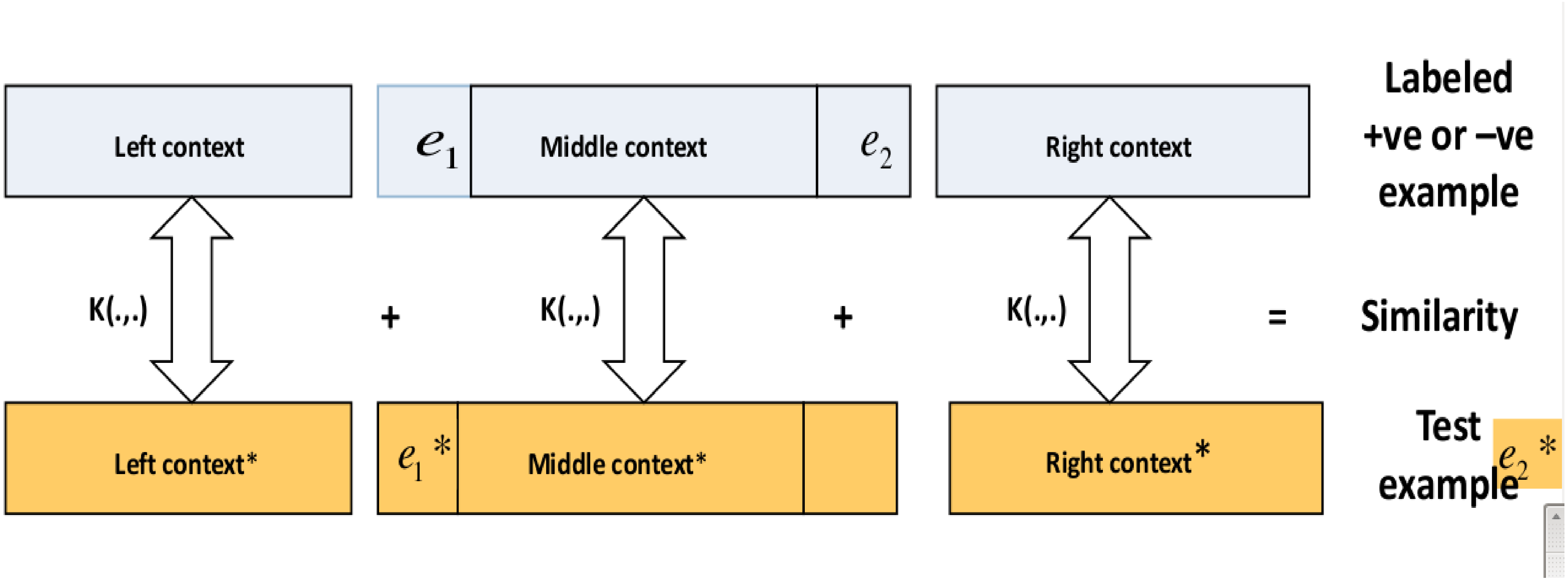
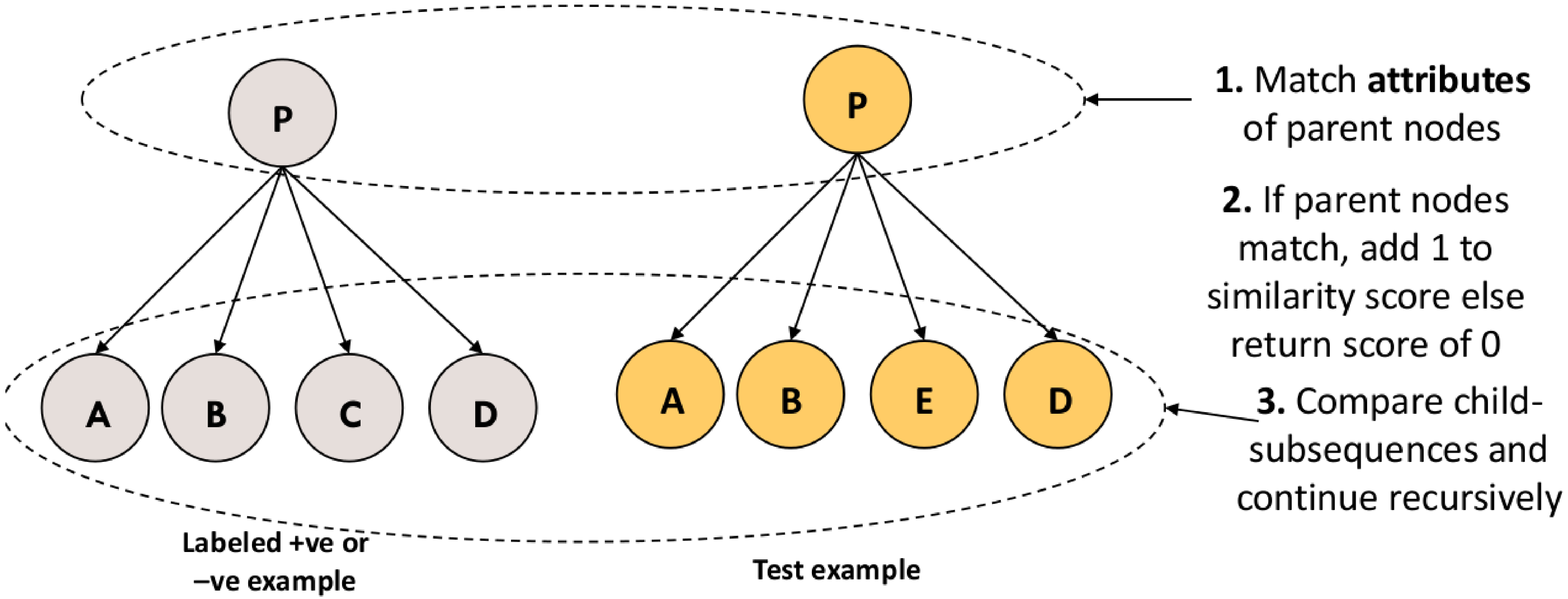


