Knowledge Semantic Representation

Han Xiao, Minlie Huang, Xiaoyan Zhu

Tsinghua University

October 26, 2016

Website: `http:\\ciia.cs.tsinghua.edu.cn`

Home Page: `http:\\www.ibookman.net`

Knowledge Semantic Representation

- **Motivations.** Geometrical positions as knowledge representation could hardly explicitly indicate the semantics.
  - The representation of the entity *Table* in TransE:

$$(0.12, -0.22, 0.55, 0.60, 0.71, -0.01, 0.00, -0.77...)$$

- **Could we tell about something semantic?**
  - being a furniture?
  - being a daily tool?
  - being not an animal?
- The **GAP** between knowledge and language remains.
- Thus, developing a *semantics-specific representation* triggers an urgent task.
- A well-fitting model for knowledge graph is encouraging, but is still insufficient for pragmatic applications.

Knowledge Semantic Representation

- **Knowledge Semantic Analysis (KSA)**
  - **Definitions**: A knowledge representation methodology that is supposed to explicitly provide human-comprehensive or at least semantics-relevant representation.
  - *(Stanford University) =*
    (University:Yes, Animal:No, Location:California, ...)
- **Knowledge Feature** is a term we introduced for describing semantic aspects of knowledge.
- **Benefits**
  - The trade-off between *Human-Comprehensive* and *Machine-Computational* Knowledge Representation.
  - At least, in this way, it is more elegant to joint multiple information sources and knowledge triples.

Knowledge Semantic Representation

- **A Naive Example** in the scenario of information retrieval.
- Query: *What private university is most famous in California?*
  1. Extracting the keywords: *private, university, famous, California*.
  2. Mapping to knowledge feature: *(University:Yes, Animal:No, Location:California, Type:Private, Famous:Very, ...)*.
  3. inferring the possible entity/relation *(Stanford University)* as the answer with link prediction task.
- Notably, our model **KSR** is a generative model, which could generate the representations, while is also capable to infer the entities/relations.

Knowledge Semantic Representation

- **Model Descriptions** KSR leverages a two-level hierarchical generative process to semantically represent the entities, relations and triples.

For each triple $(h, r, t) \in \Delta$*:*
(First-Level)
Draw a knowledge feature f_i *from* $\mathcal{P}(f_i|r)$*:*

1. **(Second-Level)**
   Draw a subject-specific category z_i *from*

$$\mathcal{P}(z_i) \propto \mathcal{P}(z_i|h)\mathcal{P}(z_i|r)\mathcal{P}(z_i|t, f_i)$$

2. **(Second-Level)**
   Draw an object-specific category y_i *from*

$$\mathcal{P}(y_i) \propto \mathcal{P}(y_i|t)\mathcal{P}(y_i|r)\mathcal{P}(y_i|z_i, f_i)$$

Knowledge Semantic Representation

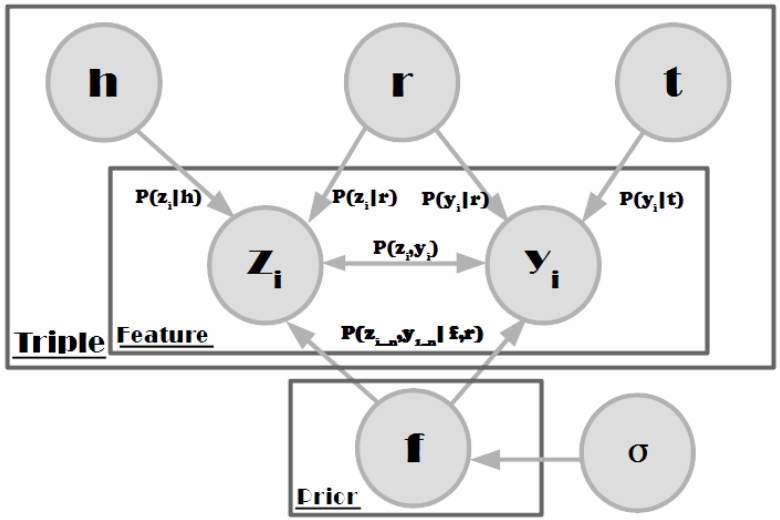
- **Model Descriptions** KSR leverages a two-level hierarchical generative process to semantically represent the entities, relations and triples.
  1. In the first level of our model, we generate some *knowledge features* such as *University(YES/NO), Animal Type, Location, etc*.
  2. In the second level of our model, we assign a corresponding category in each knowledge feature for every triple.
- For the example of *Stanford University*, we assign *Yes* in the *University* feature, *California* in *Location* feature and so on.

$$(University : Yes, Animal : No, Location : California, \\ Type : Private, Famous : Very, ...)$$

- In this way, the knowledge representation is semantically interpretable.

