On Integrating Domain Knowledge into DNN

Keh-Yih Su

Institute of Information Science
Academia Sinica, Taipei

NLPCC, Oct. 13, 2019

Formosa Grand Challenge 科技大擂台

科技大擂台與AI對話

3/23(六) 決賽評審會議

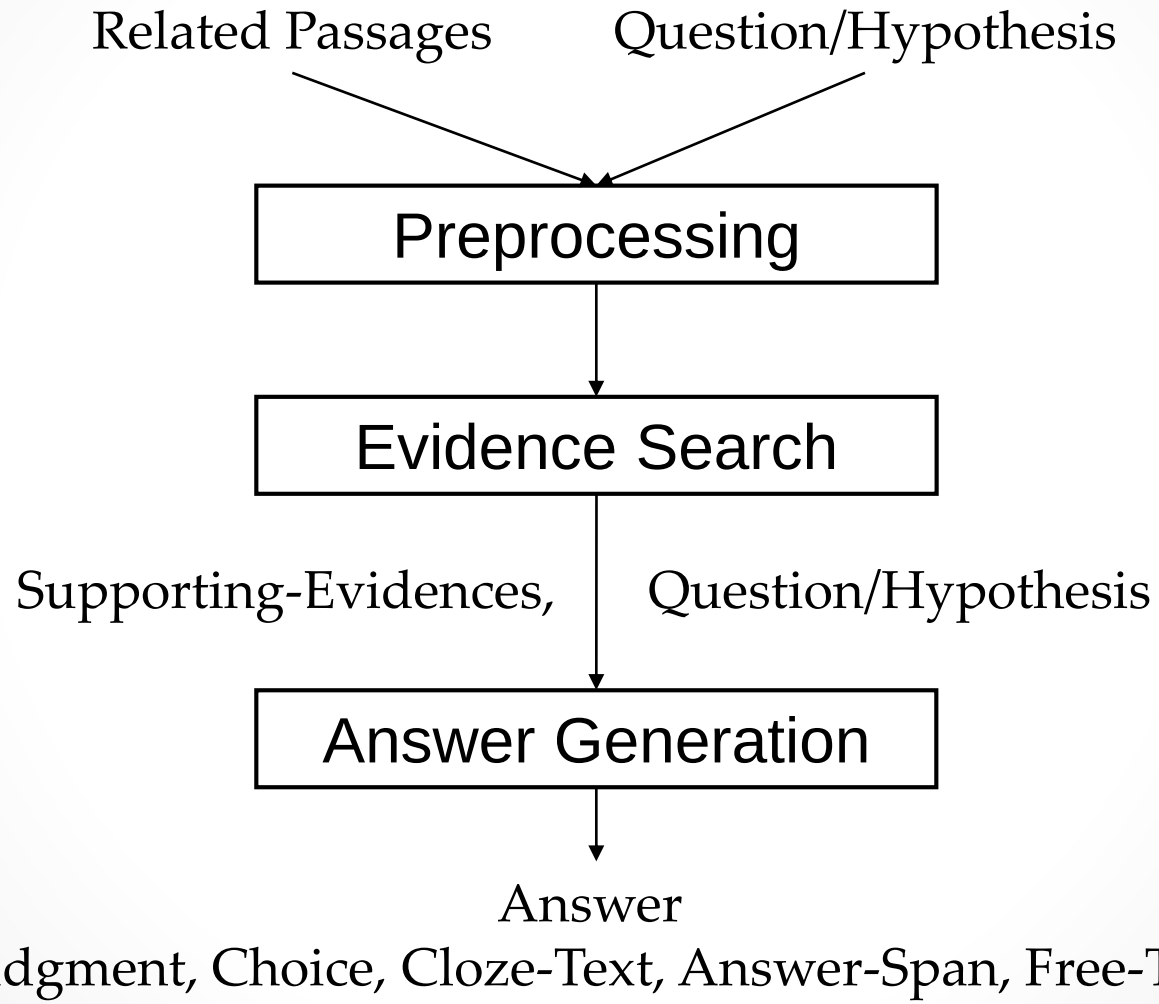
上午10:00~12:00

@西門 WESTAR

競賽獎金

決賽獎項基準	競賽獎金
最高分隊伍超過人類預試得分	第 1 名：2000 萬元 463.5萬人民幣 第 2 名：500 萬元
最高分隊伍未達人類預試得分，但達到人類預試得分之 80%	第 1 名：500 萬元（特別獎）
最高分隊伍未達人類預試得分之 80%	第 1 名：200 萬元（特別獎） 第 2 名：50 萬元（特別獎） 第 3 名：30 萬元（特別獎）

MR/QA: Natural Language Inference



團隊選擇題-答題成績總表

團隊名稱	團隊答對題數	正確率	排名
hungyilee	537	0.537	1
intellection	478	0.478	2
aieverywhere	408	0.408	6
The MUG	359	0.359	7
d204096001	422	0.422	5
weiyunma	475	0.475	3
elandtw	432	0.432	4
FireFalcons	未參賽		
Hakka	320	0.320	8

團隊簡答題-答題成績總表

團隊名稱	團隊答對題數	正確率
hungyilee	4	0.16
intellection	2	0.08
aieverywhere	4	0.16
The MUG	8	0.32
d204096001	2	0.08
weiyunma	7	0.28
elandtw	2	0.08
FireFalcons	未參賽	
Hakka	4	0.16

Outline

- DNN Limitations Observed
  - Incapable of conducting common inference
  - Need huge amount of data
  - Usually uninterpretable
  - Vulnerable to irrelevant data
- Lessons from Non-NLP Machine Learning
- Characteristics of NLP/NLU
- Various Ways to Integrate Domain Knowledge

Incapable of High Level Learning (Lin, 2017)

- **One correct prediction in the training set (Google Translation)**
  - Benchmark:Y, Prediction: Y

  (S)馬祖 位於 閩江 口 外 ， 島嶼 狹小 ， 又 無 高山 屏障 ， 全 年 多 風 少 雨 。

  Mazu is located outside the mouth of the Minjiang River. The island is small and has no alpine barrier. It is windy and rainless all year round.

  (H)馬祖 位於 閩江 口 外 ， 全 年 多 風 少 雨 。

  Mazu is located outside the mouth of the Minjiang River, with more wind and less rain all year round.

All words have valid embeddings.

- **Predictions for some created examples**
  - Benchmark:Y, Prediction: N

  (S)馬祖 位於 閩江 口 外 ， 全 年 多 風 少 雨 。

  (H)馬祖 位於 閩江 口 外 ， 全 年 多 風 少 雨 。
  - Benchmark:N, Prediction:Y

    (S)馬祖 位於 閩江 口 外 ， 島嶼 狹小 ， 又 無 高山 屏障 ， 全 年 多 風 少 雨 。

    (H)馬祖 位於 閩江 口 外 ， 全 年 少 風 少 雨 。
  - Benchmark:N, Prediction:Y

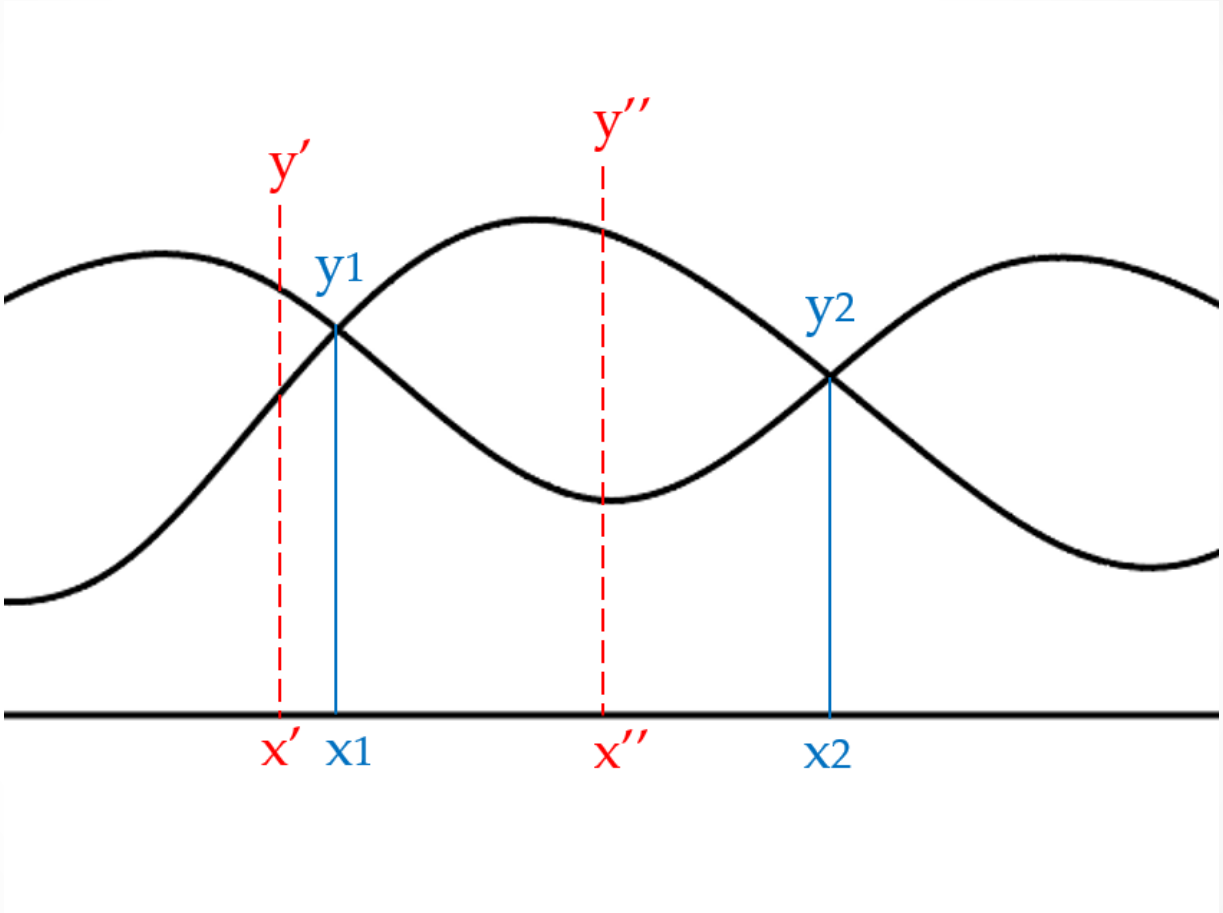
    (S)馬祖 位於 閩江 口 外 ， 島嶼 狹小 ， 又 無 高山 屏障 ， 全 年 多 風 少 雨 。

    (H)馬祖 位於 愛河 口 外 ， 全 年 多 風 少 雨 。

DNN: a continuous mapping function

Initial Point #1

Initial Point #2



Need Aggregative Features Sometimes

- **Not easy for DNN to learn comparison, aggregative features**
  - Benchmark:Y, Prediction: N

    (S)馬祖 位於 閩江 口 外 ， 全 年 多 風 少 雨 。

    (H)馬祖 位於 閩江 口 外 ， 全 年 多 風 少 雨 。
  - Without aggregative feature
    - Training Data: <A, A> -> Y; <B, B> -> Y; <C, C> -> Y
    - Test Data: <D, D> -> ?
  - With aggregative feature
    - Training Data: <A, A, A-A=0> -> Y; <B, B, B-B=0> -> Y; <C, C, C-C=0> -> Y
    - Test Data: <D, D, D-D=0> -> Y

Incapable of High Level Learning (Lin, 2017)

- **One correct prediction in the training set**
  - Benchmark:Y, Prediction: Y
    (S)馬祖 位於 閩江 口 外 ， 島嶼 狹小 ， 又 無 高山 屏障 ， 全 年 多 風 少 雨 。
    (H)馬祖 位於 閩江 口 外 ， 全 年 多 風 少 雨 。
- **Predictions for some created examples**
  - Benchmark:Y, Prediction:N
    (S)馬祖 位於 閩江 口 外 ， 全 年 多 風 少 雨 。
    (H)馬祖 位於 閩江 口 外 ， 全 年 多 風 少 雨 。

All words have valid embeddings.

