# Understanding Short Texts
## ACL 2016 Tutorial

Zhongyuan Wang (Microsoft Research)

Haixun Wang (Facebook Inc.)

**Tutorial Website**:
http://www.wangzhongyuan.com/tutorial/ACL2016/Understanding-Short-Texts/

# Outline

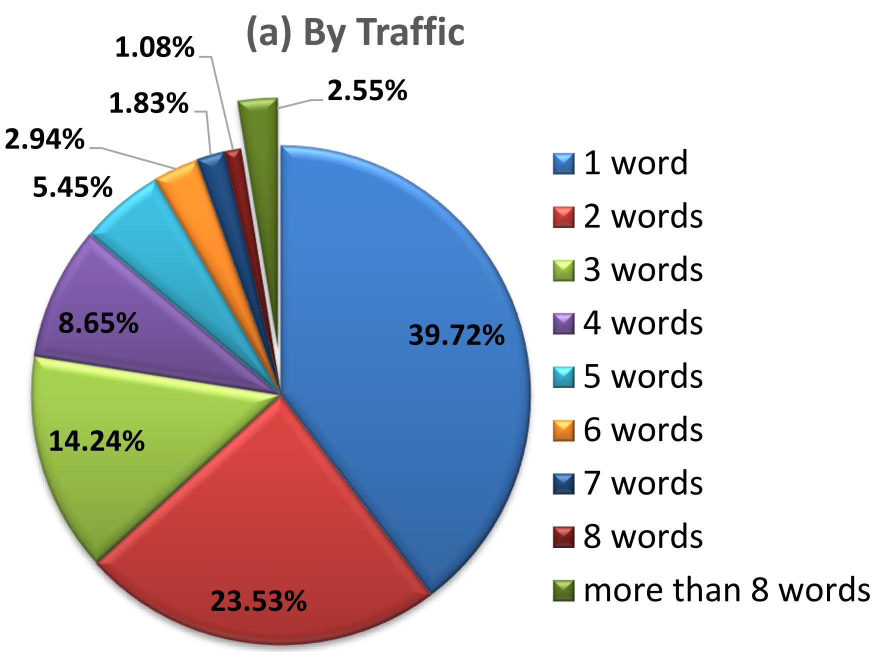
- Part 1: Challenges
- Part 2: Explicit representation
- Part 3. Implicit representation
- Part 4: Conclusion