Knowledge Semantic Representation

- **Probabilistic Graph Model**



$$[h, r, t, z_k, y_k | f_k, \sigma] = [z_k|h][z_k|r][y_k|t][y_k|r][z_k, y_k|f_k, r, \sigma]$$

$$[h, r, t] = \sum_{k=1}^{n} [f_k|\sigma] \left\{ \sum_{i,j=1}^{d} [h, r, t, z_k = i, y_k = j | f_k, \sigma] \right\}$$

$$= \overbrace{\sum_{k=1}^{n} [f_k|\sigma] \overbrace{\left\{ \sum_{i,j=1}^{d} [z_k = i, y_k = j | f_k, \sigma][h, r, t | z_k = i, y_k = j, f_k, \sigma] \right\}}^{\text{Second—Level:Category Mixture}}}^{\text{First—Level:Feature Mixture}}$$

Knowledge Semantic Representation

- **Semantic Representation** are generated by adopting the most possible category in the specific knowledge features as the semantic representation.

$$\mathbf{S_e} = (\mathbf{S_{e,1}}, \mathbf{S_{e,2}}, ..., \mathbf{S_{e,n}})$$

$$\mathbf{S_{e,i}} = \arg\max_{\mathbf{c=1}}^{\mathbf{d}}[\mathbf{z_i} = \mathbf{c}|\mathbf{e}]$$

$$\mathbf{S_r} = (\mathbf{S_{r,1}}, \mathbf{S_{r,2}}, ..., \mathbf{S_{r,n}})$$

$$\mathbf{S_{r,i}} = \arg\max_{\mathbf{c=1}}^{\mathbf{d}}[\mathbf{z_i} = \mathbf{c}|\mathbf{r}][\mathbf{y_i} = \mathbf{c}|\mathbf{r}]$$