  - Benchmark:N, Prediction: Y
    (S)馬祖 位於 閩江 口 外 ， 島嶼 狹小 ， 又 無 高山 屏障 ， 全 年 多 風 少 雨 。
    (H)馬祖 位於 閩江 口 外 ， 全 年 少 風 少 雨 。
    Mazu is located outside the mouth of the Minjiang River, with less wind and less rain all year round. (Google Translation)
  - Benchmark:N, Prediction: Y
    (S)馬祖 位於 閩江 口 外 ， 島嶼 狹小 ， 又 無 高山 屏障 ， 全 年 多 風 少 雨 。
    (H)馬祖 位於 愛河 口 外 ， 全 年 多 風 少 雨 。
    Mazu is located outside the mouth of Aihe, with more wind and less rain all year round. (Google Translation)

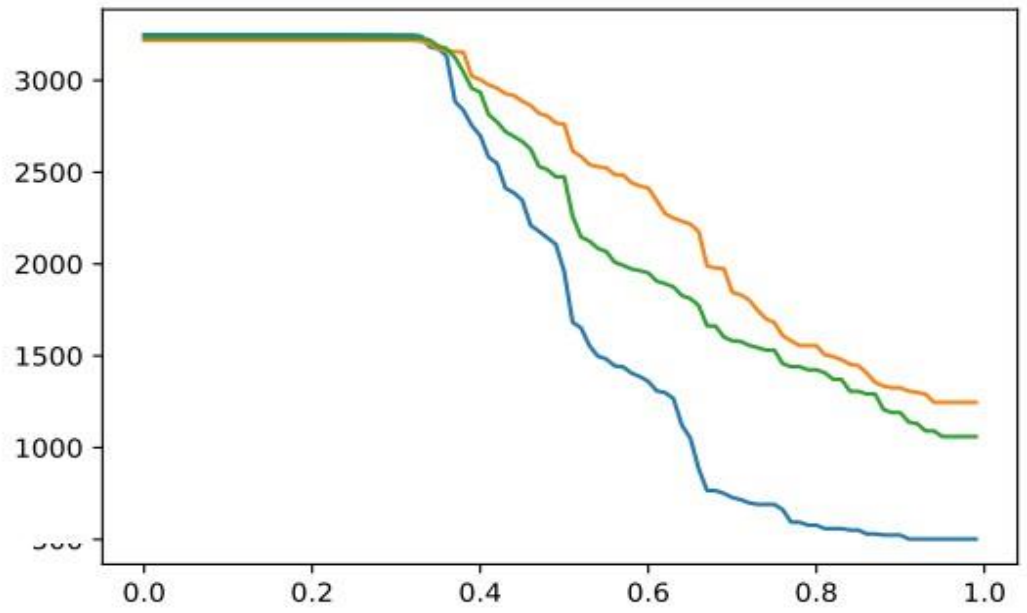
DNN: Effective Surface Learner (1/3)

- **Hypothesis Only Baselines in Natural Language Inference [Poliak et al., 2018]**

Dataset	DEV Hyp-Only	DEV MAJ	DEV |Δ|	DEV Δ%	TEST Hyp-Only	TEST MAJ	TEST |Δ|	TEST Δ%	Baseline	SOTA
Recast										
DPR	50.21	50.21	0.00	0.00	49.95	49.95	0.00	0.00	49.5	49.5
SPR	86.21	65.27	+20.94	+32.08	86.57	65.44	+21.13	+32.29	80.6	80.6
FN+	62.43	56.79	+5.64	+9.31	61.11	57.48	+3.63	+6.32	80.5	80.5
Human Judged										
ADD-1	75.10	75.10	0.00	0.00	85.27	85.27	0.00	0.00	92.2	92.2
SciTail	66.56	50.38	+16.18	+32.12	66.56	60.04	+6.52	+10.86	70.6	77.3
SICK	56.76	56.76	0.00	0.00	56.87	56.87	0.00	0.00	56.87	84.6
MPE	40.20	40.20	0.00	0.00	42.40	42.40	0.00	0.00	41.7	56.3
JOCI	61.64	57.74	+3.90	+6.75	62.61	57.26	+5.35	+9.34	–	–
Human Elicited										
SNLI	69.17	33.82	+35.35	+104.52	69.00	34.28	+34.72	+101.28	78.2	89.3
MNLI-1	55.52	35.45	+20.07	+56.61	–	35.6	– –		72.3	80.60
MNLI-2	55.18	35.22	+19.96	+56.67	–	36.5	–	–	72.1	83.21

DNN: Effective Surface Learner (2/3)

- **Hypothesis Only Baselines in Natural Language Inference [Poliak et al., 2018]**



(a) SNLI

Figure 2: Plots showing the number of sentences per ϵ $p(l|w) >= x$ for at least one label l. Colors indicate

DNN: Effective Surface Learner (3/3)

- **Hypothesis Only Baselines in Natural Language Inference [Poliak et al., 2018]**

SNLI								
Word	Score	Freq	Word	Score	Freq	Word	Score	Freq
instrument	0.90	20	tall	0.93	44	sleeping	0.88	108
touching	0.83	12	competition	0.88	24	driving	0.81	53
least	0.90	10	because	0.83	23	Nobody	1.00	52
Humans	0.88	8	birthday	0.85	20	alone	0.90	50
transportation	0.86	7	mom	0.82	17	cat	0.84	49
speaking	0.86	7	win	0.88	16	asleep	0.91	43
screen	0.86	7	got	0.81	16	no	0.84	31
arts	0.86	7	trip	0.93	15	empty	0.93	28
activity	0.86	7	tries	0.87	15	eats	0.83	24
opposing	1.00	5	owner	0.87	15	sleeps	0.95	20
(a) entailment			(b) neutral			(c) contradiction		

Outline

- DNN Limitations Observed
  - Incapable of conducting common inference
  - Need huge amount of data
  - Usually uninterpretable
  - Vulnerable to irrelevant data
- Lessons from Non-NLP Machine Learning
- Characteristics of NLP/NLU
- Various Ways to Integrate Domain Knowledge

Solving Algebraic Word Problems

- **Purely statistical approaches (Kushman etc., 2014; Roy et al., 2015):**
  - Linear algebraic questions
  - Mapping problem to a set of equations
  - Generate numerical answers

Derivation 2	
Word problem	A motorist drove 2 hours at one speed and then for 3 hours at another speed. He covered a distance of 252 kilometers. If he had traveled 4 hours at the first speed and 1 hour at the second speed, he would have covered 244 kilometers. Find two speeds?
Aligned template	$n_1 \times u_1^1 + n_2 \times u_2^1 - n_3 = 0$ $\quad$ $n_4 \times u_1^2 + n_5 \times u_2^2 - n_6 = 0$
Instantiated equations	$2x + 3y - 252 = 0$ $\quad$ $4x + 1y - 244 = 0$
Answer	$x = 48$ $y = 52$

n = number variable

u = unknown variable

Template-based (Kushman et al., 2014) : (1/2)

- **Dataset:** Algebra.com
  - Linear algebra question (#514)
- **Manually label equation**

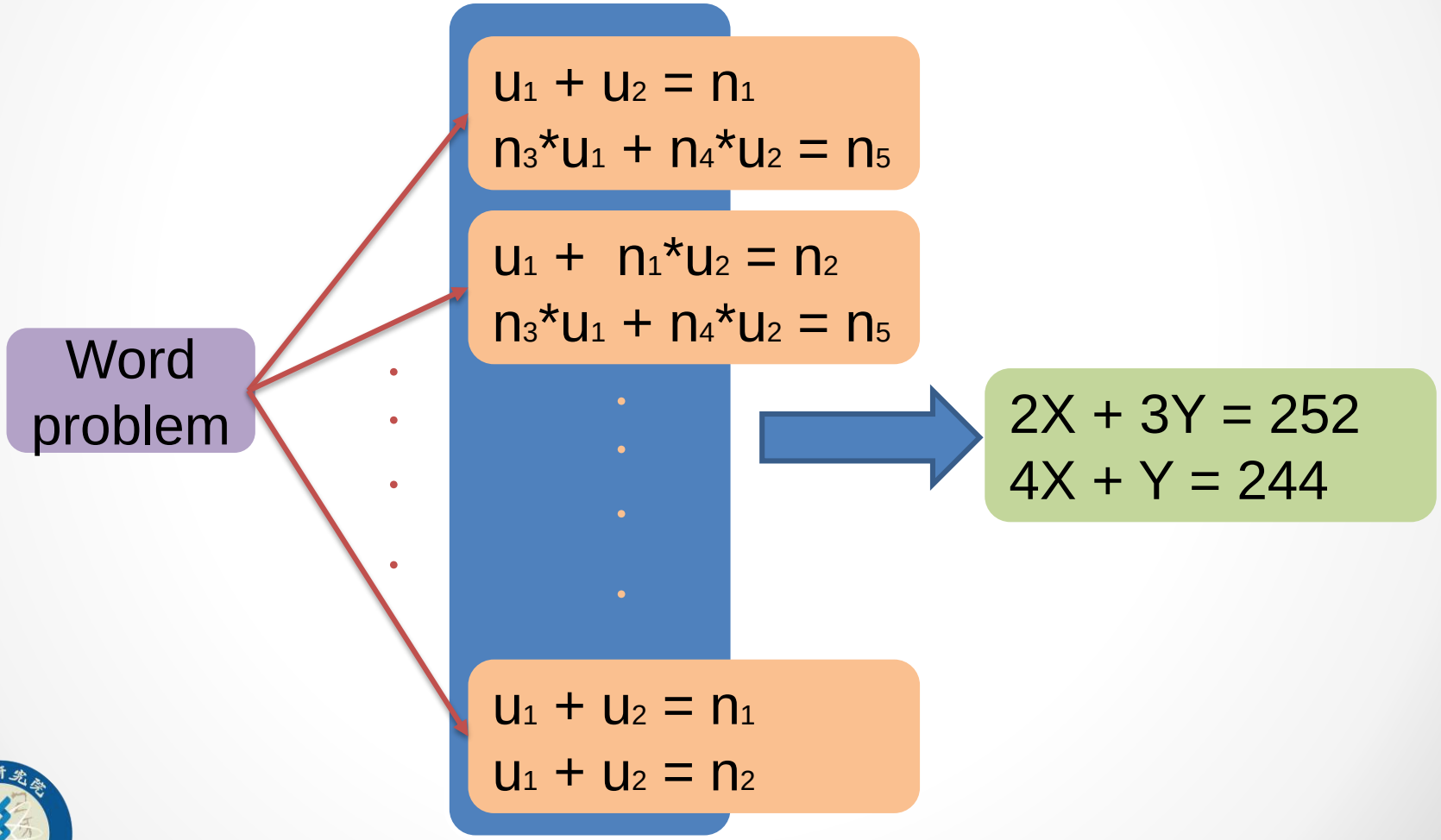
```
lEquations [2]
    0  : children+adults=278.0
    1  : (1.5*children)+(4.0*adults)=792.0
```

- **Generalize equation to a template with two rules**

$$\begin{aligned} u_1^1 + u_2^1 - n_1 &= 0 \\ n_2 \times u_1^2 + n_3 \times u_2^2 - n_4 &= 0 \end{aligned}$$

Template-based (Kushman et al., 2014) : (2/2)

- Try each template and variable alignment with a score



Solving Algebraic Word Problems

- **Purely statistical approaches (Kushman etc., 2014; Roy et al., 2015):**
  - Linear algebraic questions
  - Mapping problem to a set of equations
  - Generate numerical answers

Derivation 2	
Word problem	A motorist drove 2 hours at one speed and then for 3 hours at another speed. He covered a distance of 252 kilometers. If he had traveled 4 hours at the first speed and 1 hour at the second speed, he would have covered 244 kilometers. Find two speeds?
Aligned template	$n_1 \times u_1^1 + n_2 \times u_2^1 - n_3 = 0$ $\quad$ $n_4 \times u_1^2 + n_5 \times u_2^2 - n_6 = 0$
Instantiated equations	$2x + 3y - 252 = 0$ $\quad$ $4x + 1y - 244 = 0$
Answer	$x = 48$, $y = 52$

n = number variable

u = unknown variable

Experiment and Result

- 5-fold cross-validation (Train #400, Test #100)
- Semi-supervised
  - 5EQ: 5 seed questions annotated with equation
  - 5EQ+ANS: 5 seed questions annotated with equation + answers of remaining problems
- **Fully supervised**
  - ALLEQ: full equation systems

	Equation accuracy	Answer accuracy
5EQ	20.4	20.8
5EQ+ANS	45.7	46.1
ALLEQ	66.1	68.7

Need Huge Amount of Data (1/2)

- **A DNN approach for AWP (Ling et al., ACL-2017)**
  - Training data-set: 100,949 MWPS

Model	Perplexity	BLEU	Accuracy
Seq2Seq	524.7	8.57	20.8
+Copy Input	46.8	21.3	20.4
+Copy Output	45.9	20.6	20.2
Our Model	**28.5**	**27.2**	**36.4**

Table 3: Results over the test set measured in Perplexity, BLEU and Accuracy.

Need Huge Amount of Data (2/2)

- Learning efficiency and performance of DNN are far less than template-based approach
  - #100,949 vs. #514 MWPs
  - 36.4% vs. 68.7%

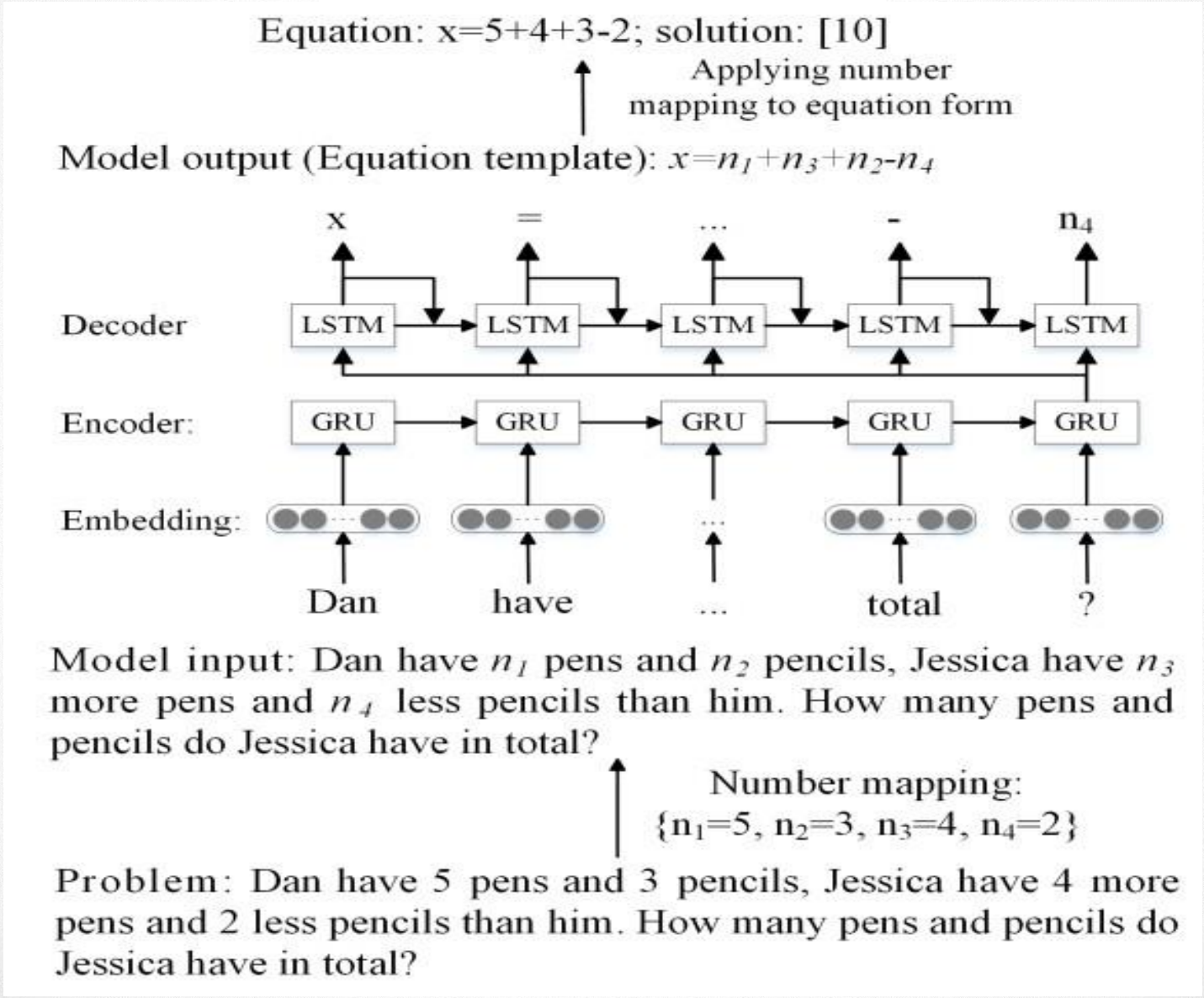
	Equation accuracy	Answer accuracy
5EQ	20.4	20.8
5EQ+ANS	45.7	46.1
ALLEQ	66.1	68.7

