# Probabilistic Reasoning via Deep Learning: Neural Association Models

**Quan Liu**[†]

[†]University of Science and Technology of China

Joint work with **Hui Jiang**[‡], **Zhen-Hua Ling**[†], **Si Wei**[§] and **Yu Hu**[§]
[‡]York University, Canada
[§]iFLYTEK Research, Hefei, China

July 10, 2016

# Outline

1. Neural Association Model (NAM)
   - Motivation
   - Model
   - Experiments

2. NAM for Winograd Schemas
   - Winograd Schemas
   - Data Collection
   - NAM for Winograd Schemas

# Neural Association Model

Neural **Association** Model

$\Uparrow$

Main work

### Motivation: Neural Model to Associate between Events

- Events emerge everywhere ($\rightarrow$ massive) in our diary life.
- Events are discrete ($\rightarrow$ sparse).
- Commonsense reasoning relies on the Association between Events.
- Association relationships
  - Causality, Temporal, Taxonomy, Entailment, etc.

# Examples

- What are the possible events Associated with event “Play basketball”?


.......

Association $\neq$ Classification!

# Motivation: Main Method

**Neural Association Model: a neural model for probabilistic reasoning**

- Associating two events via deep learning techniques:
- Predicting the conditional association probability $\Pr(E_2|E_1)$ of two different events, $E_1$ and $E_2$.

| Application | $E_1$ | $E_2$ |
|---|---|---|
| Causal-Effect reasoning | *cause* | *effect* |
| Recognize lexical entailment | $W_1$ | $W_2$ |
| Recognize textual entailment | $D_1$ | $D_2$ |
| Language modeling | $h$ | $w$ |
| Knowledge link prediction | $(e_i, r_k)$ | $e_j$ |

E.g. Causal-Effect reasoning

- $E_1$ = cause event
- $E_2$ = effect event

How likely $E_2$ is caused by $E_1$?

# Advantages vs. Disadvantages

## Advantages of NNs for reasoning

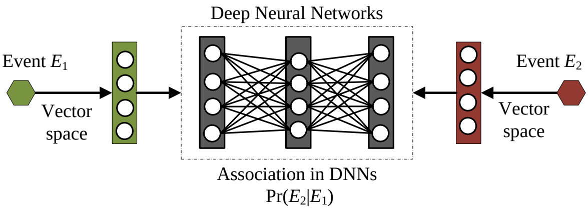
- Neural networks make universal approximation (Hornik et al., 1990).
  - Linear models can hardly do this.
  - Nickel, Murphy et al. (2015)
- Associating in continuous spaces improve scalability.
  - Graphical models suffer from the scalability issue.
  - Jensen (1996); Richardson and Domingos (2006)

## Disadvantages

- Deep learning need big data, i.e., KBs.
  - Automated Knowledge Acquisition
  - Transfer Learning

# 2. Neural Association Model

A neural model for modeling the association probability of two events.



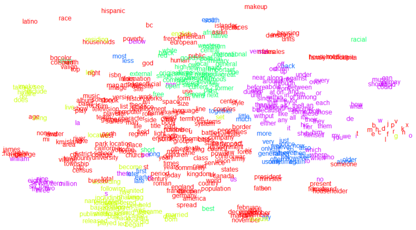

## Key modules

- **Representation**: Represent discrete events into continuous vectors
- **Association**: Predict the association probability via deep learning

# Association via DNNs

## Distributed representations

All discrete events are represented in continuous vector spaces.



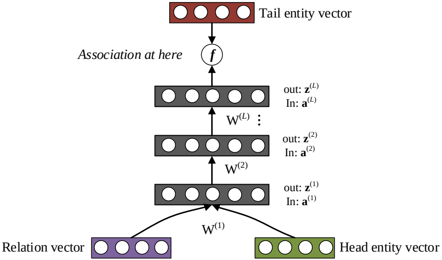

## Two model structures for Association

1. Deep Neural Networks (DNN)
2. Relation-modulated Neural Networks (RMNN)

# 2.1 Deep Neural Networks

## Deep Neural Networks (DNN)

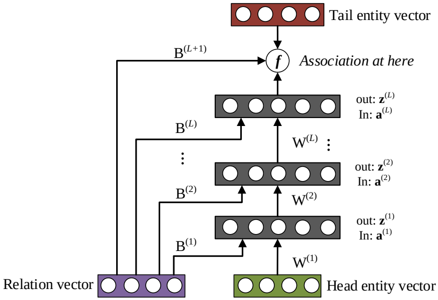
- Associating two events through deep neural networks
- For a multi-relation data $x_n = (e_i, r_k, e_j)$:
  - Entity vector: $e_i \rightarrow \mathbf{v}_i^{(1)}$, $e_j \rightarrow \mathbf{v}_j^{(2)}$ (Different embedding matrices)
  - Relation code: $r_k \rightarrow \mathbf{c}_k$