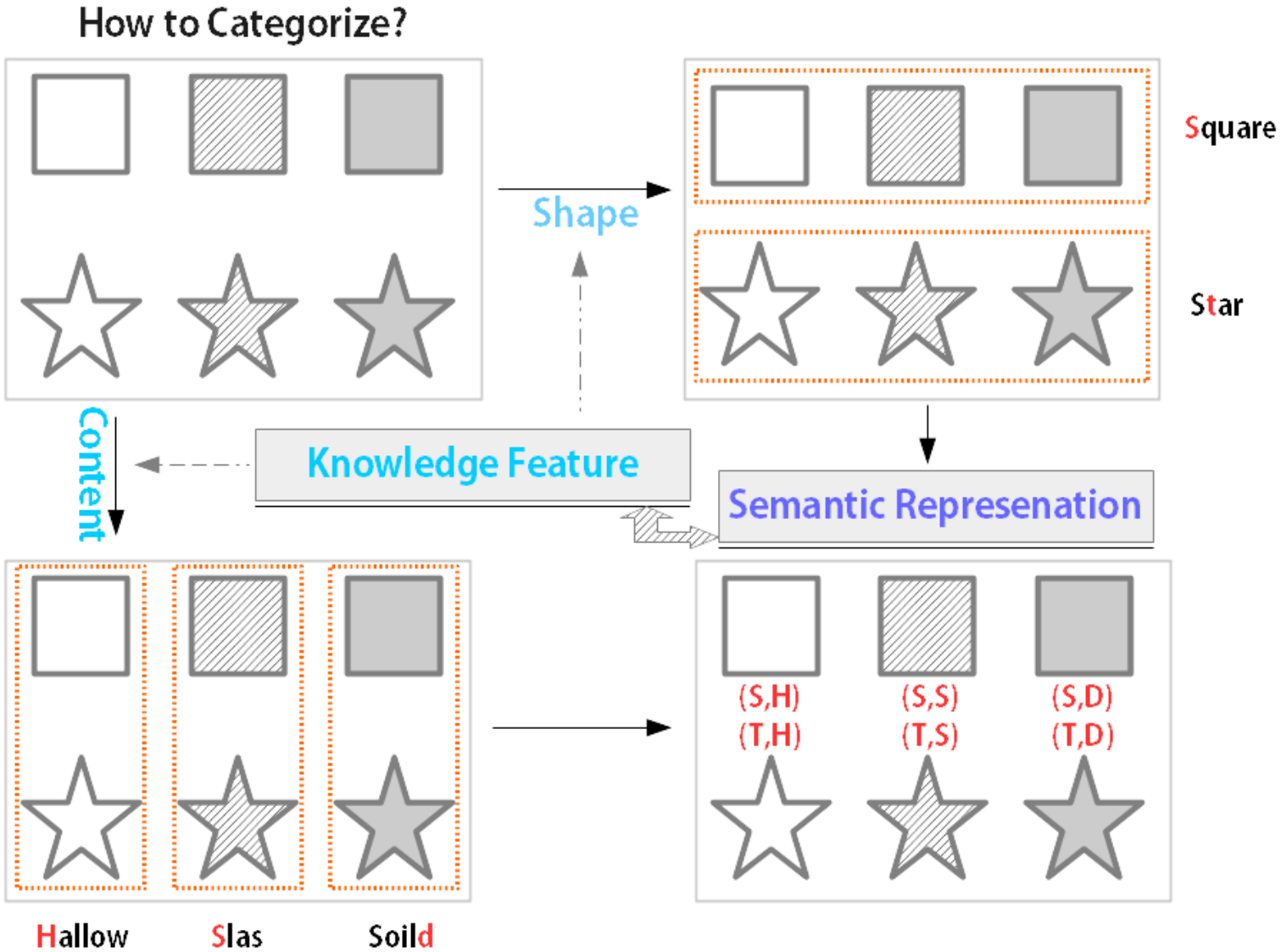
- *(Star Trek) =(Film:Related, American:Related, Sports:Unrelated, Person:Unrelated, Location :Unrelated, Drama:Related)*.
- **Entity Inference** is conducted as followed, approximately.

$$\mathbf{e}|\mathbf{c_{1..n}} = \arg\max_{\mathbf{e}\in\mathbf{E}} \prod_{\mathbf{i=1}}^{\mathbf{n}}[\mathbf{z_i} = \mathbf{c_i}|\mathbf{e}]$$

$$\mathbf{r}|\mathbf{c_{1..n}} = \arg\max_{\mathbf{r}\in\mathbf{R}} \prod_{\mathbf{i=1}}^{\mathbf{n}}[\mathbf{z_i} = \mathbf{c_i}|\mathbf{r}][\mathbf{y_i} = \mathbf{c}|\mathbf{r}]$$