- People do not learn math concepts directly from those examples
  - People first learn math concepts separately, then apply them to the problems
- Rationale learnt here is not the math concept

Outline

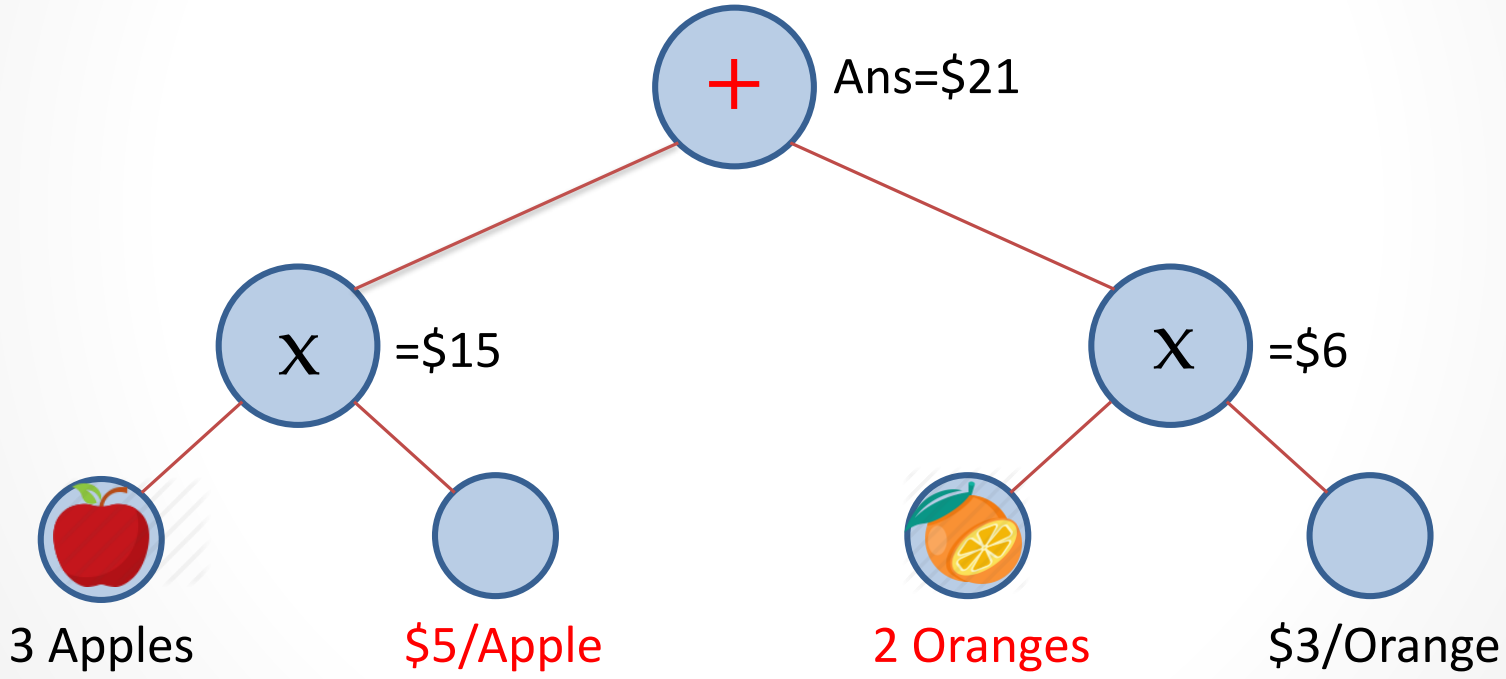
- DNN Limitations Observed
  - Incapable of conducting common inference
  - Need huge amount of data
  - Usually uninterpretable
  - Vulnerable to irrelevant data
- Lessons from Non-NLP Machine Learning
- Characteristics of NLP/NLU
- Various Ways to Integrate Domain Knowledge

Seq2Seq MWP Solver (Wang et al., 2017)



An uninterpretable case

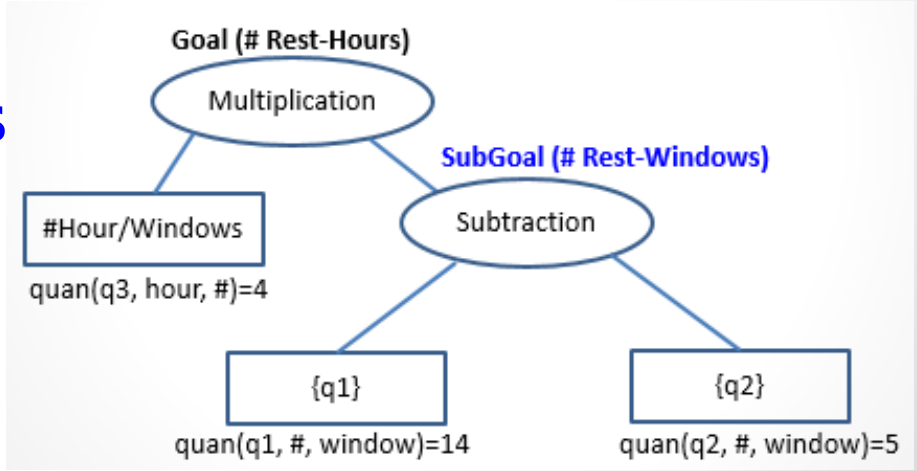
Mary bought three apples and two oranges. Each apple is priced at $5 and an orange is priced at $3. How much money that Mary need to pay?



Multi-Step MWPs

A new building *needed 14 windows*.
The builder had already *installed 5 of them*.
If it takes *4 hours to install each window*,
how long will it take him to install the rest?

- **The ANSWER could be Solved by**
  - (#Rest-Windows)
    =14-windows – 5-windows
  - (#Rest-Hours)
    =(#Rest-Windows)
    x 4 hours/window



Outline

- DNN Limitations Observed
  - Incapable of conducting common inference
  - Need huge amount of data
  - Usually uninterpretable
  - Vulnerable to irrelevant data
- Lessons from Non-NLP Machine Learning
- Characteristics of NLP/NLU
- Various Ways to Integrate Domain Knowledge

NLU Noisy Dataset

- (a.1)Additional Irrelevant quantity to a new ***Subject***
- (a.2)Additional Irrelevant quantity to a new ***Entity***
- (a.3)Additional Irrelevant quantity to a new ***Modifier***
- (a.4)***Re-Order*** the Quantities

(ORG) Tim has 10 yellow flowers and 12 red flowers. How many flowers does Tim has?
(a.1) Tim has ... **Mary** has 3 yellow flowers. How many …
(a.2) Tim has ... Tim also has 3 **books**. How many …

(ORG1) Tom has 9 yellow balloons. Sara has 8 yellow balloons. How many yellow balloons do they have in total ?
(a.3) Tom has 8 **red** balloons and …. Sara has 8 yellow balloons . How many …
(a.4) Tom has 9 yellow balloons and 8 red balloons. Sara has 8 yellow balloons . How many …

Single-step Dataset Performance

	AI2	ILDS	NDS1
Our System*[1]	**81.5**	**81.0**	**82.1**
Roy2015 (MSMWP*[2])	78.0	73.9	28.5*[4]
Hosseini 2014	77.7	-	-
Roy2015 (SMWP*[3])	-	52.7	-
Kushman 2014	64.0	73.7	-

*1: ([Liang *et al.*, 2017-2])
*2: MSMWP: multiple-step MWP solving system ([Roy and Roth, 2015])
*3: SMWP: single-step MWP solving system ([Roy *et al.*, 2015])
*4: We submit MWPs to Illinois Math Solver (https://cogcomp.cs.illinois.edu/page/demo_view/Math) in May and June, 2017.

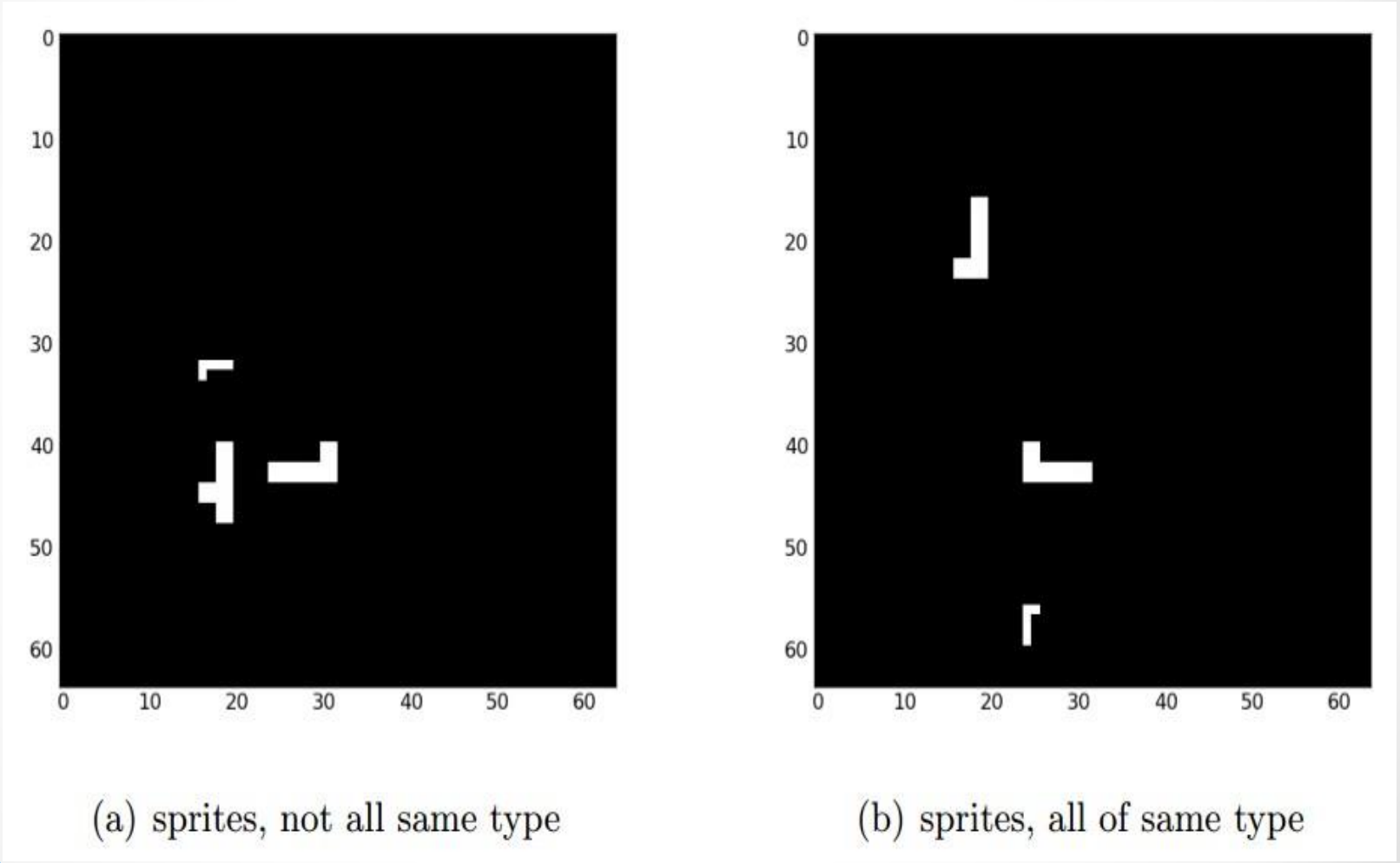
Outline

- DNN Limitations Observed
  - Incapable of high level learning
  - Need huge amount of data
  - Usually uninterpretable
  - Vulnerable to irrelevant data
- Lessons from Non-NLP Machine Learning
- Characteristics of NLP/NLU
- Various Ways to Integrate Domain Knowledge

Knowledge Matters (Gulcehre and Bengio, 2013)

- Humans and animals can perform much more complex tasks than they can acquire using pure trial and error learning
  - This gap is filled by teaching
  - Shaping (Krueger and Dayan, 2009): a teacher decomposes a complete task into sub-components, thereby providing an easier path to learning
  - Curriculum Learning (Bengio et al., 2009): Humans and animals learn much better when the examples are not randomly presented but organized in a meaningful order
    - It achieves faster training in the online setting, and guides training towards better regions in parameter space.