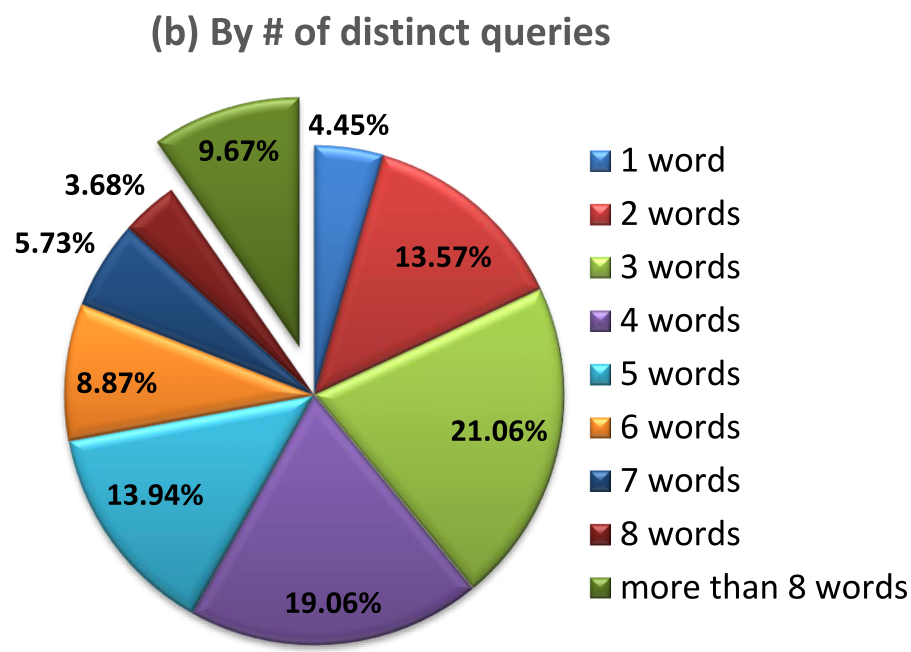
# Short Text

- Search Query
- Ad keyword
- Anchor text
- Image Tag
- Document Title
- Caption
- Question
- Tweet/Weibo

# Challenges

- *First, short texts contain limited context*




*Based on Bing query log between 06/01/2016 and 06/30/2016*

# Challenges

- *Second, "telegraphic": no word order, no function words, no capitalization, …*

*Query "*Distance between Sun and Earth*" can also be expressed as:*

- "how far" earth sun
- "how far" sun
- "how far" sun earth
- average distance earth sun
- average distance from earth to sun
- average distance from the earth to the sun
- distance between earth & sun
- distance between earth and sun
- distance between earth and the sun
- distance from earth to the sun
- distance from sun to earth
- distance from sun to the earth
- distance from the earth to the sun
- distance from the sun to earth
- distance from the sun to the earth
- distance of earth from sun
- distance between earth sun
- how far away is the sun from earth
- how far away is the sun from the earth
- how far earth from sun
- how far earth is from the sun
- how far from earth is the sun
- how far from earth to sun
- how far from the earth to the sun
- distance between sun and earth

# Challenges

- *Second, “telegraphic”: no word order, no function words, no capitalization, …*

| **Short Text 1** | **Short Text 2** | **Term Match** | **Semantic Match** |
|---|---|---|---|
| china kong *(actor)* | china hong kong | partial | no |
| hot dog | dog hot | yes | no |
| the big apple tour | new york tour | almost no | yes |
| Berlin | Germany capital | no | Yes |
| DNN tool | deep neural network tool | almost no | Yes |
| wedding band | band for wedding | partial | no |
| why are windows so expensive | why are macs so expensive | partial | no |

# Challenges

- *Sparse, noisy, ambiguous*

**watch for kids**

i) 


ii) 

iii) 

# Short Text Understanding

- Many applications
  - Search engines
  - Automatic question answering
  - Online advertising
  - Recommendation systems
  - Conversational bot
  - …

- Traditional NLP approaches not sufficient

# The big question

- Humans are much powerful than machines in understanding short texts.
- Our minds build rich models of the world and make strong generalizations from input data that is *sparse, noisy, and ambiguous* – in many ways far too limited to support the inferences we make.
- How do we do it?

If the mind goes beyond the data given, *another source of information* must make up the difference.

Science AAAS

*Science* **331**, 1279 (2011);

**How to Grow a Mind: Statistics, Structure, and Abstraction**

Joshua B. Tenenbaum,[1*] Charles Kemp,[2] Thomas L. Griffiths,[3] Noah D. Goodman[4]