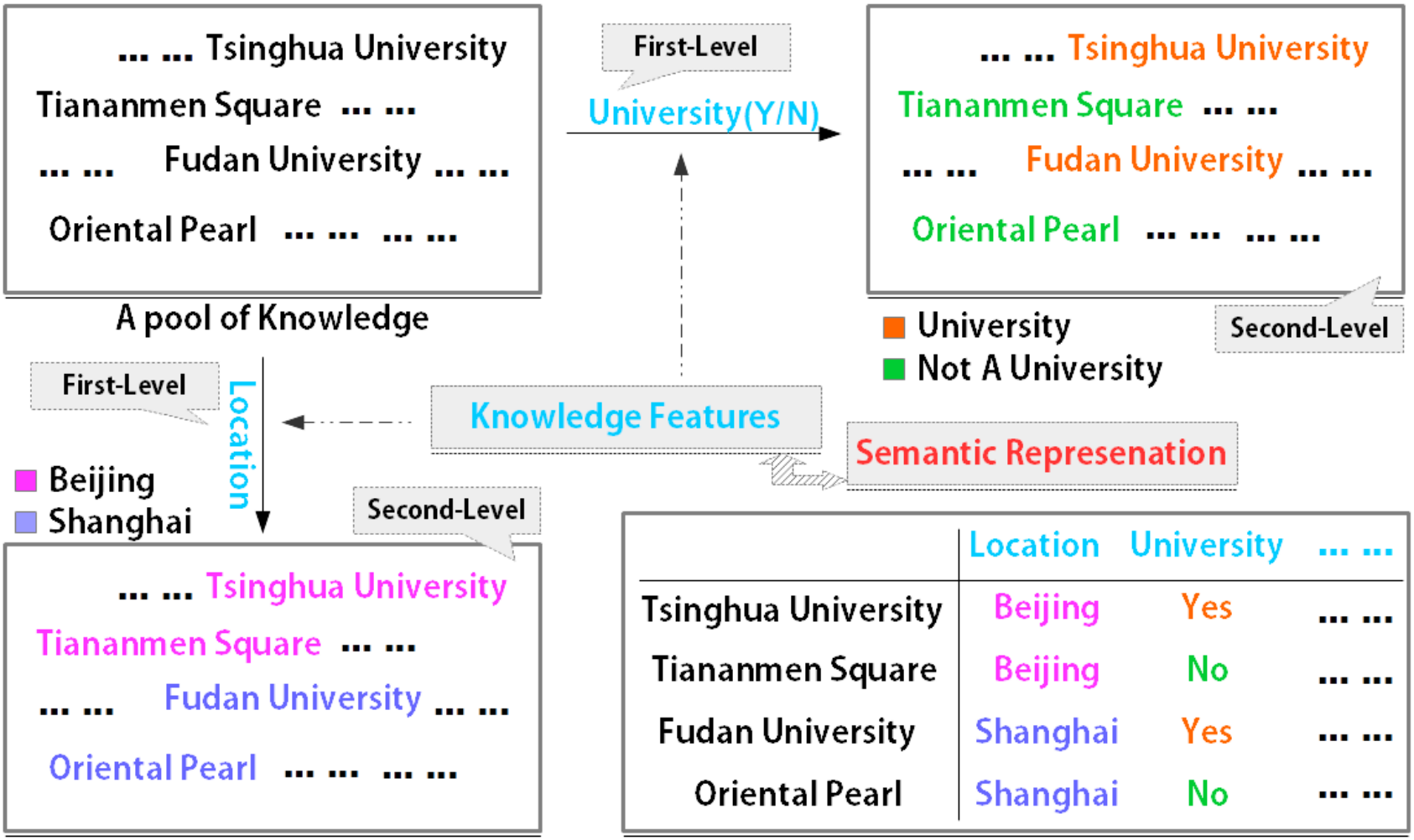
Knowledge Semantic Representation

- **Clustering Perspective (Basic Idea)**



Knowledge Semantic Representation

- **Clustering Perspective (Simple Illustration)**



Knowledge Semantic Representation

- **Identification Perspective**
  1. **Second-Level**. The false triple is supposed to be assigned to any category in a low probability.
  2. **First-Level**. Even if some features of this negative one holds high certainty, the corresponding relation also weights the feature with $[z_i, y_i, f|r]$ to filter out these noisy information.
- **Summarizing**: Our model could discriminate the plausibility of triples in *a two-level filtering form*, leading to a better performance.

Knowledge Semantic Representation

- **Experiments**: Knowledge Graph Completion

FB15K	**Mean Rank**		**HITS@10(%)**	
Methods	Raw	Filter	Raw	Filter
RESCAL	828	683	28.4	44.1
LFM	283	164	26.0	33.1
TransE	210	119	48.5	66.1
TransH	212	**87**	45.7	64.4
KSR(S1)	**178**	**87**	**55.6**	**75.7**
HOLE	-	-	-	73.9
KSR(S2)	**170**	**86**	**56.9**	**80.4**
TransR	198	77	48.2	68.7
CTransR	199	75	48.4	70.2
KG2E	183	69	47.5	71.5
ManifoldE	-	-	55.2	86.2
KSR(S3)	**159**	**66**	**57.2**	**87.2**
KSR(K4)	1	1	1	1

Knowledge Semantic Representation

- **Experiments**: Entity Classification

Metrics	Type@25	Type@50	Type@75
Random	39.5	30.5	26.0
TransE	82.7	77.3	74.2
TransH	82.2	71.5	71.4
TransR	82.4	76.8	73.6
ManifoldE	86.4	82.2	79.6
KSR(S1)	90.7	85.6	83.3
KSR(S2)	**91.4**	**87.6**	**85.1**
KSR(S3)	90.2	86.1	83.1

Knowledge Semantic Representation

- **Experiments**: Semantic Analysis: **Basic Setting**.
- **Task Description**.
- **Consideration**.