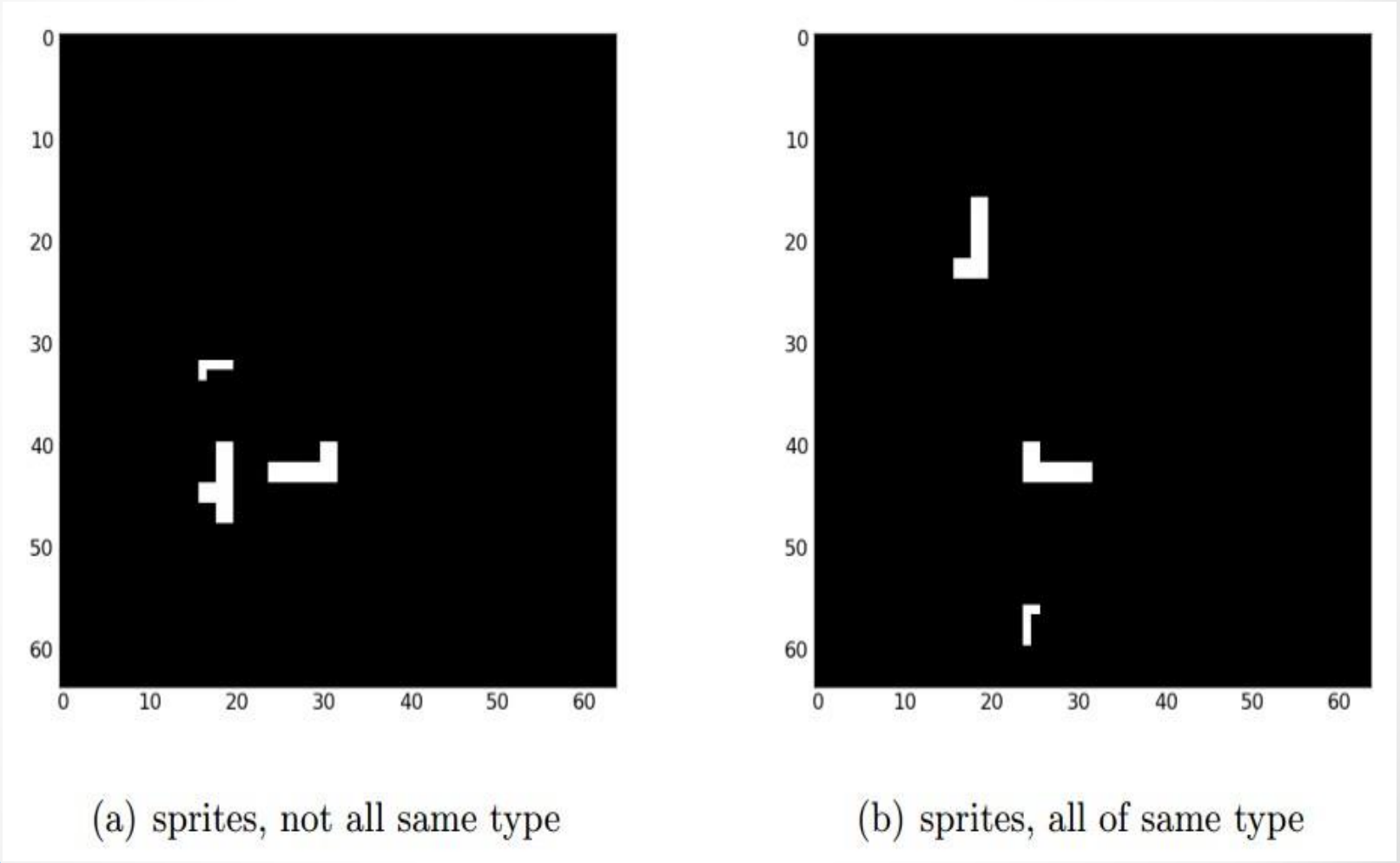
Task (1/3) (Knowledge Matters; Gulcehre and Bengio, 2013)



Task(2/3) (Knowledge Matters; Gulcehre and Bengio, 2013)

- **五格骨牌**（Pentomino），又稱**五連塊**，
  - 每塊以五個全等的正方形連成，反射或旋轉視作同一種共有十二種，可以英文字母代表。

Task (1/3) (Knowledge Matters; Gulcehre and Bengio, 2013)



SMLP (2/3; Gulcehre and Bengio, 2013)

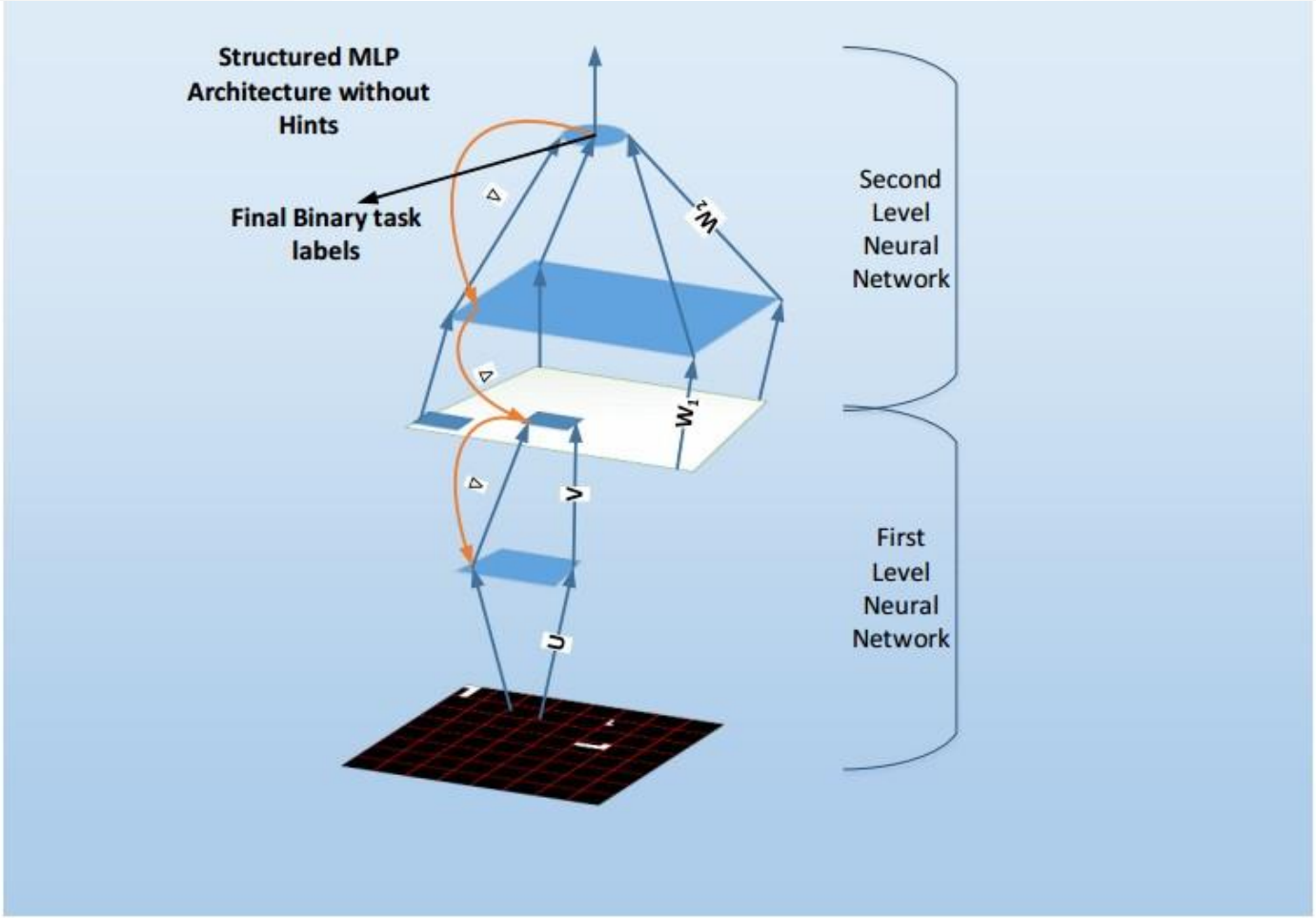


Figure 7: Structured MLP architecture, used without hints (SMLP-nohints). It is the same architecture as SMLP-hints (Figure 5) but with both parts (P1NN and P2NN) trained jointly with respect to the final binary classification task.

SMLP (3/3; Gulcehre and Bengio, 2013)

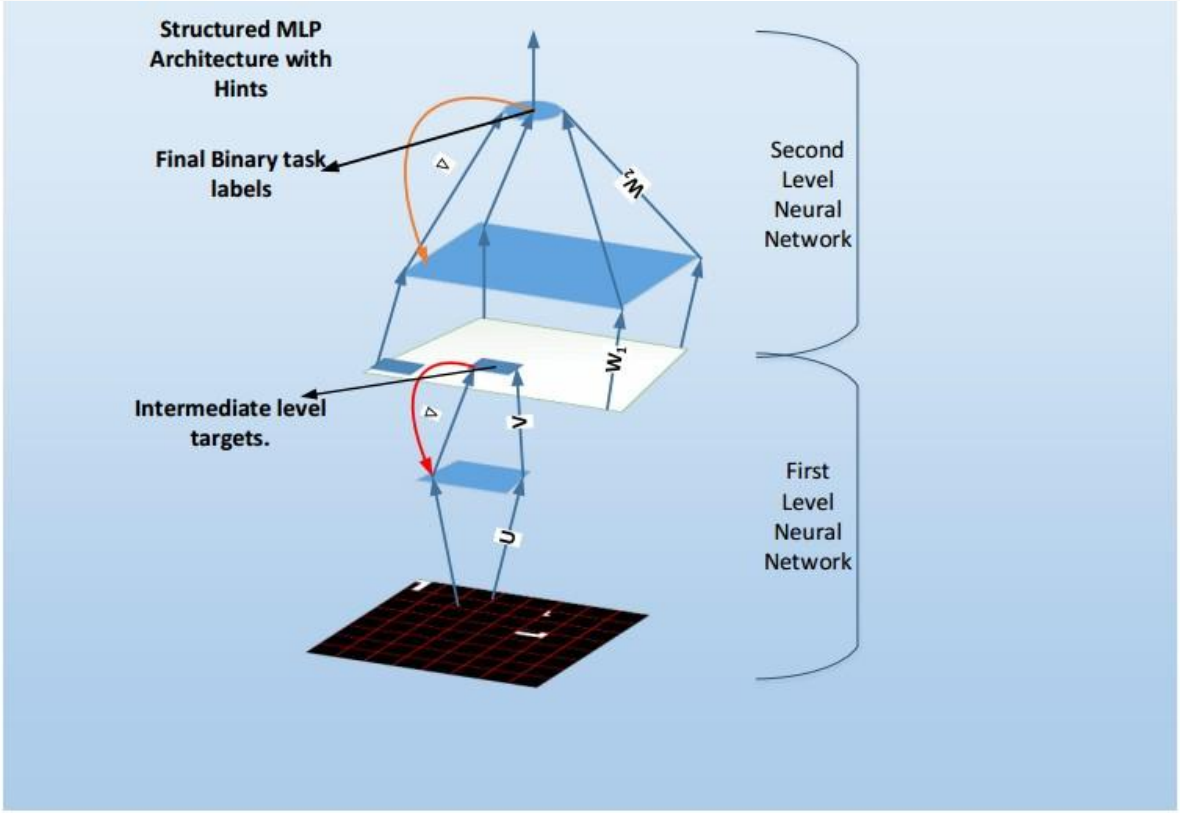


Figure 5: Structured MLP architecture, used with hints (trained in two phases, first P1NN, bottom two layers, then P2NN, top two layers). In SMLP-hints, P1NN is trained on each 8x8 patch extracted from the image and the softmax output probabilities of all 64 patches are concatenated into a 64×11 vector that forms the input of P2NN. Only U and V are learned in the P1NN and its output on each patch is fed into P2NN. The first level and the second level neural networks are trained separately, not jointly.

Knowledge Matters (5/8; Gulcehre and Bengio, 2013)

Algorithm	20k dataset		40k dataset		80k dataset	
	Training Error	Test Error	Training Error	Test Error	Training Error	Test Error
SVM RBF	26.2	50.2	28.2	50.2	30.2	49.6
K Nearest Neighbors	24.7	50.0	25.3	49.5	25.6	49.0
Decision Tree	5.8	48.6	6.3	49.4	6.9	49.9
Randomized Trees	3.2	49.8	3.4	50.5	3.5	49.1
MLP	26.5	49.3	33.2	49.9	27.2	50.1
Convnet/Lenet5	50.6	49.8	49.4	49.8	50.2	49.8
Maxout Convnet	14.5	49.5	0.0	50.1	0.0	44.6
2 layer sDA	49.4	50.3	50.2	50.3	49.7	50.3
Struct. Supervised MLP w/o hints	0.0	48.6	0.0	36.0	0.0	12.4
Struct. MLP+CAE Supervised Finetuning	50.5	49.7	49.8	49.7	50.3	49.7
Struct. MLP+CAE+DAE, Supervised Finetuning	49.1	49.7	49.4	49.7	50.1	49.7
Struct. MLP+DAE+DAE, Supervised Finetuning	49.5	50.3	49.7	49.8	50.3	49.7
Struct. MLP with Hints	0.21	30.7	0	3.1	0	0.01

Table 1: The error percentages with different learning algorithms on Pentomino dataset with different number of training examples.