Explicit
(Logic)
Representation

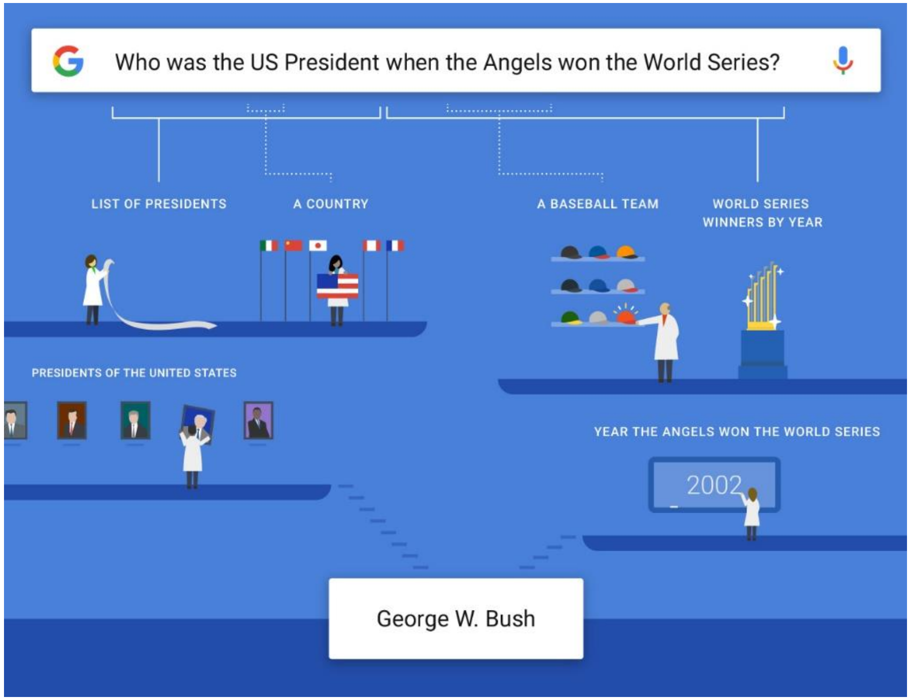
How?

Implicit
(Embedding)
Representation

Symbolic knowledge
(**Explicit**)

Distributional semantics
(**Implicit**)

# Explicit Knowledge Representation

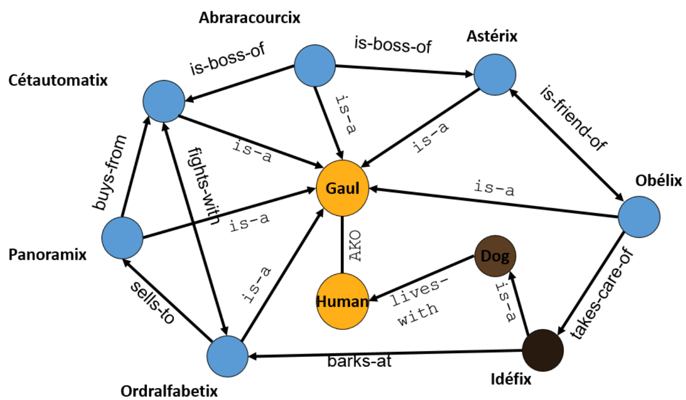



- First, understand **superlatives**—"tallest," "largest," etc.—and ordered items. So you can ask:
  *"Who are the tallest Mavericks players? "*
  *"What are the largest cities in Texas? "*
  *"What are the largest cities in Iowa by area? "*
- Second, have you ever wondered about a **particular point in time**? Google now do a much better job of understanding questions with dates in them. So you can ask:
  *"What was the population of Singapore in 1965?"*
  *"What songs did Taylor Swift record in 2014?"*
  *"What was the Royals roster in 2013?"*
- Finally, Google starts to understand some **complex combinations**. So Google can now respond to questions like:
  *"What are some of Seth Gabel's father-in-law's movies?"*
  *"What was the U.S. population when Bernie Sanders was born?"*
  *"Who was the U.S. President when the Angels won the World Series?"*

http://insidesearch.blogspot.com/2015/11/the-google-app-now-understands-you.html

# Explicit Knowledge Representation

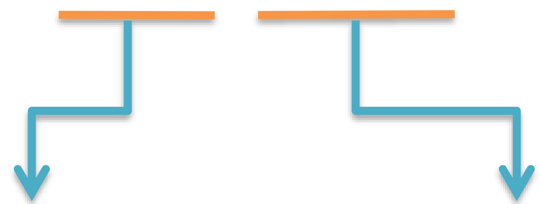
- Logic Representation (First-order-logic)
  - Freebase, Google knowledge Graph…




**True or False**

- Vector Representation
  - ESA: Mapping text to Wikipedia article titles
  - Conceptualization: Mapping text to concept space