Knowledge Semantic Representation

- **Experiments**: Semantic Analysis: Case Study of **Features**

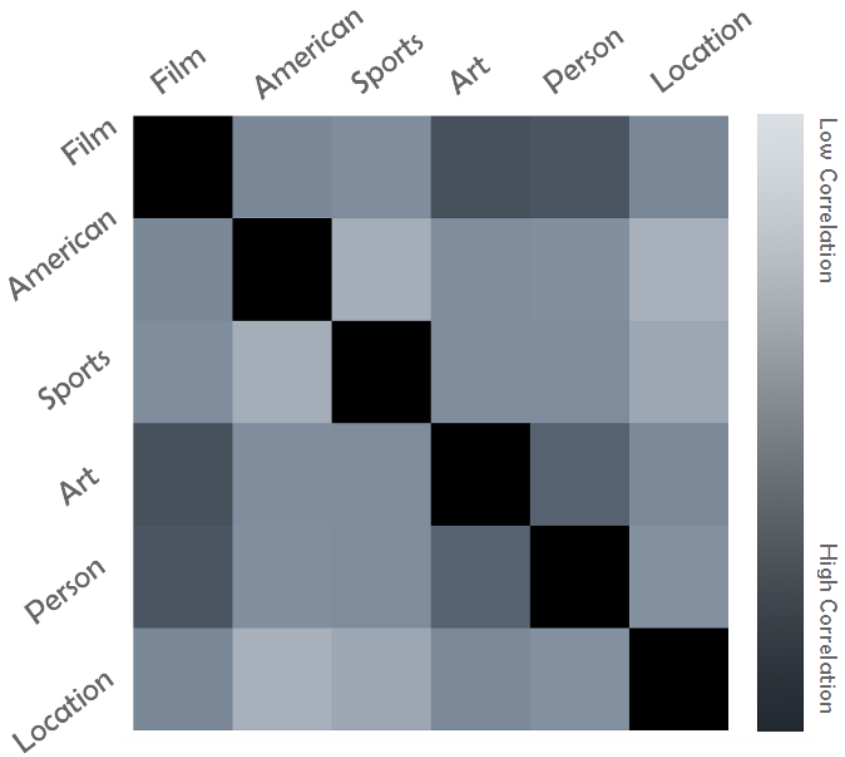
No.	Semantics	Categories(Significant Words)
1	**Film-Related**	**Yes** (Film, Director, Season), **Yes** (Producer, Actor), **No**
2	**American-Related**	**No**, **No**, **Yes** (United, States, Country, Area)
3	**Sports-Related**	**No**, **No**, **Yes** (Football, Basketball, World Cup)
4	**Art-Related**	**Yes**(Drama, Voice, Acting), **Yes** (Film, Story, Play), **No**
5	**Persons-Related**	**Multiple** (Team, League, Roles), **Single**(Actress, Director, Singer), **No**
6	**Location-Related**	**Yes** (British, London, England), **No**, **No**

Knowledge Semantic Representation

- **Experiments**: Semantic Analysis: Case Study of **Entity**
  1. *(Star Trek) =(Film:Related, American:Related, Sports:Unrelated, Person:Unrelated, Location :Unrelated, Drama:Related).*
  2. *(Football Club Illichivets Mariupol) = (Film:Unrelated, American:Unrelated, Sports:Related, Art:Unrelated, Persons:Multiple, Location:Related).*
  3. *(Johnathan Glickman)=(Film:Related, American:Unrelated, Sports:Unrelated, Art:Unrelated, Person:Single, Location:Unrelated).*
- **Experiments**: Semantic Analysis Case Study of **Relation**
  1. *(Country Capital) = (Film:Unrelated, American:Unrelated, Sports:Unrelated, Art:Unrelated, Person:Unrelated, Location:Related).*

Knowledge Semantic Representation

- **Experiments**: Semantic Analysis: **Statistical Justification**
  1. Semantic Correction:
     - 68% Totally Correct.
     - 19% Incorrect at Only One Feature.
     - 13% Incorrect at More Than One Feature.
  2. Correlation Heatmap:



Knowledge Semantic Representation

- **Experiments**: Description to Entity
- **Task Description.**
- **Case Study for Film**: An American 2011 biographical sports drama film directed by Bennett Miller from a screenplay by Steven Zaillian and Aaron Sorkin.
  1. **MoneyBall**
  2. Social Network
  3. Schindler's List
- **Case Study for Scientific Definition**: The social science of human social behavior and its origins, development, organizations, and institutions
  1. **Sociology**
  2. Economics
  3. Philosophy

Knowledge Semantic Representation

- **Case Study for Sports**:A professional baseball team located in Chicago, Illinois, USA.
  1. **Chicago Cub**
  2. Chicago White Sox
  3. Minnesota Twins
- **Case Study for Person**: A man who is a contemporary writer, playwright, screenwriter, actor and movie director in Kannada language. His rise as a playwright in 1960s, marked the coming of age of Modern Indian playwriting in Kannada, just as Badal Sarkar did in Bengali, Vijay Tendulkar in Marathi, and Mohan Rakesh in Hindi.
  1. **Girish Raghunath Karnad**
  2. Tanikella Bharani
  3. Sunny Deol
- **Case Study for Others**: An infectious illness of the liver caused by the hepatitis B virus that affects hominoidea, including humans.
  1. **Hepatitis B**

Knowledge Semantic Representation

- **Case Study for Partial Focus**:A ball sport played by tall men or tall boys, famous for NBA
  1. Basketball (Concept)
  2. Dwyane Tyrone Wade, jr. (NBA player)
  3. Los Angeles Lakers. (Team)
  4. Kobe Bean Bryant, nicked as Black Mamba. (NBA player)
  5. Lebron Raymone James, nicked as King James. (NBA player)

Knowledge Semantic Representation

- **Question Answering for Factoids**
  1. Who proposes the Relativity Theory? *Albert Einstein*
  2. Who proposes the Hawking Radio? *Stephen William Hawking*

  ---

  3. Which company is best at chips? *Intel Corporation*
  4. Which company is best at MultiMedia? *Adobe Systems Incorproated*.
  5. Which company is best at selling operating system? *Microsoft Corporation*
  6. Which company is best at selling phones? *Samsung Electronics, co*.
  7. Which company is best at selling sport shoes? *Nike, inc*.

  ---

  8. Which are the most famous cities(city) of Germany? *Hambury,Bonn*

  ---

  9. What is the Chinese province with capital Taiyuan? *Shan-Xi*
  10. What is the Chinese province with capital Changsha? *Hu-Nan*

Knowledge Semantic Representation

- **Question Answering for List-Type**
- Which provinces are the neighbor of Beijing?
  1. Beijing
  2. Tianjin
  3. He-Bei
  4. Peking University
- Which provinces are the neighbor of Macao?
  1. Macao
  2. Guang-Dong
  3. Jiang-Su
  4. Su-Zhou
- Which countries are the neighbor of China?
  1. China
  2. Burma
  3. Vietnam
  4. India

Knowledge Semantic Representation

- **Question Answering with Latent Semantic Results**
- The people in which Chinese province is richest?
  1. China
  2. Macao
  3. ~~Su-Zhou~~
  4. Hong Kong
  5. Jiang-Su
  6. Zhe-Jiang
- Which universities are famous in Kyoto?
  1. Kyoto University
  2. Kyoto
  3. Fukuoka
  4. Kansai
  5. The University of Tokyo

Knowledge Semantic Representation

- **Question Answering for Fun**
- Which programming language is the best in the world?
  1. Python
  2. **C++**
  3. C
  4. Java
  5. Perl

Knowledge Representation Framework

- An Open-Source Knowledge Representation Framework.
- Contributed by XiaoHan in Tsinghua University.
- The functions contained in this project are listed as
  - Testing Framework.
  - Dataset Loading Framework.
  - Logging Framework.
  - Parallelism Framework based on OpenMP.
  - Report Framework.
  - ...
- URL: `https://github.com/BookmanHan/Embedding`.

Knowledge Semantic Representation

XiaoHan 肖寒

Home Page: `http://www.ibookman.net`

CV for Faculty: `http://www.ibookman.net/res/CV.pdf`

Thanks for your attention.