Knowledge Matters (6/8; Gulcehre and Bengio, 2013)

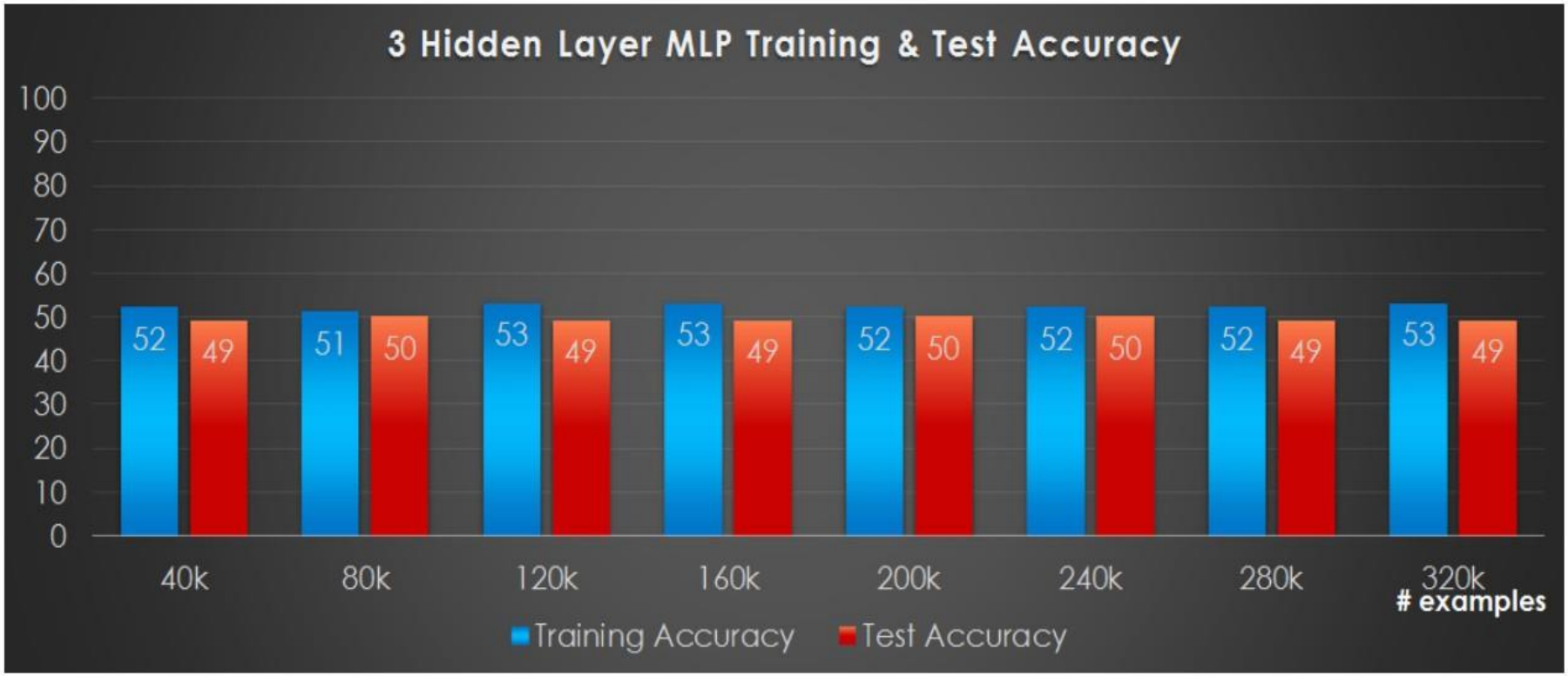


Figure 14: Training and test error bar charts for a regular MLP with 3 hidden layers. There is no significant improvement on the generalization error of the MLP as the new training examples are introduced.

Knowledge Matters (7/8; Gulcehre and Bengio, 2013)

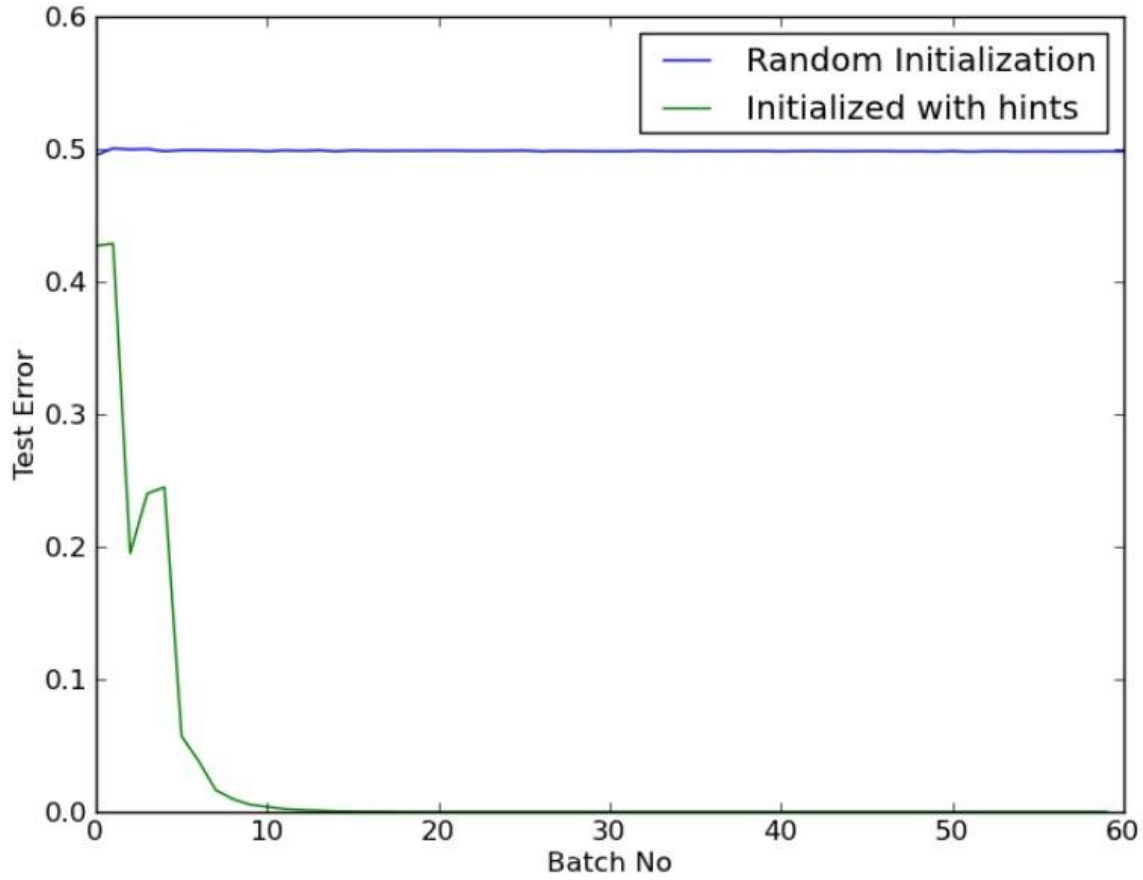


Figure 15: Plots showing the test error of SMLP with random initialization vs initializing with hint based training.

Some Observations (Bengio, 2009; 2013): 1/2

- Training deep architectures is easier if hints are provided about the function that intermediate levels should compute
- It is much easier to teach a network with supervised learning (i.e., concept) than to expect unsupervised learning to discover the concept
- It is more difficult to train deeper architectures with more abstract concepts

Some Observations (Gulcehre and Bengio, 2013): 2/2

- Directly training all the layers of a deep network together makes it difficult to exploit all the extra modeling power of a deeper architecture
- Black-box machine learning algorithms only got chance
  - It is hypothesized that the learning difficulty is due to the composition of two highly non-linear tasks
- More abstract learning tasks are more likely to yield effective local minima for neural networks
- Using a particular structure and guiding the learner allows to nail the task.

Select Learning Algm. (Bengio and LeCun, 2007)

- The quest for a completely general learning method is doomed to failure
  - We need to search for learning models that are well suited for a particular type of tasks.
- Prefer a deep network, with small dose of prior knowledge embedded in the architecture
  - Combined with a learning algorithm that can deal with millions of examples

Outline

- DNN Limitations Observed
  - Incapable of conducting common inference
  - Need huge amount of data
  - Usually uninterpretable
- Lessons from Non-NLP Machine Learning
- Characteristics of NLP/NLU
- Various Ways to Integrate Domain Knowledge

NLP vs. Speech/Vision Recognition

- Animal can watch/listen, but only human can communicate with language.
- Symbolic Processing
  - Features and clues are more comprehensible
- Require more external knowledge
  - Data alone is not enough
- More knowledge is learnt from teachers
  - Domain Concept is essential
- Inference is more Abstract
  - More difficult for DNN

Lexicon Relationship

- Three girls and two boys enter a classroom. How many people are there?
  - Need to know: "girl -> people"
    "boy -> people"
- Weakly supervised learning can catch the relation from the training data
  - Coverage rate is small
- WordNet could supplement the training data

Common Sense (facts)

- 小綠和阿夫猜拳，小綠出剪刀，阿夫出布，兩人共伸出了幾根手指頭？
- Xiao-Green and A-fu play "Rock Paper Scissors", Xiao-Green out of the scissors, A-fu out of the paper, how many fingers did the two stick out? (Google Translation)
  - Need to know: “scissor -> 2 fingers”
    “paper -> 5 fingers”

- **姐姐搭車到圖書館，從8點20分上車到9點10分下車，經過了幾分鐘？**
- The sister took a ride to the library and got on at 8:20 to get off at 9:10. How many minutes have passed? (Google Translation)
  - Need to know: “1 hr = 60 minutes”

Common Sense (reasoning): 1/2

- train.G1.03.000390
  有9人排隊讓老師改作業，老師改好了4個小朋友的作業，剩下幾人在排隊？
- There were 9 people waiting in line for the teacher to modify their homework. The teacher modified the homework of 4 children. How many people are waiting in line? (Google Translation)
  - Need to know: “改好作業” 表示”離開排隊的隊伍”

    “modified the homework” means “Leaving the queued team”

- Dan 's cat had kittens. He gave 7 to Tim and 4 to Jason . He now has 5 kittens . How many kittens did he have to start with ?
  - Implicit implication: the meaning of “Dan's cat had kittens” equals “Dan had kittens” in this problem

Common Sense (reasoning): 2/2

- Beijing office called us yesterday (Analogy derived from Jerry Hobb's example)
  - It is not a building in Beijing that gave us a call
  - But the people who is in Beijing branch (of our company) gave us a call

- This is the key problem that AI/DNN needs to solve (for NLU task)

Domain Knowledge

- 長３公尺、寬２公尺的長方形，面積是多少平方公尺？
- A rectangle of 3 meters in length and 2 meters in width. What is the square meter? (Google Translation)
  - Need to know: "Rectangle-Area = Length x Width"

- 1包鐵蛋有8個，點點分**8分之3**包鐵蛋給妹妹，妹妹可以分到幾個鐵蛋？
- 1 package iron egg have eight, Little-Little divided 3/8 package iron egg to the sister, how many iron egg can the sister be assigned? (Google Translation)
  - Need to know: Physical meaning of "Fraction"

NLP vs. Speech/Vision Recognition

- Animal can watch/listen, but only human can communicate with language.
- Symbolic Processing
  - Features and clues are more comprehensible
- Require more external knowledge
  - Data alone is not enough
- More knowledge is learnt from teachers
  - Domain Concept is essential
- Inference is more Abstract
  - More difficult for DNN

NLP Remains the Domain Problems

- Christ Manning made the following claims (2015):
  - "*The gains so far have not so much been from true Deep Learning as from the use of distributed word representations*"
  - "*problems in higher-level language processing have not seen the dramatic error rate reductions from deep learning that have been seen in speech recognition and in object recognition in vision….and this remains the case*"
  - "*Our field is the domain science of language technology; it's not about the best method of machine learning—the central issue remains the domain problems*"

Outline

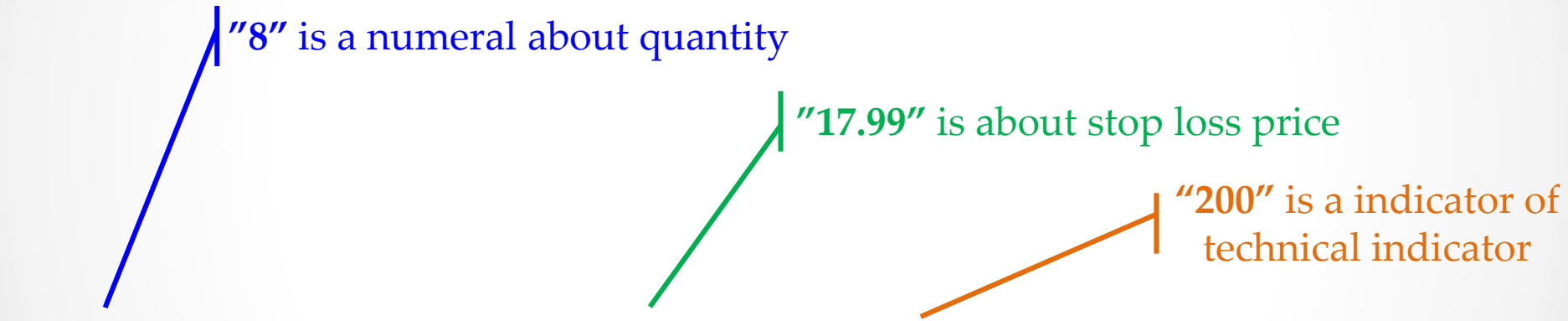
- DNN Limitations Observed
  - Incapable of conducting common inference
  - Need huge amount of data
  - Usually uninterpretable
  - Vulnerable to irrelevant data
- Lessons from Non-NLP Machine Learning
- Characteristics of NLP/NLU
- Various Ways to Integrate Domain Knowledge

Integrate Domain-Concept into DNN (1/3)

- Preprocess the input text to handle OOV (UNK)
  - E.g., convert personal-names and numerical values into meta-tokens in MWP
- Incorporate additional human-crafted features into DNN
  - E.g., add Exactly-the-Same feature in Q&A task
- Adopt existing linguistic resources as external knowledge
  - E.g., *WordNet* and *ConceptNet*
- Adopt a better intermediate representation to train DNN
  - Decompose the task into several meaningful sub-tasks
  - E.g., *Logic predicate* for MWP

FinNum Classification Task

- Purpose: To understand the fine-grained numeral information in financial Tweet



Pre-processing

- Apply Normalization Processes before the diversity evaluation
  - Stop word removal
  - Named entity (person names and quantity values) normalization

"Sara has 16 red flowers and John has 24 yellow flowers."

"PERSON-1 has Quantity-1 red flowers and PERSON-2 has Quantity-24 yellow flowers."

Experimental Results

	CNN		RNN		RNN+CNN	
	Micro	Macro	Micro	Macro	Micro	Macro
None	81.83	69.54	84.22	73.36	82.71	69.63
+POS&NE	88.21	79.14	88.45	78.63	**89.72**	**80.93**

Task-1 testing set performance

- OOVs provide no useful Information
  - OOVs: 30+% on Development and Test sets
- Linguistic Information (POS&NE) attached to OOVs improved the performance significantly (4% ~ 10%).