P(concept | short text)



a domain millions of concepts used in day to day communication

search query, anchor text twitter, ads keywords, …

**Probabilistic Model**

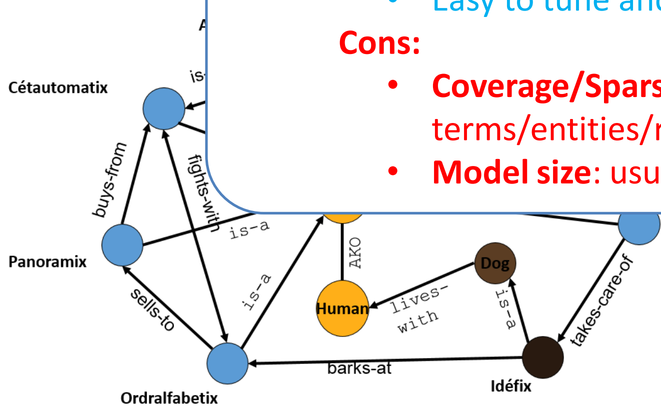
# Explicit Knowledge Representation

- Logic Representation (First-order-logic)
  - Freeb… know…

- Vector Representation
  - ESA: Mapping text to Wikipedia article titles
  - …apping text
  - …)

**Pros:**
- The results are easy to understand for human beings
- Easy to tune and customize

**Cons:**
- **Coverage/Sparse model**: can't handle unseen terms/entities/relations
- **Model size**: usually very large




a domain millions of concepts used in day to day communication

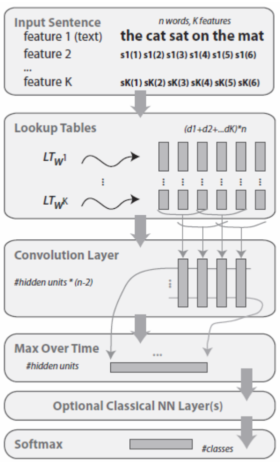
search query, anchor text twitter, ads keywords, …

**True or False**

**Probabilistic Model**

# Implicit Knowledge Representation: Embedding

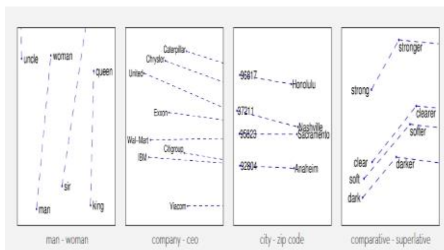
**CW08**

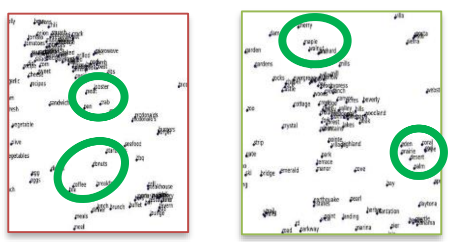



**Input units:** word
**Vocabulary:** 130k

Collobert, Ronan, et al. "Natural language processing (almost) from scratch." *The Journal of Machine Learning Research* 12 (2011): 2493-2537.

**GloVe**



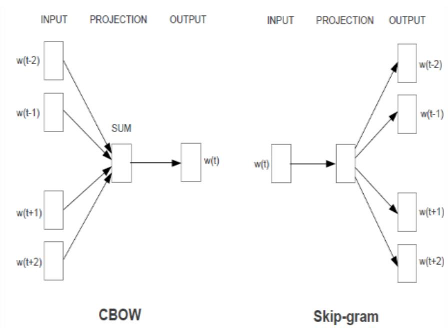



**Input units:** word
**Training size:** > 42B tokens
**Vocabulary:** > 400K

J Pennington, R Socher, CD Manning "Glove: Global Vectors for Word Representation." EMNLP 2014.

word2vec
Tool for computing continuous distributed representations of words.