- $\mathbf{z}^{(0)} = [\mathbf{v}_i^{(1)}, \mathbf{c}_k]$
- $\mathbf{a}^{(\ell)} = \mathbf{W}^{(\ell)}\mathbf{z}^{(\ell-1)} + \mathbf{b}^{\ell}, \quad \ell = 1...L,$
- ReLU hidden layer activation:
  $\mathbf{z}^{(\ell)} = \max\left(0, \mathbf{a}^{(\ell)}\right), \ell = 1...L,$
- The associative probability:
  $f(x_n; \mathbf{\Theta}) = \sigma\left(\mathbf{z}^{(L)} \cdot \mathbf{v}_j^{(2)}\right),$
  $\sigma(x) = 1/(1+e^{-x}).$

## Relation-modulated Neural Networks (RMNN)

- Improved over DNN
- Define and connect relation codes to all the layers of DNN




- $\mathbf{z}^{(0)} = [\mathbf{v}_i^{(1)}, \mathbf{c}_k]$
- $\mathbf{a}^{(\ell)} = \mathbf{W}^{(\ell)}\mathbf{z}^{(\ell-1)} + \mathbf{B}^{(\ell)}\mathbf{c}^{(k)}, \quad \ell = 1 \ldots L,$
- ReLU hidden layer activation: $\mathbf{z}^{(\ell)} = \max\left(0, \mathbf{a}^{(\ell)}\right), \ell = 1 \ldots L,$
- The associative probability: $f(x_n; \boldsymbol{\Theta}) = \sigma\left(\mathbf{z}^{(L)} \cdot \mathbf{v}_j^{(2)} + \mathbf{B}^{(L+1)} \cdot \mathbf{c}^{(k)}\right).$

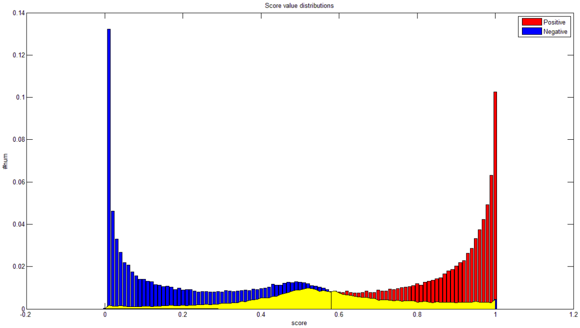
# NAM: Final Training Objectives

Training sample: event pair $x = (E_1, E_2)$; score: $f(x; \mathbf{\Theta}) = \Pr(E_2|E_1)$

## Training objective

For each positive sample $x_n^+$ and negative sample $x_n^-$, To **minimize**:

$$\mathcal{Q}(\mathbf{\Theta}) = -\sum_{x_n^+ \in \mathcal{D}^+} \ln f(x_n^+; \mathbf{\Theta}) - \sum_{x_n^- \in \mathcal{D}^-} \ln(1 - f(x_n^-; \mathbf{\Theta})) \quad (1)$$



# 3. Experiments

## Experiments

- Recognizing textual entailment
- Commonsense reasoning

# 3.1 Recognizing Textual Entailment (RTE)

## Recognizing Textual Entailment

- Recognizing the entailment relationship between two sentences
  - Premise: "*The man was assassinated.*"
  - Hypothesis: "*The man is dead.*"
- Datasets
  - The Stanford Natural Language Inference (SNLI) Corpus
- Experiments: 2-class recognition

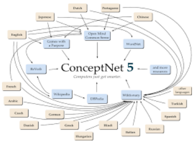
| Model | Accuracy (%) |
|---|---|
| Edit Distance Based | 71.9 |
| Classifier Based | 72.2 |
| With Lexical Resources | 75.0 |
| Neural Association Model | **84.7** |

- NAM model performs better than many traditional methods.