Integrate Domain-Concept into DNN (1/3)

- Preprocess the input text
  - E.g., convert personal-names and numerical values into meta-tokens in MWP
- Incorporate additional human-crafted features into DNN
  - E.g., add Exactly-the-Same feature in Q&A task
- Adopt existing linguistic resources as external knowledge
  - E.g., *WordNet* and *ConceptNet*
- Adopt a better intermediate representation to train DNN
  - Decompose the task into several meaningful sub-tasks
  - E.g., *Logic predicate* for MWP

Add Aggregative Features

- "*If it can be exactly matched to a word in the Question, either in its original, lowercase or lemma form*" (Chen et al., 2017)
  - For their DNN-based open domain Q&A task.
- "*If physical units of operands match each other*"
  - For the DNN in the MWP task
- "*If the evidence and the hypothesis are exactly the same*"
  - For the DNN in the MRC task

DNN: a continuous mapping function (2)

- Training data: <S, H, A>
  - S: Supporting-Evidence; H: Hypothesis; A: Answer (Entail, Neural and Contradictory)
  - S = [Left, John, Mary], H = [Left, Mary, John], A = Contradictory (i.e., [0, 0, 1])
- Test data
  - S-1 = [Left, Jack, Jane], H-1 = [Left, Jane, Jack], A-1 = Contradictory ([0, 0, 1])]); Prediction: [0, 0, 1]
  - S-2 = [Left, Jack, Jane], H-2 = [Left, Jane, John], A-2 = Neural ([0, 1, 0]); Prediction: [0, 0, 1]
  - S-3 = [Left, table, chair], H-3 = [Left, chair, table], A-3 = Contradictory ([0, 0, 1]); Prediction: Unpredictable
  - S-4 = [Above, table, chair], H-4 = [Above, chair, table], A-4 = Contradictory ([0, 0, 1]); Prediction: Unpredictable

How fast can BERT learn Binary Predicate?

- Fine-tune BERT to understand the predicate "left(x, y)"
  - "left(john, mary)" means "John is on the left side of Mary"
  - Trained from "BERT-Base, Uncased"
    - 12-layer, 768-hidden, 12-heads, 110M parameters
- Name sets
  - η_L, η_H, η_c are sets of personal names
    - "nikki" is the centroid of 1465 collected personal names
    - η_L is a set of 147 names with low similarities to "nikki".
    - η_H is a set of 147 names selected from the remaining 1318 names which have high similarities to "nikki"
    - η_c is a set of the remaining 1171 names
  - η_F is a set of 10 words related to food and tableware
  - η_A is a set of 10 words related to animals

Learning Efficiency w/o Aggr. Feat.

Separator token

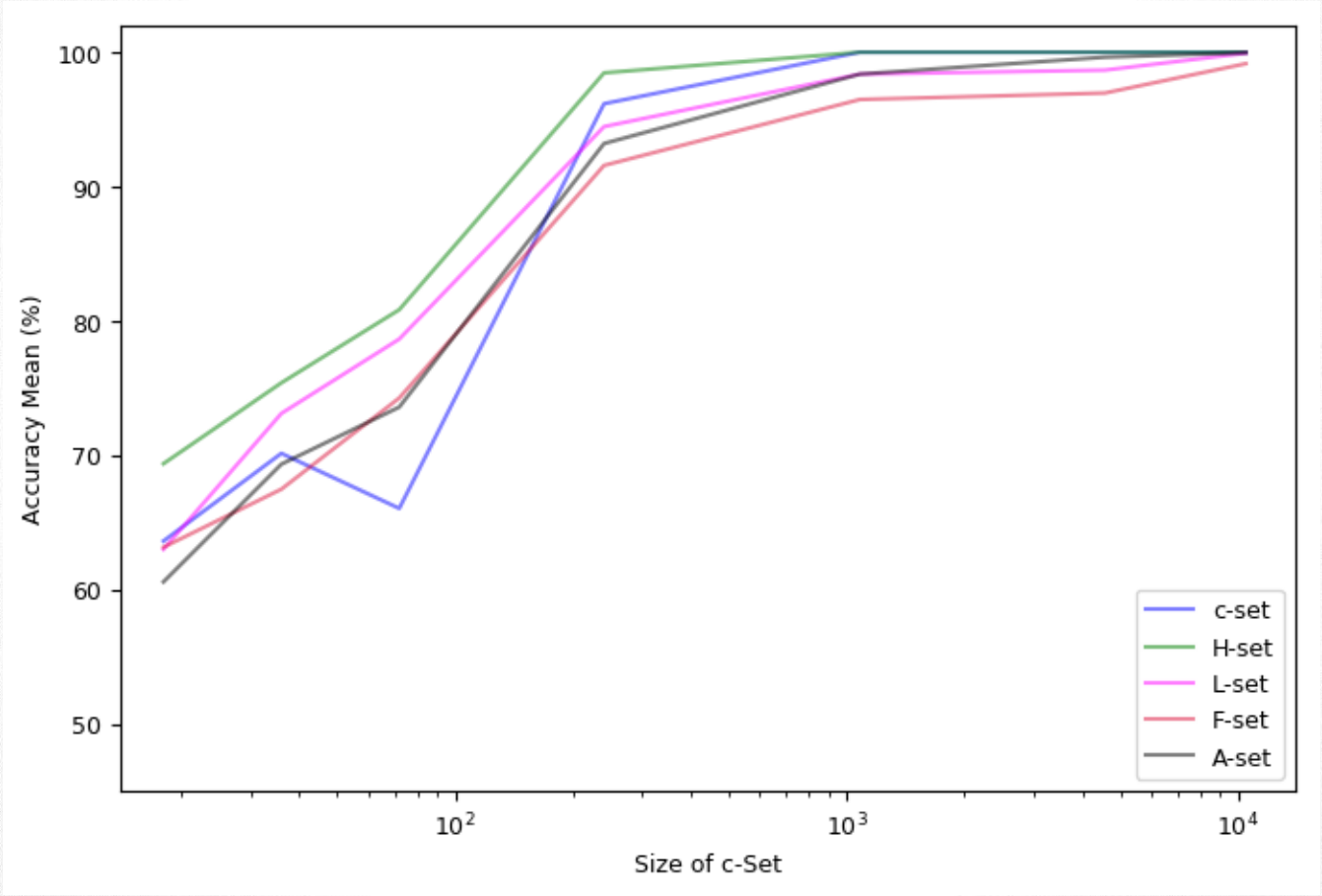
E: Entailment
C: Contradiction
N: Neutral

- Exp-B
  - Templates (supporting-evidence [sep] hypothesis ➔ label)
    - left x y [sep] left x y ➔ E, left x y [sep] left y x ➔ C, left x y [sep] left y z ➔ N, …
  - Datasets
    - c/H/L/F/A-set is generated by fill the templates with $x, y, z \in \eta_{\mathrm{c/H/L/F/A}}$
- Exp-C2
  - Templates (supporting-evidence [sep] hypothesis [sep] F_1 F_2 ➔ label)
    - left x y [sep] left x y [sep] true true ➔ E, left x y [sep] left y x [sep] false false ➔ C, left x y [sep] left y z [sep] false false ➔ N, …
- Exp-C1
  - Templates (supporting-evidence [sep] hypothesis [sep] F ➔ label)
    - left x y [sep] left x y [sep] true ➔ E, left x y [sep] left y x [sep] false ➔ C, left x y [sep] left y z [sep] fuzzy ➔ N, …

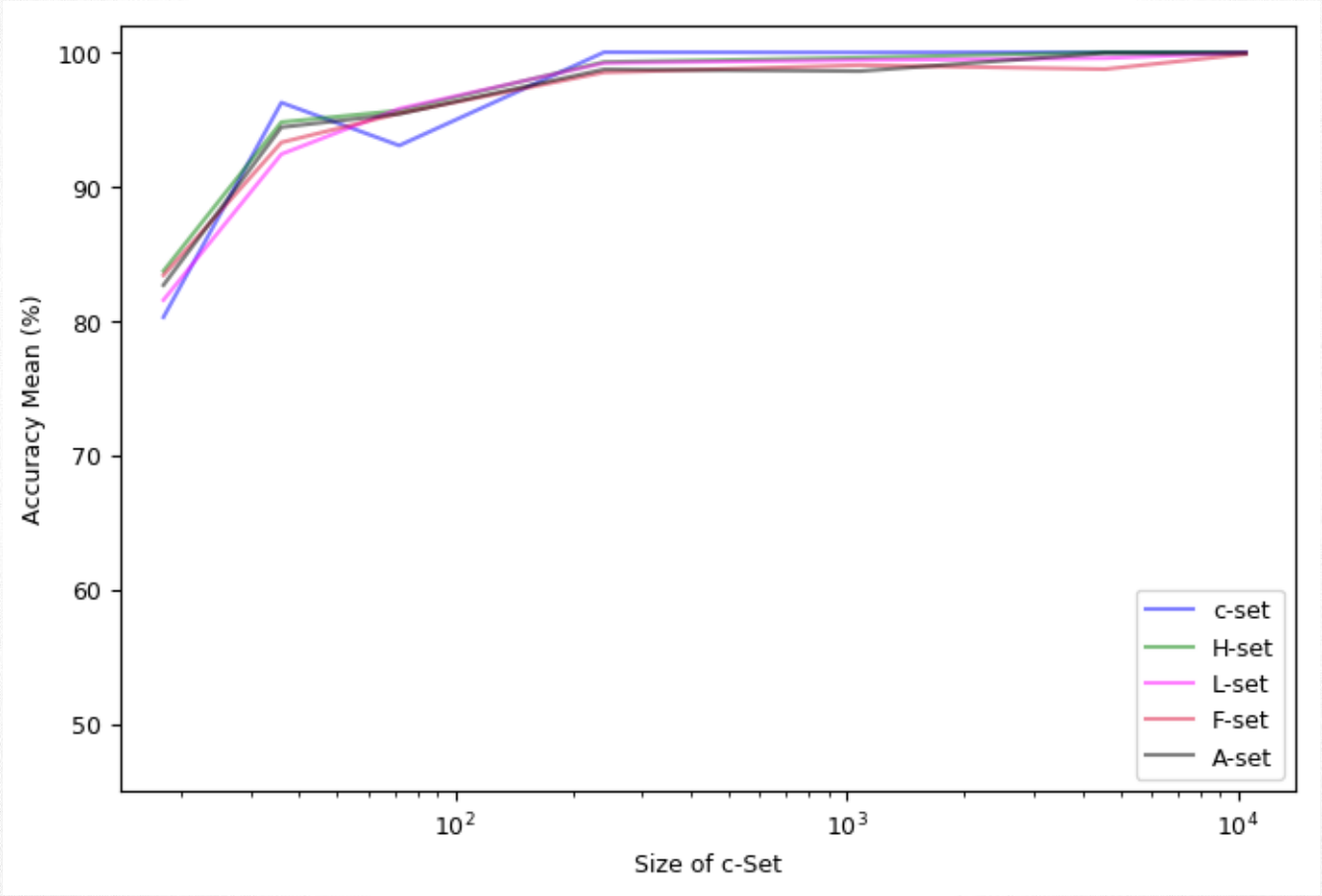
F_i is "true" if the i-th predicate arguments of the premise and the hypothesis are equal. Otherwise, F_i is "false".

F is "true" if the argument pairs are equal. F is "false" if argument pairs are swapped, Otherwise, F is "fuzzy".