Image Source:A survey on relation extraction, Nguyen Bach, Carnegie Mellon University

KERNEL BASED APPROACHES



DEPENDENCY KERNELS

1. '**his** actions in **Brcko**', and
2. '**his** arrival in **Beijing**'.

1.Actual Sentences

1. '**his** $\rightarrow$ actions $\leftarrow$ in $\leftarrow$ **Brcko**', and
2. '**his** $\rightarrow$ arrival $\leftarrow$ in $\leftarrow$ **Beijing**'.

2. Dependency Graph

1. $\boldsymbol{x} = [x_1\ x_2\ x_3\ x_4\ x_5\ x_6\ x_7]$, where $x_1 =$ {his, PRP, PERSON}, $x_2 = \{\rightarrow\}$, $x_3 =$ {actions, NNS, Noun}, $x_4 = \{\leftarrow\}$, $x_5 =$ {in, IN}, $x_6 = \{\leftarrow\}$, $x_7 =$ {Brcko, NNP, Noun, LOCATION}
2. $\boldsymbol{y} = [y_1\ y_2\ y_3\ y_4\ y_5\ y_6\ y_7]$, where $y_1 =$ {his, PRP, PERSON}, $y_2 = \{\rightarrow\}$, $y_3 =$ {arrival, NN, Noun}, $y_4 = \{\leftarrow\}$, $y_5 =$ {in, IN}, $y_6 = \{\leftarrow\}$, $y_7 =$ {Beijing, NNP, Noun, LOCATION}