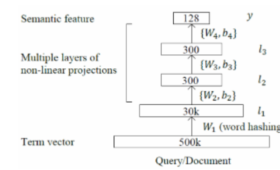
https://code.google.com/p/word2vec/




**Input units:** word
**Training size:** > 100B sequence (Freebase)
**Vocabulary:** > 2M

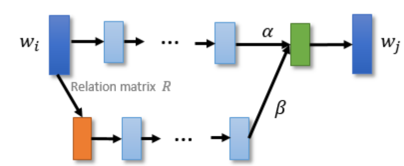
Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013.

**Deep Structured Semantic Model (DSSM)**




**Input units:** Tri-letter
**Training size:** ~20B clicks (Bing + IE log)
**Vocabulary:** 30K **Parameter:** ~10M

**KNET**




$$\max \Pr(w_j|w_i) = \frac{\exp\left(c_1(w_i) {v'_{w_j}}^T v_{w_i} + c_2(w_i) {v'_{w_j}}^T v_{R_i}\right)}{\sum_{w \in V} \exp\left(c_1(w_i) {v'_w}^T v_{w_i} + c_2(w_i) {v'_w}^T v_{R_i}\right)}$$

Huang, Po-Sen, et al. "Learning deep structured semantic models for web search using clickthrough data." in CIKM. ACM, 2013.

**Count + Predict**

**Predict**

# Implicit Knowledge Representation: Embedding

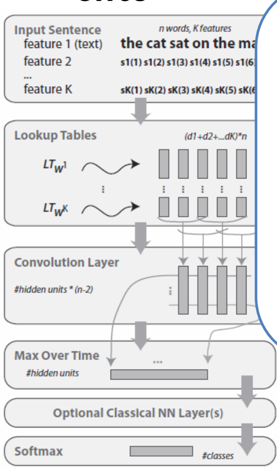
**Pros:**

- Dense semantic encoding
- A good representation framework
- Facilitates computation (similarity measure)

**Cons:**

- Perform poorly for rare words and new words
- Missing relations (e.g, isA, isPropertyOf)
- Hard to tune since it's not nature for human beings

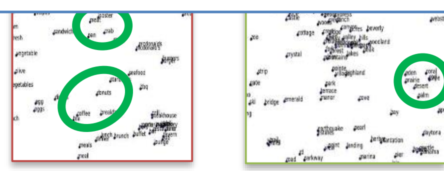
**CW08**




**Input units:** word
**Vocabulary:** 130k

Collobert, Ronan, et al. "Natural language processing (almost) from scratch." *The Journal of Machine Learning Research* 12 (2011): 2493-2537.

GloVe



**Input units:** word
**Training size:** > 42B tokens
**Vocabulary:** > 400K

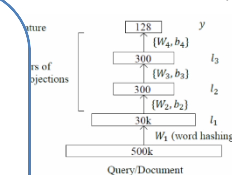
J Pennington, R Socher, CD Manning "Glove: Global Vectors for Word Representation." EMNLP 2014.

word2vec


**Input units:** word
**Training size:** > 100B sequence (Freebase)
**Vocabulary:** > 2M

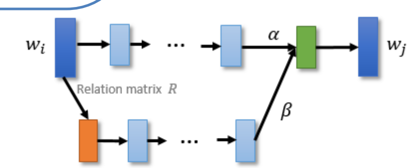
Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013.