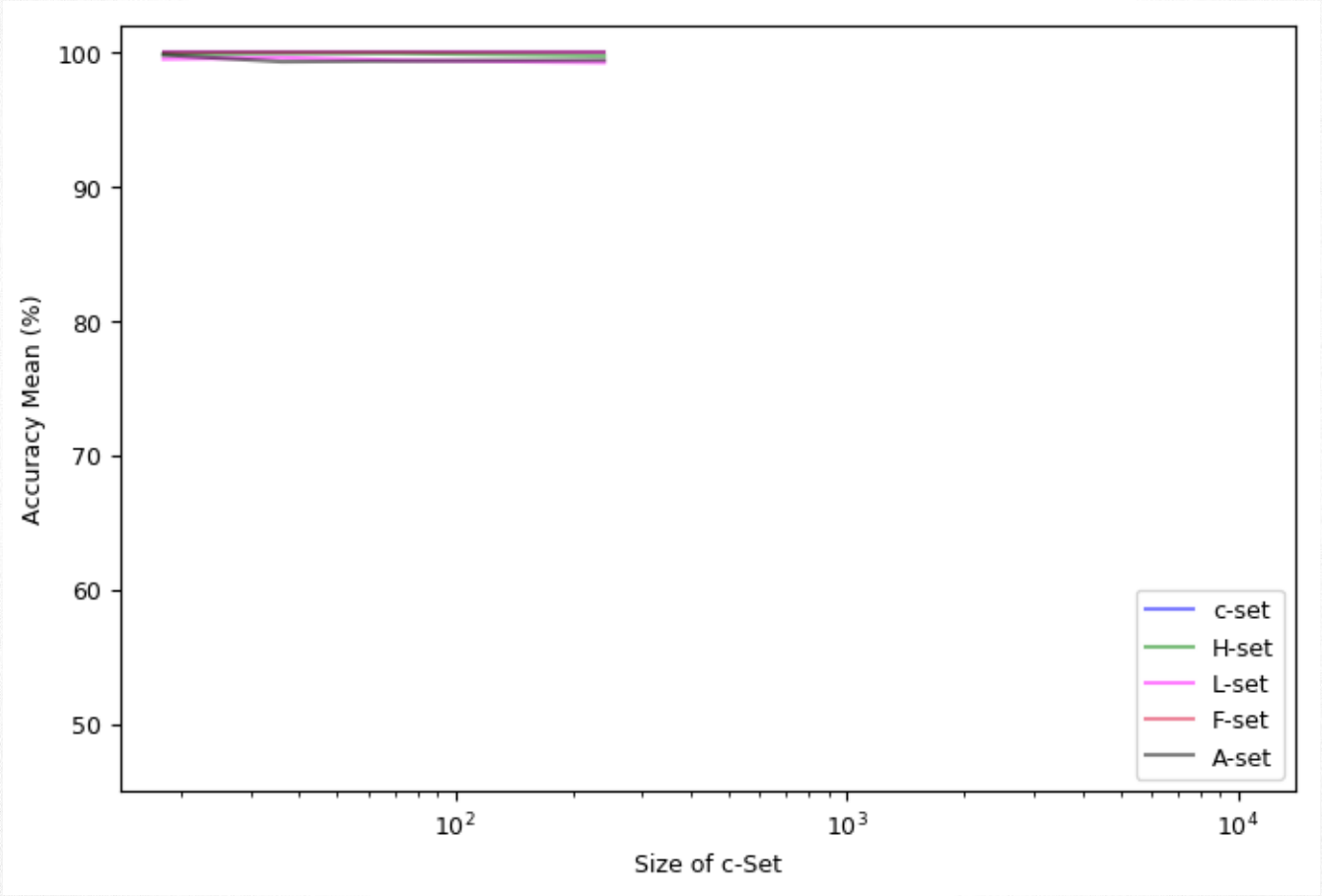
F-Best Performances of Exp-B



F-Best Performances of Exp-C2



F-Best Performances of Exp-C1



F-Best Performances of Exp-B/C2/C1

Num. of training examples	Set	Accuracy (%)					
		Exp-B		Exp-C2		Exp-C1	
		μ	σ	μ	σ	μ	σ
18	c-set	63.6	*5.2*	80.3	*4.8*	100.0	*0.0*
	H-set	69.4	*3.8*	83.7	*2.6*	99.9	*0.1*
	L-set	63.0	*4.0*	81.6	*2.7*	99.5	*0.2*
	F-set	63.1	*3.6*	83.4	*2.3*	100.0	*0.0*
	A-set	60.6	*3.6*	82.7	*2.5*	99.8	*0.1*
36	c-set	70.1	*5.8*	96.3	*1.8*	100.0	*0.0*
	H-set	75.4	*3.8*	94.8	*1.5*	99.9	*0.1*
	L-set	73.1	*3.4*	92.4	*1.6*	99.6	*0.2*
	F-set	67.5	*3.0*	93.3	*1.3*	100.0	*0.0*
	A-set	69.3	*3.1*	94.4	*1.0*	99.3	*0.5*
72	c-set	66.0	*5.8*	93.1	*2.8*	100.0	*0.0*
	H-set	80.8	*3.3*	95.7	*1.4*	99.9	*0.1*
	L-set	78.6	*3.2*	95.8	*1.1*	99.4	*0.3*
	F-set	74.3	*2.4*	95.4	*1.0*	100.0	*0.0*
	A-set	73.6	*2.8*	95.4	*1.0*	99.3	*0.4*
240	c-set	96.2	*3.3*	100.0	*0.0*	100.0	*0.0*
	H-set	98.5	*1.2*	99.3	*0.5*	99.7	*0.2*
	L-set	94.5	*1.6*	99.2	*0.4*	99.2	*0.5*
	F-set	91.6	*1.1*	98.5	*0.4*	100.0	*0.0*
	A-set	93.2	*1.7*	98.7	*0.5*	99.4	*0.4*

F-Best Performances of Exp-B/C2/C1 (cont'd.)

Num. of training examples	Set	Accuracy (%)					
		Exp-B		Exp-C2		Exp-C1	
		μ	σ	μ	σ	μ	σ
1080	c-set	100.0	*0.0*	100.0	*0.0*		
	H-set	100.0	*0.0*	99.6	*0.4*		
	L-set	98.4	*0.2*	99.4	*0.2*		
	F-set	96.5	*1.2*	99.0	*0.4*		
	A-set	98.4	*0.7*	98.6	*0.8*		
4560	c-set	100.0	*0.0*	100.0	*0.0*		
	H-set	100.0	*0.0*	100.0	*0.0*		
	L-set	98.7	*0.3*	99.6	*0.1*		
	F-set	97.0	*0.9*	98.7	*0.4*		
	A-set	99.6	*0.2*	99.9	*0.0*		
10440	c-set	100.0	*0.0*	100.0	*0.0*		
	H-set	100.0	*0.0*	100.0	*0.0*		
	L-set	99.9	*0.0*	100.0	*0.0*		
	F-set	99.1	*0.3*	99.9	*0.1*		
	A-set	100.0	*0.0*	99.9	*0.0*		

Integrate Domain-Concept into DNN (1/3)

- Preprocess the input text
  - E.g., convert personal-names and numerical values into meta-tokens in MWP
- Incorporate additional human-crafted features into DNN
  - E.g., add Exactly-the-Same feature in Q&A task
- Adopt existing linguistic resources as external knowledge
  - E.g., *WordNet* and *ConceptNet*
- Adopt a better intermediate representation to train DNN
  - Decompose the task into several meaningful sub-tasks
  - E.g., *Logic predicate* for MWP

Common-Sense Q&A (Talmor et al., 2019)

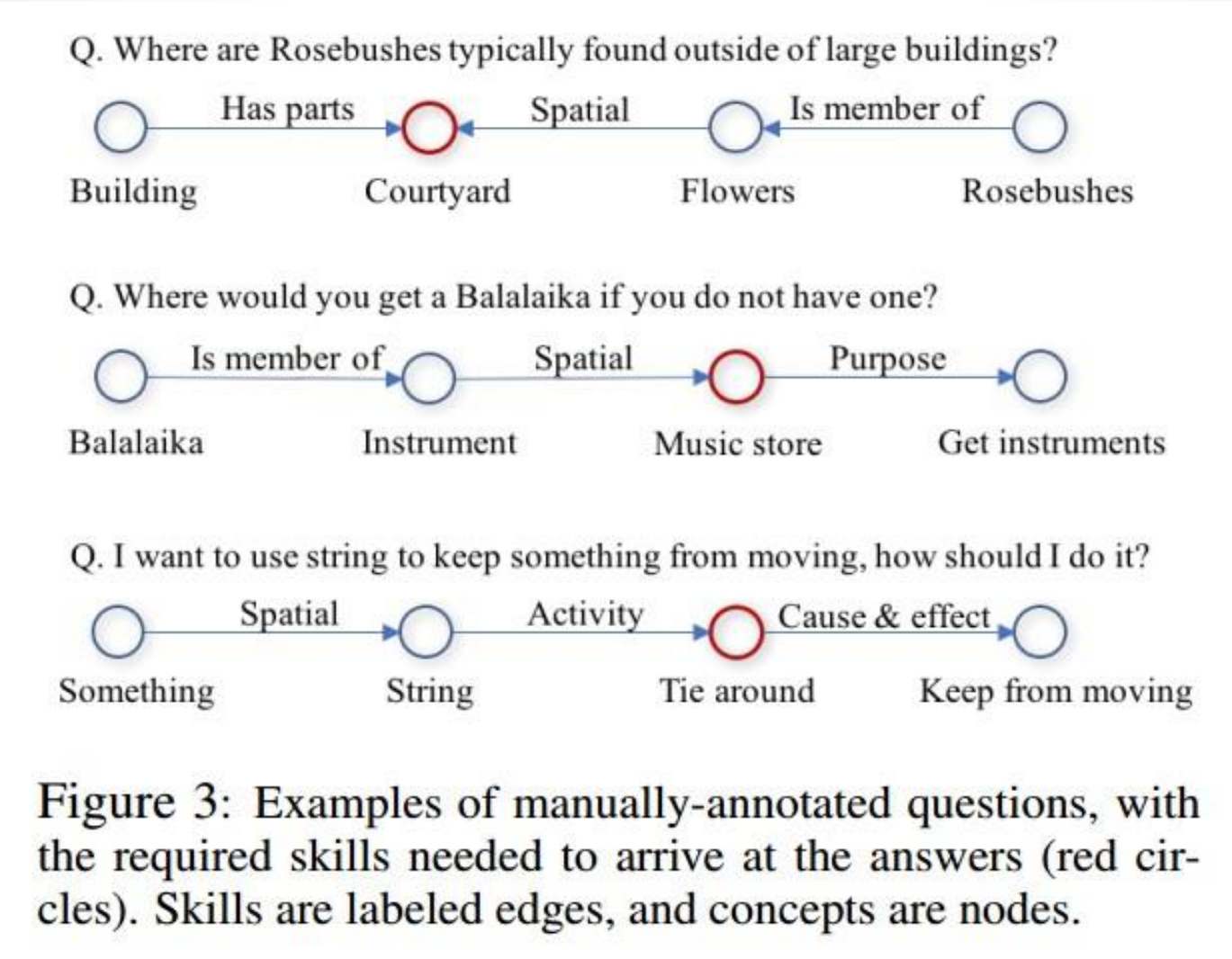


Figure 3: Examples of manually-annotated questions, with the required skills needed to arrive at the answers (red circles). Skills are labeled edges, and concepts are nodes.

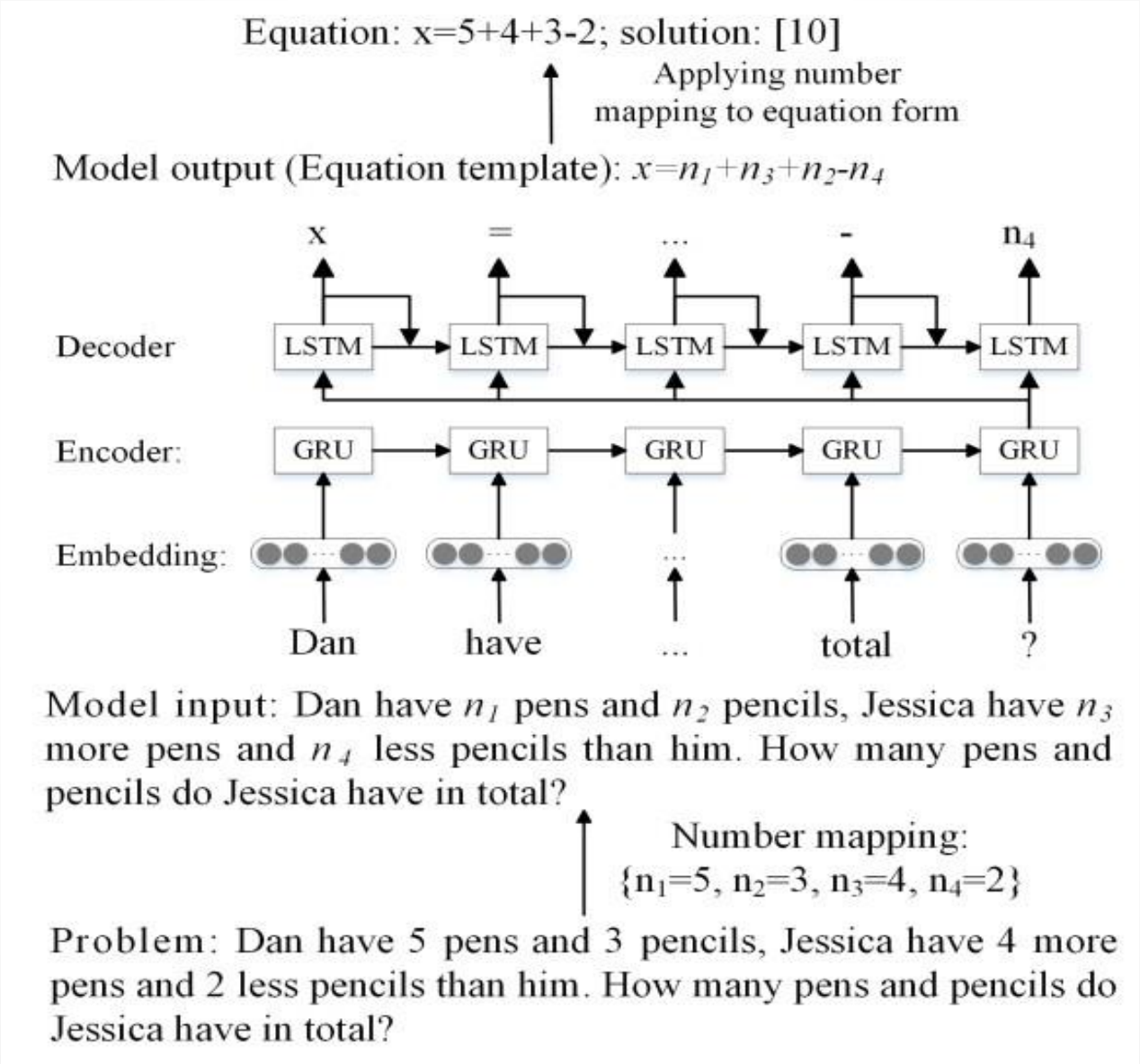
Integrate Domain-Concept into DNN (1/3)

- Preprocess the input text
  - E.g., convert personal-names and numerical values into meta-tokens in MWP
- Incorporate additional human-crafted features into DNN
  - E.g., add Exactly-the-Same feature in Q&A task
- Adopt existing linguistic resources as external knowledge
  - E.g., *WordNet* and *ConceptNet*
- Adopt a better intermediate representation to train DNN
  - Decompose the task into several meaningful sub-tasks
  - E.g., *Logic predicate* for MWP

If Intermediate Repr. is Meaningful

- Inference will be explainable
  - E.g., alignment
- Learning will be more efficient
  - Shaping Learning
- Generalization Capability will be higher
  - Better Initialization
- Debugging will be easier
  - Performance could be checked stage by stage

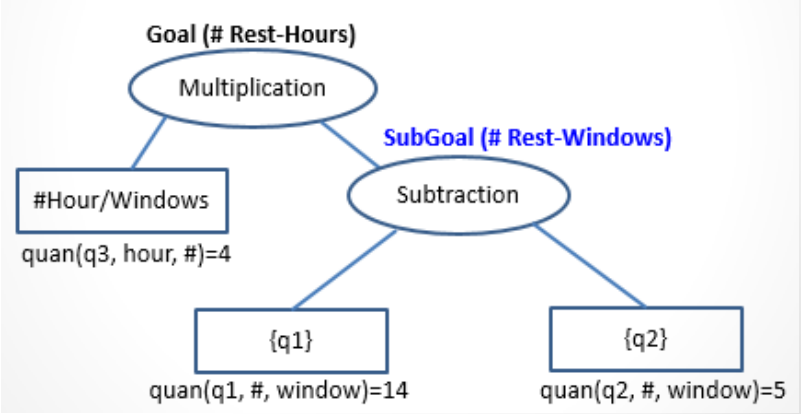
Seq2Seq MWP Solver (Wang et al., 2017)



Multi-Step MWPs (2)

A new building *needed 14 windows*.
The builder had already *installed 5 of them*.
If it takes *4 hours to install each window*,
how long will it take him to install the rest?