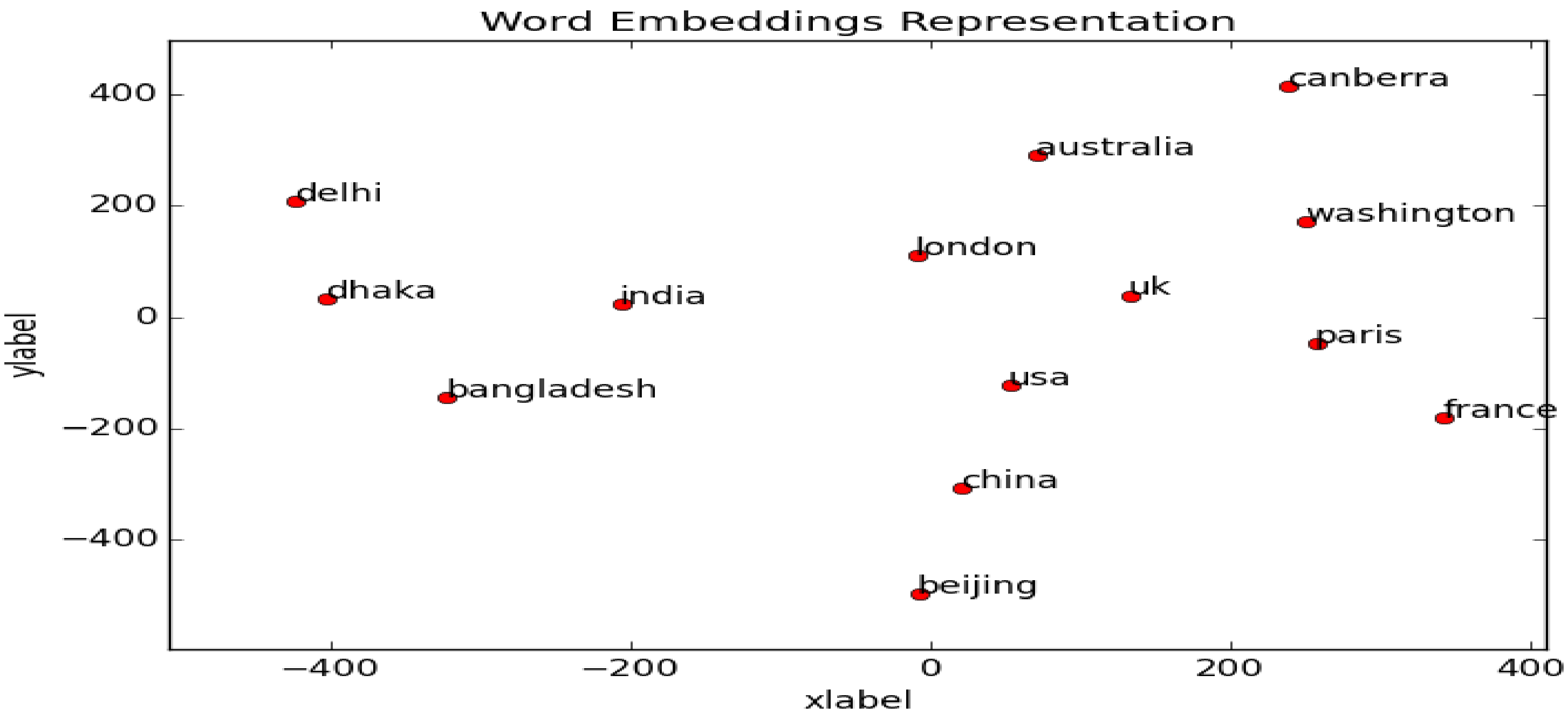
3.Kernel Computation

Kernel:

$K(x,y)=3\times1\times1\times1\times2\times1\times3$

$= 18$

PRELIMINARY RESULTS



Word Vector Embedding: Wikipedia Corpus

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

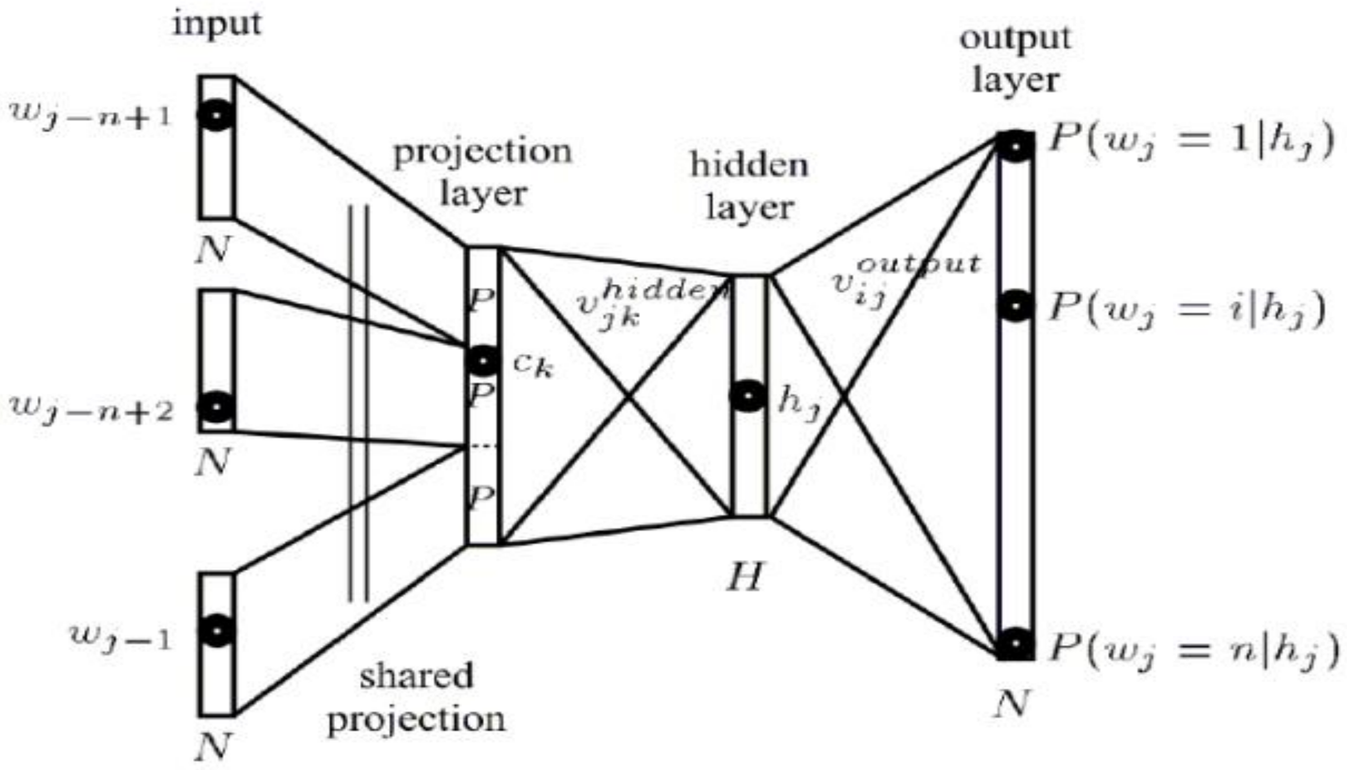
1)Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic Regularities in Continuous Space Word Representations. In Proceedings of NAACL HLT, 2013.

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<=p<=1
- Weights learnt using back-propagation
- The projection matrix from the projection layer to the hidden layer give the word vector embeddings

Image Source: Linguistic Regularities in Continuous Space Word Representations,Mikolov,et.al 2013