**Deep Structured Semantic Model (DSSM)**




: Tri-letter
e: ~20B clicks (Bing + IE log)
: 30K **Parameter:** ~10M

**KNET**



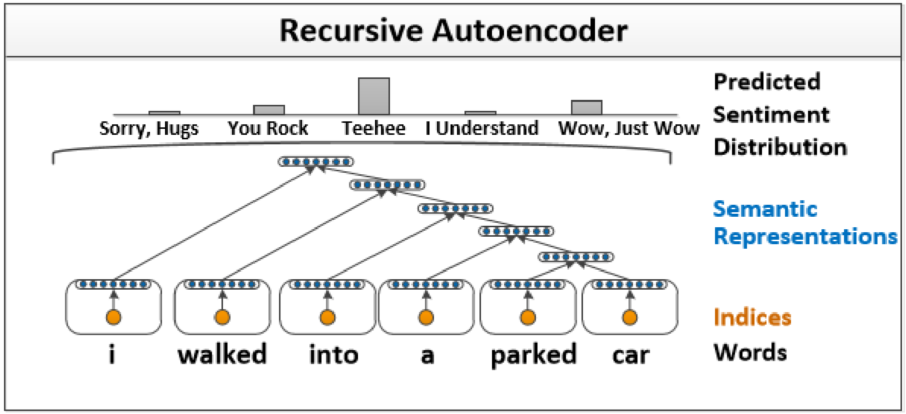

$$\max \Pr(w_j|w_i) = \frac{\exp\left(c_1(w_i) v'^{T}_{w_j} v_{w_i} + c_2(w_i) v'^{T}_{w_j} v_{R_i}\right)}{\sum_{w \in V} \exp\left(c_1(w_i) v'^{T}_{w} v_{w_i} + c_2(w_i) v'^{T}_{w} v_{R_i}\right)}$$

Huang, Po-Sen, et al. "Learning deep structured semantic models for web search using clickthrough data." in CIKM. ACM, 2013.

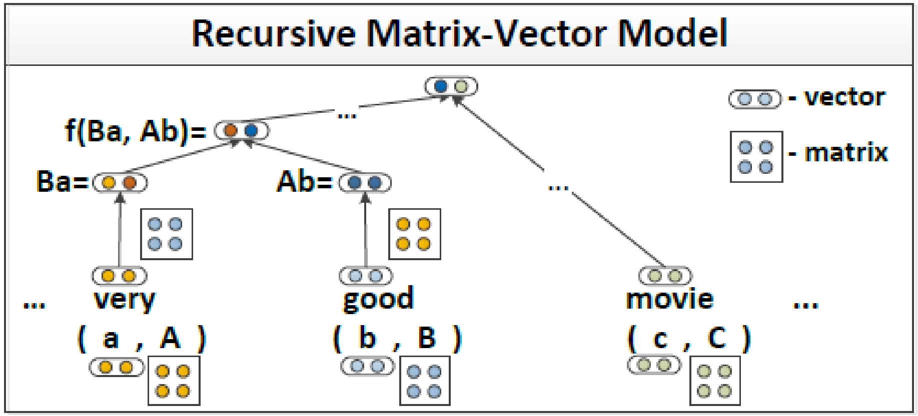
**Count + Predict**

**Predict**

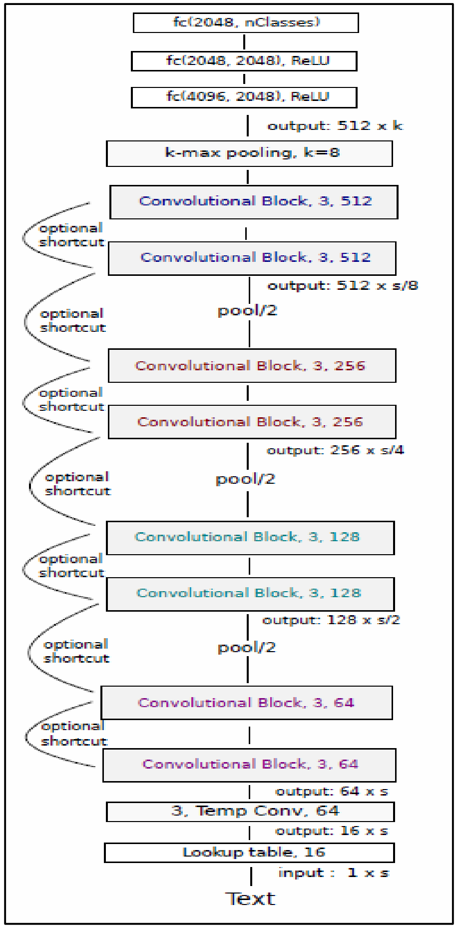
# Implicit Knowledge Representation: DNN