- **The ANSWER could be Solved by**
  - (#Rest-Windows)
    =14-windows – 5-windows
  - (#Rest-Hours)
    =(#Rest-Windows)
    x 4 hours/window



Integrate Domain-Concept into DNN (2/3)

- Adjust DNN architecture toward human reasoning style
  - E.g., add an alignment layer to DNN in entailment judgment
- Add preferences implied by domain knowledge during DNN training process
  - E.g., add human preferred tendency to train word-embeddings
  - $W_{bad} = W_{good}$ (Li et al., IJCAI-2018)
- Pre-store the domain-knowledge into a memory network
  - E.g., add a *Memory Network*
- Integrate DNN into original operation flow (instead of end-to-end)
  - E.g., replace statistical approach with DNN for the STC module in MWP

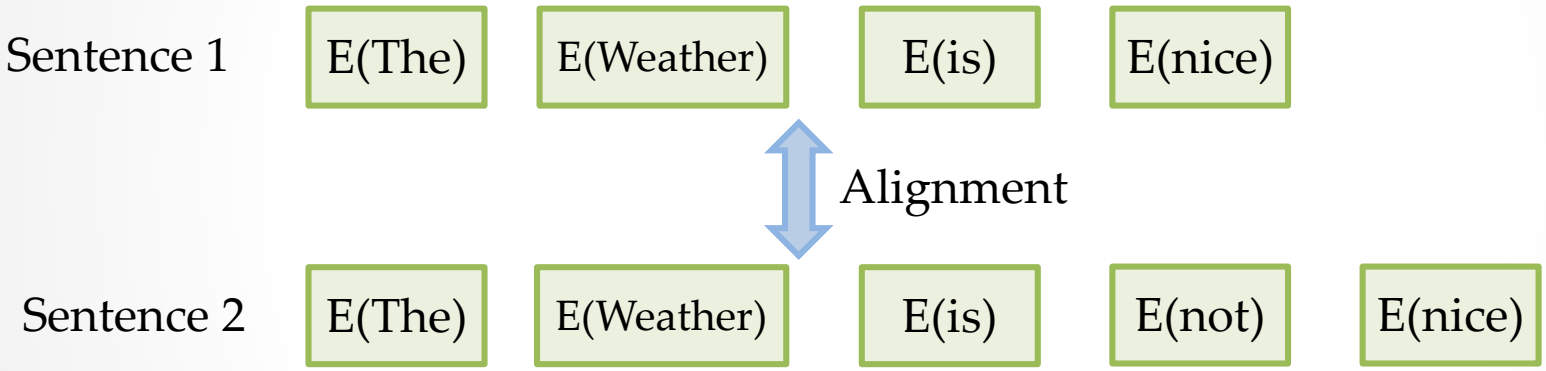
Observations (1/3)

- Previous DNN approaches for entailment judgment
  - Form two *sentence-embedding* vectors independently
  - sentence-embedding is formed by simply summing associated word-embeddings

- One major problem – *Information Loss!*
  - Overlapping feature space upon vector addition.
  - *Less discriminative power* for NN models that treat the sentence as a unit.
    - Scenario A: Distance between sentence-embeddings is *large*
    - Scenario B: Distance between sentence-embeddings is *small*

Alignment/Attention: only check words in H

Instead, Align sentential units (words, phrases, etc.):



S: *Bob is in his room, but because of the thunder and lightning outside, he cannot sleep.*
H1: *Bob is awake.* **H2:** *It is sunny outside.*

A Decomposable Attention Model for Natural Language Inference, Ankur P. Parikh et al., EMNLP-2016

Decomposable DNN (Parikh et al., 2016)

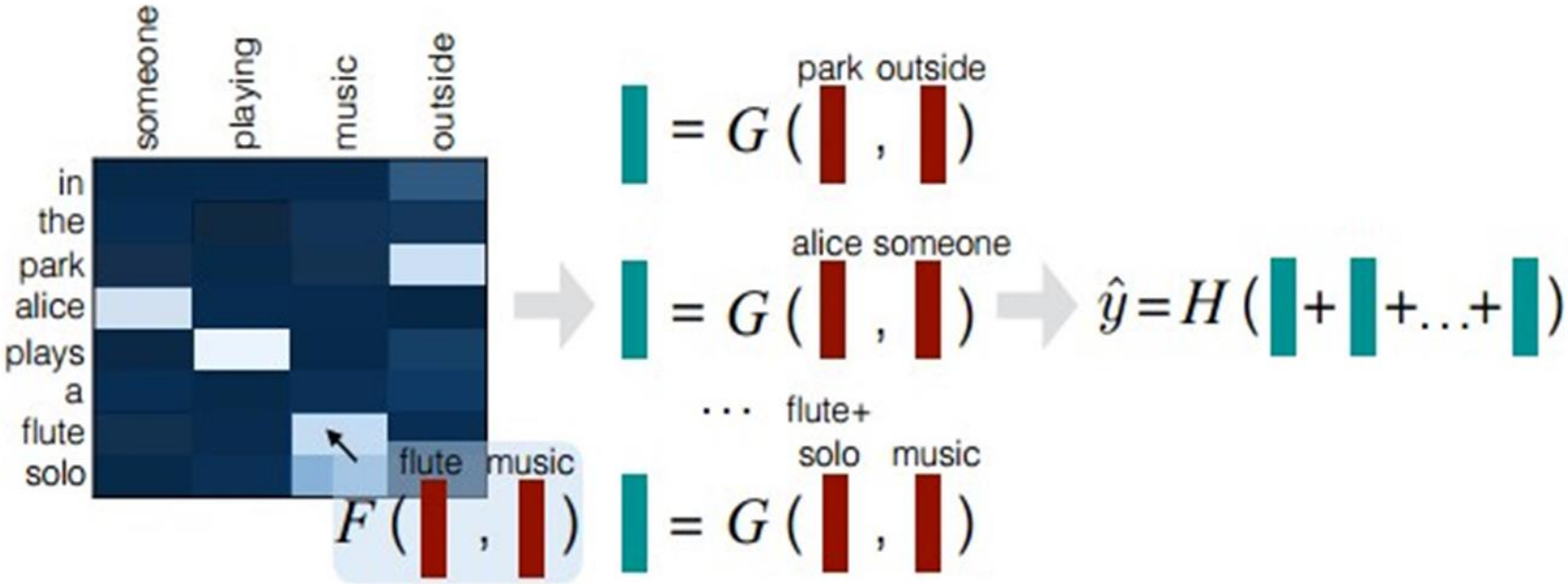


Figure 1: Pictoral overview of the approach, showing the *Attend* (left), *Compare* (center) and *Aggregate* (right) steps.

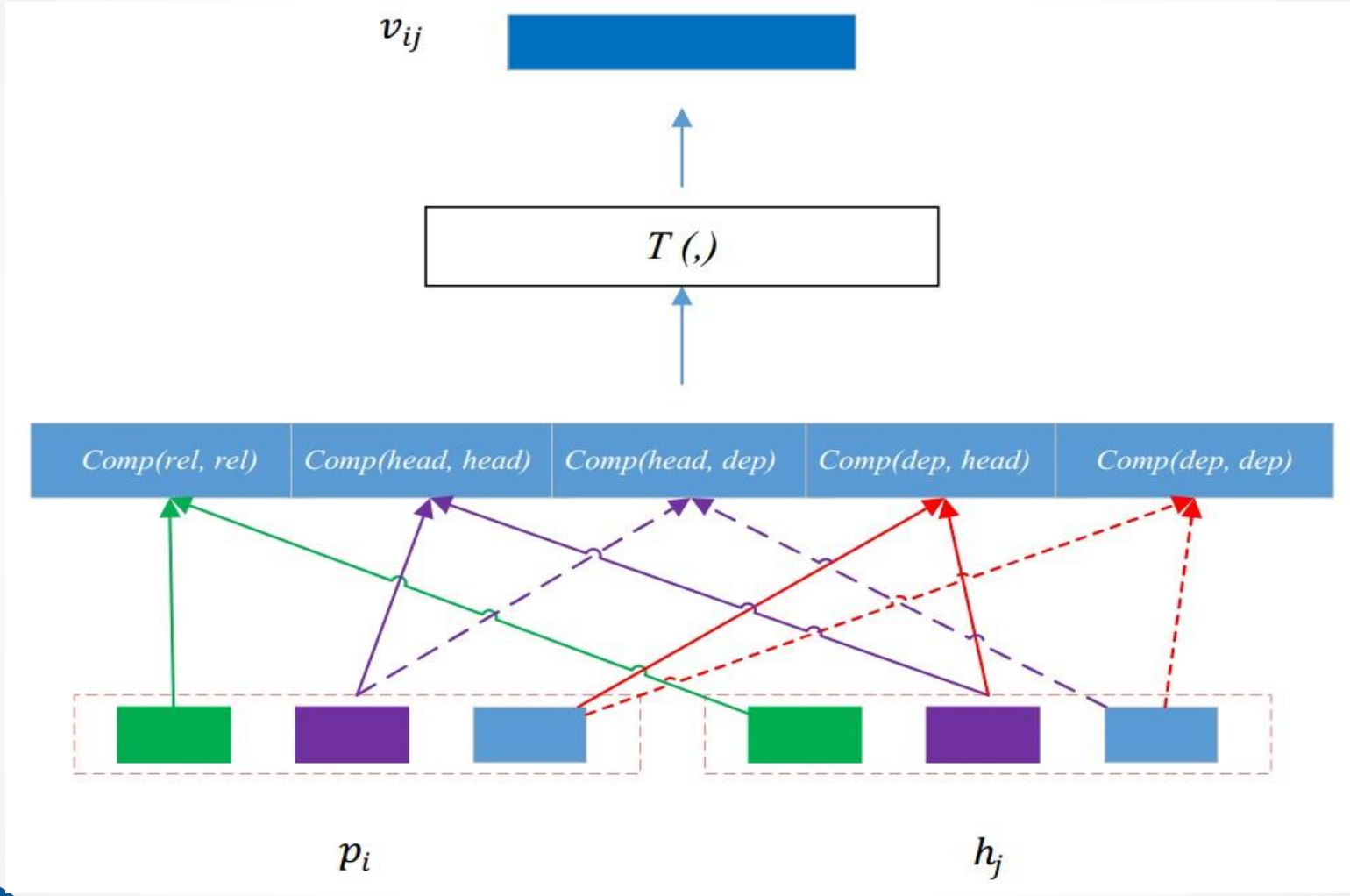
Word-Pair Dependency (1/3)

- **Premise**: An older man sits with his orange juice at a **small** table in a coffee **shop** while employees in bright colored shirts smile in the background.

- **Hypothesis**: An elderly man sitting in a **small** **shop**.

- Relation(Head, Dependent) (Du et al., 2018)

Word-Pair Dependency (2/3)



Integrate Domain-Concept into DNN (2/3)

- Adjust DNN architecture toward human reasoning style
  - E.g., add an alignment layer to DNN in entailment judgment
- Add preferences implied by domain knowledge during DNN training process
  - E.g., add human preferred tendency to train word-embeddings
  - $W_{king} - W_{queen} = W_{husband} - W_{wife}$ (Li et al., IJCAI-2018)
- Pre-store the domain-knowledge into a memory network
  - E.g., add a *Memory Network*
- Integrate DNN into original operation flow (instead of end-to-end)
  - E.g., replace statistical approach with DNN for the STC module in MWP

Adding Constraints (Li et al., IJCAI-18)

- Learning Word Vectors with Linear Constraints (19,544 Q)
  - $W_{king} - W_{queen} = W_{husband} - W_{wife}$

Models			Analogy		
Name	Dim	Size	SEM	SYN	Overall
SVD	100	1.0B	23.4	31.3	27.4
GloVe	100	1.0B	46.3	38.4	42.3
CBOW	100	1.0B	34.8	37.0	36.0
SG	100	1.0B	49.7	39.4	44.1
Additive(1%)	100	1.0B	74.2	66.3	70.2
Additive(2%)	100	1.0B	89.2	91.3	90.3
Additive(3%)	100	1.0B	100.0	99.7	99.9
SVD	200	1.0B	23.4	31.2	27.3
GloVe	200	1.0B	58.7	49.5	54.0
CBOW	200	1.0B	40.8	42.9	42.0
SG	200	1.0B	63.5	47.0	54.5
Additive(1%)	200	1.0B	74.2	66.3	70.2
Additive(2%)	200	1.0B	89.3	91.3	90.3
Additive(3%)	200	1.0B	100.0	99.8	99.9
SVD	200	4.2B	38.3	46.6	42.5
GloVe	200	4.2B	71.1	60.5	65.7
CBOW	200	4.2B	52.9	55.3	54.1
SG	200	4.2B	70.8	63.3	67.0
Additive(1%)	200	4.2B	74.2	66.3	70.2
Additive(2%)	200	4.2B	88.9	90.7	89.8
Additive(3%)	200	4.2B	100.0	99.9	99.9

Integrate Domain-Concept into DNN (2/3)

- Adjust DNN architecture toward human reasoning style
  - E.g., add an alignment layer to DNN in entailment judgment
- Add preferences implied by domain knowledge during DNN training process
  - E.g., add human preferred tendency to train word-embeddings
  - $W_{bad} = W_{good}$ (Li et al., IJCAI-2018)
- Pre-store the domain-knowledge into a memory network
  - E.g., add a *Memory Network*
- Integrate DNN into original operation flow (instead of end-to-end)
  - E.g., replace statistical approach with DNN for the STC module in MWP

Integrate Domain-Concept into DNN (3/3)

- Train the DNN model *iteratively/jointly* with domain knowledge representation
  - Enhance generalization and improve interpretability
  - E.g., ***iterative distillation*** (Hu et al., 2016a), which transfer the structured information of first-order logic rules into the weights of neural networks.
  - E.g., ***mutual distillation*** (Hu et al., 2016b), which transfers information between DNN and structured constraints for effective knowledge learning.

Who will win?

- History
  - Around 1990: NN vs. HMM in speech recognition
    - "Training Data Only" versus "Training Data + Human Knowledge"
  - Now: Researchers are adding more domain knowledge into vanilla DNN
    - NMT: Seq2Seq → Attention → Tree-based → Tree-Coverage measure
- History will repeat
  - "Domain Knowledge Guided DNN" vs. "Vanilla DNN"
  - To outperform other approaches in the same data-set, you should add domain knowledge
    - NLP/NLU needs more guidance from domain knowledge

Q&A

Keh-Yih Su

Institute of Information Science
Academia Sinica, Taipei

NLPCC, Oct. 13, 2019