WORD VECTOR MODEL

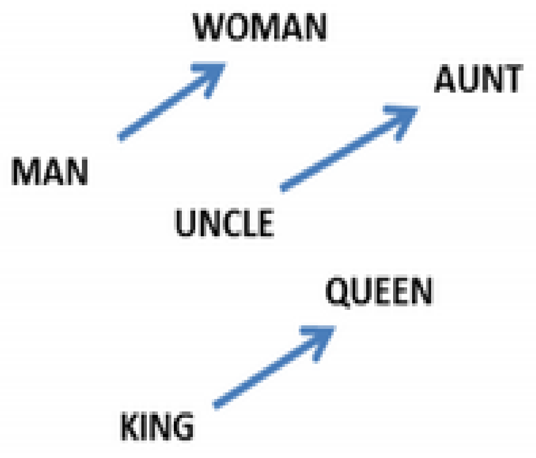
FRANCE	JESUS	XBOX	REDDISH	SCRATCHED	MEGABITS
AUSTRIA	GOD	AMIGA	GREENISH	NAILED	OCTETS
BELGIUM	SATI	PLAYSTATION	BLUISH	SMASHED	MB/S
GERMANY	CHRIST	MSX	PINKISH	PUNCHED	BIT/S
ITALY	SATAN	IPOD	PURPLISH	POPPED	BAUD
GREECE	KALI	SEGA	BROWNISH	CRIMPED	CARATS
SWEDEN	INDRA	PSNUMBER	GREYISH	SCRAPED	KBIT/S
NORWAY	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

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



From Mikolov *et al.* (2013a)

KERNEL BASED APPROACHES

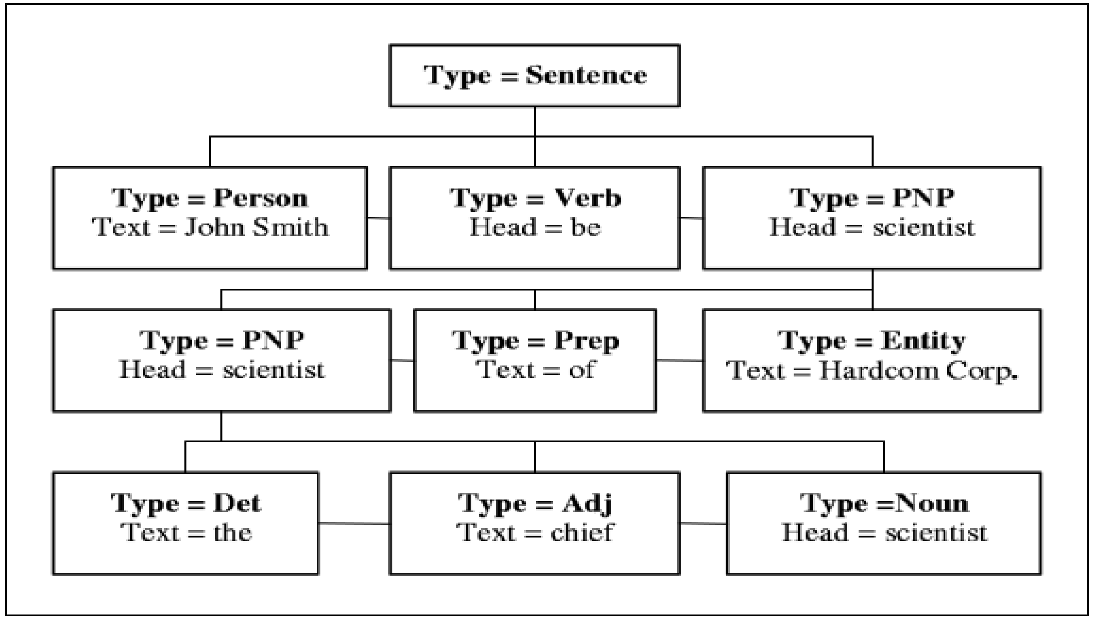


Figure 1: The shallow parse representation of the the sentence "John Smith is the chief scientist of the Hardcom Corporation".The types "PNP", "Det", "Adj", and "Prep" denote "Personal Noun Phrase", "Determiner", "Adjective", and "Preposition", respectively.

Image Source:Kernel Methods for Relation Extraction,Zelenko,et al