# 3.2 Commonsense Reasoning

## Commonsense reasoning

- Task investigated in this work
  - Answering simple commonsense questions
  - Judge the truth of commonsense triples
    - "*Is a camel capable of journey across desert?*"
    - Triple: (camel, capable of, journey across desert).
- Datasets
  - From ConceptNet 5, a commonsense KB (Speer and Havasi 2012). `http://conceptnet5.media.mit.edu/`
  - We extract 14 popular commonsense relations (CN14).



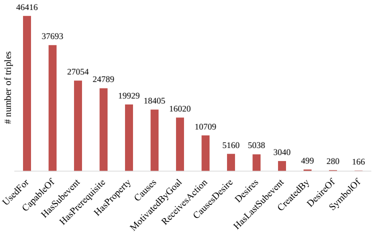

| Dataset | #Rel | #Entities | # Train | # Dev | # Test |
|---|---|---|---|---|---|
| CN14 | 14 | 159,135 | 200,198 | 5,000 | 10,000 |

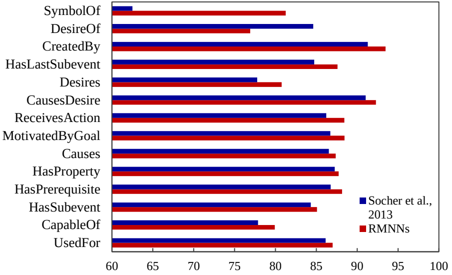
# Results

- Overall results on CN14

| Model | Accuracy (%) |
|---|---|
| DNN | **85.7** |
| RMNN | **86.1** |

- Results on different relations




- NAM shows some potentials for commonsense reasoning.

# Application:
# NAM for Winograd Schemas

# Winograd schemas

## Typical Winograd schemas example

Co-reference cannot be resolved without commonsense knowledge.

- Statement: Marry made sure to thank Susan for all the help she had received.
- Q: who had received the help?
- Answer: Marry

Commonsense knowledge: **receive help** $\rightarrow$ **thank**

$$\Downarrow$$

Association between Events:

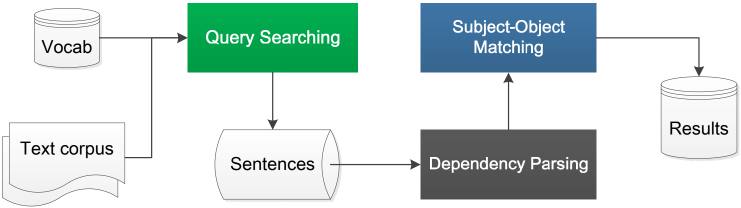
$$\Pr(\text{thank}|\text{receive help}) > \Pr(\text{thank}|\text{give help})$$

# NAM for Winograd Schemas

## Modules for Solving Winograd Schemas

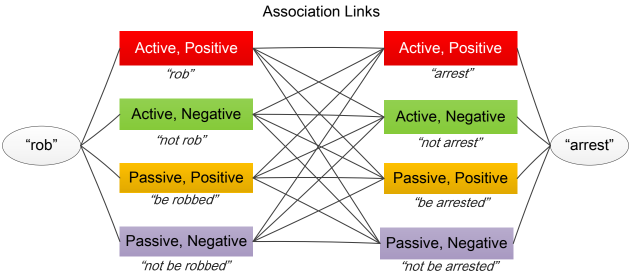
- Neural Association Model
- Data Collection: how to collect training data for NAM?

- System framework for data collection



## Query Search in Text Corpus

- Search query: keyword pairs formed from a common vocabulary.
  - Vocabulary: 7500 common verbs and adjectives.
  - E.g. (*arrest* ... *because* ... *rob*); (*decide* ... *because* ... *explain*)
- Each word/phrases have 4 variations → 16 patterns for each query.



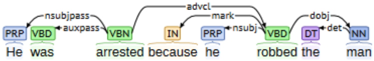

We want to gather the **number** of active association links.