Stanford Deep Autoencoder for Paraphrase Detection [Soucher et al. 2011]



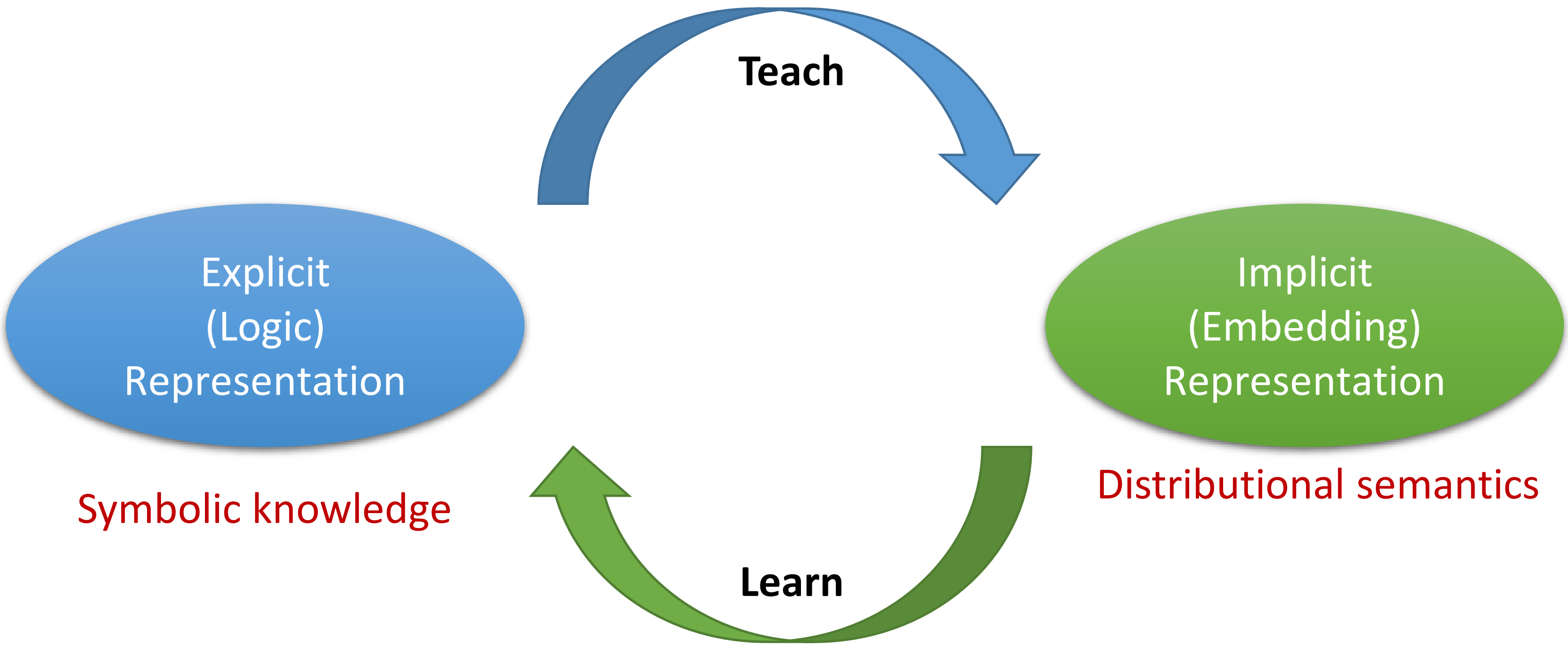

Stanford MV-RNN for Sentiment Analysis [Soucher et al. 2012]




Facebook DeepText classifier [Zhang et al. 2015]

# New Trend: Fusion of Explicit and Implicit knowledge

- Relationship
- Rules of inference




- Learn more similar rules, enrich logic representation