# 1. Data Collection

## Association knowledge from dependency parsing

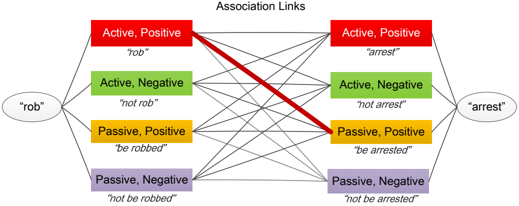
- Subject/Object Matching ⇒ Assigning Association links
- Collect the number of active links

*"He was arrested because he robbed the man."*




- (he, nsubjpass, arrest), (he, nsubj, rob)
- "rob" and "arrest" share a same subject "he"
- "nsubjpass" ⇒ passive

"rob" ⇒ "be arrested"



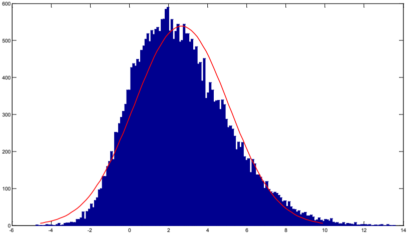
# Data collection results

## Copora for data collection

- BookCorpus (Zhu et al., 2015)
- CBTest corpus (Hill et al., 2016)
- Gigaword 5 (Parker, Robert, et al., 2011)

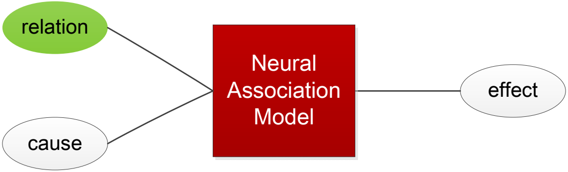
## Results: highly associated pairs

- We extracted about 100,000 highly associated pairs.
  - (know $\Rightarrow$ clear), (believe $\Rightarrow$ not disagree), (be released $\Rightarrow$ not hold).
- Typical PMI distributions

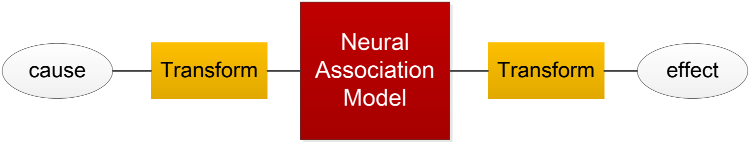


# 2. NAM for Winograd Schemas

- NAM RelationCode: Treat the 16 dimensions as distinct relations




- NAM TransMatrix: Do linear transformation for each word/phrases



# 2. NAM for Winograd Schmeas

## Datasets

- From http://www.cs.nyu.edu/faculty/davise/papers/WS.html
- We labelled 70 schemas related to cause effect reasoning.
  - Available at http://home.ustc.edu.cn/~quanliu/

## Results

- We now achieve 61.4% accuracy on the Winograd CE datasets.

| Model | Accuracy (%) |
|---|---|
| NAM TransMatrix | **59.6** |
| NAM RelationCode | **61.4** |

Table: Performance of NAM.

# Answering examples

- "tasty" $\rightarrow$ "be eaten"

============================================================

>> **Schema_51-A**: the fish ate the worm. it was tasty

- **Ques**: what was tasty?
- **Cand**: the worm/the fish
- **Answer**: the worm

**[*] DNN Scores**

| Candidate | Cause | Effect | Gold Label | Score |
|---|---|---|---|---|
| 1 | tasty | be eat | YES | 0.875606 |
| 2 | tasty | eat | NO | 0.304860 |

**[*] Final Decisions**

- Score for A: 0.875606
- Score for B: 0.304860
- System decision: **A**

============================================================

# Answering examples

- "hungry" $\rightarrow$ "eat"

==================================================

>> **Schema_51-B**: the fish ate the worm. it was hungry

- **Ques**: what was hungry?
- **Cand**: the worm/the fish
- **Answer**: the fish

**[*] DNN Scores**

| Candidate | Cause | Effect | Gold Label | Score |
|---|---|---|---|---|
| 1 | hungry | be eat | NO | 0.227095 |
| 2 | hungry | eat | YES | 0.929566 |

**[*] Final Decisions**

- Score for A: 0.227095
- Score for B: 0.929566
- System decision: **B**

==================================================

# Future works

## Data level

- Collect more useful data for commonsense reasoning
- Automatic construction from text/KBs
- Human labelling

## Model level

- Toward more complex probabilistic reasoning problems
- Neural association model for transfer learning

Thank You!

(Q&A)