Uphill Battles in NLP/AI: Knowledge About the World

Yejin Choi
Computer Science & Engineering

UNIVERSITY of WASHINGTON
What Begin to Work!

• Image description
• Video description
• Visual QA

- Very large dataset: MS CoCo, VQA, ImSitu…
Image Captioning (it works!?)

a man riding a surfboard on top of a wave

a man jumping on a swing at a tennis ball.
Image Captioning (or Not …?)

a child is being pulled by a small boy on a surfboard.

a young boy in a blue shirt is jumping.
MSR CoCo Dataset

- 120,000 images, 5 captions for each image
- 92 objects

- sports (10 categories):
  - frisbee, skies, snowboard, kite, sport balls, baseball bat, baseball gloves, skateboard, surf board, tennis racket (3561 images).

- street (5 categories)
  - traffic light (4330 images), fire hydrant (1797 images), stop sign (1803 images), parking meters (742 images), bench (5805 images)

- person (6 categories)
  - tie (3955 images), umbrella (4142 images)

Is this Data problem? Or Modeling problem?
Reasoning about the Event

Image captioning is an emblematic task, not the end goal

• What’s happening?
• How / why did this happen?
• What are the intent / goal of the participants?
• Sentiment: are they happy?
• Reaction: do we need to act on them (e.g., dispatching help)?
What Remain to be Hard

Goals: broad-coverage *grounding* and reasoning

- Image description
- Video description
- Visual QA
- ...


What Remain to be Hard

Goals: broad-coverage grounding and reasoning

- Image description
- Video description
- Visual QA
- ...

- Despite very large dataset: MS CoCo, VQA, ImSitu…

- Fundamental challenges with data and knowledge
Need: Knowledge about the World

- Propositional knowledge
  - knowledge of “that”
  - Encyclopedic knowledge:
    - E.g., Baltimore is a major city in Maryland with a long history as an important seaport. Fort McHenry, birthplace of the U.S. national anthem, “The Star-Spangled Banner,” sits at the mouth of Baltimore’s Inner Harbor….
  - Everyday functional knowledge (commonsense)
    - E.g., Bananas are usually yellow, elephants are larger than butterflies….

- Procedural knowledge
  - knowledge of “how”
  - e.g., how to ride a bicycle, how to brew beer

Our recent attempts on “reverse engineering” knowledge: EMNLP ’15, AAAI ’16 ICCV ’16, ACL ’16
Are Elephants Bigger than Butterflies?
Knowledge on Size Useful for

• Vision:
  – Prune out implausible detections

• Language:
  – The trophy would not fit in the brown suitcase because it was too **big**. What was too **big**?
    Answer 0: the trophy
    Answer 1: the suitcase
Support from Psychology Studies

- A familiar-size stroop effect. (Konkle, T., and Oliva, A. 2012.)
- Knowledge on size (Konkle & Oliva, 2011; Linsen, Leyssen, Sammartino, & Palmer, 2011).
Working around Reporting Bias

• Reporting bias: do not state the obvious
• Use both language and images!
• Elephants bigger than butterflies?
  ➔ Need multi-hop inference
In Sum, We Tried to Learn …

- Attempt to learn some relative physical knowledge from language and vision (at a small scale)
A horse is eating.
Is that horse standing or sitting?
a horse eating => a horse standing

• Reporting bias: do not state the obvious (Gordon and Benjamin Van Durme. 2013)
• Another case where language + vision can help!
Entailment $X \Rightarrow Y$

$T(Y)$

$T(X)$
Entailment $X \Rightarrow Y$

$\text{entail}(X \models Y) := Sim_{R2I}^{\rightarrow}(X, Y) - Sim_{R2I}^{\rightarrow}(Y, X)$

$Sim_{R2I}^{\rightarrow}(X, Y) =$ average asymmetric region-to-image similarity measure (Kim and Grauman 2010) using top K segmentation masks
In Sum, We Tried to Learn ...

1. Visual Entailment
2. Visual Paraphrasing
3. Semantic Similarity
Prototypical Event Knowledge
Learned Events:
- Dance
- Kiss
- Cut the cake
- Vows
- Exchange rings

Temporal Knowledge:
1. Ring time.
2. Exchanging our rings.
3. Rings and promises.
4. Sealed with a kiss.
5. Our first ever kiss.

Prototypical Captions:
- Dancing excitement.
- First dance.
- Ballroom dancing.
- Cake cutting.
- The cake was so solid.
- Reading our vows.
- Our vows.
Procedural Language and Knowledge

Kiddon et al. @ EMNLP 2015
Interpreting Natural Language Instructions as Action Diagrams

Smart devices and personal robots executing commands in natural language instructions not just one line command, but a sequence of commands

Step 1: interpret instructions as action diagrams
Blueberry Muffins

Ingredients
1 cup milk
1 egg
1/3 cup vegetable oil
2 cups all-purpose flour
2 teaspoons baking powder
1/2 cup white sugar
1/2 cup fresh blueberries

Procedure
1. Preheat oven to 400 degrees F. Line a 12-cup muffin tin with paper liners.
2. In a large bowl, stir together milk, egg, and oil. Add flour, baking powder, sugar, and blueberries; gently mix the batter with only a few strokes. Spoon batter into cups.

http://allrecipes.com/Recipe/Blueberry-Muffins-I/
Blueberry Muffins

Ingredients
1 cup milk
1 egg
1/3 cup vegetable oil
2 cups all-purpose flour
2 teaspoons baking powder
1/2 cup white sugar
1/2 cup fresh blueberries

Procedure
1. Preheat oven to 400 degrees F. Line a 12-cup muffin tin with paper liners.
2. In a large bowl, stir together milk, egg, and oil. Add flour, baking powder, sugar, and blueberries; gently mix the batter with only a few strokes. Spoon batter into cups.

http://allrecipes.com/Recipe/Blueberry-Muffins-1/
Blueberry Muffins

Ingredients
1 cup milk
1 egg
1/3 cup vegetable oil
2 cups all-purpose flour
2 teaspoons baking powder
1/2 cup white sugar
1/2 cup fresh blueberries

Procedure
1. Preheat oven to 400 degrees F (205 degrees C). Line a 12-cup muffin tin with paper liners.
2. In a large bowl, stir together milk, egg, and oil. Add flour, baking powder, sugar, and blueberries; gently mix the batter with only a few strokes. Spoon batter into cups.
Can do without annotated text?

- Yes, if with physical simulator
  - Branavan et al., 2009, Chen and Mooney, 2011, Bollini et al., 2013

- Can do without simulator, if with redundant data!
  - Our work (Kiddon et al., 2015)
  - 400 variations (!!@#!) on “macaroni and cheese” on allrecipes.com
Unsupervised Learning

- Chicken and Egg
  - Parsing (unstructured text $\rightarrow$ action graph) requires knowledge
  - Knowledge requires parsing

- Model:
  - Probabilistic Model

- Learning:
  - Expectation-Maximization
Probability model $P(C, R)$

- **Input**: A set of connections $C$ and a recipe $R$ segmented (Sec. 6) into its actions $\{e_1 = (v_1, a_1), \ldots, e_n = (v_n, a_n)\}$
- The joint probability of $C$ and $R$ is $P(C, R) = P(C)P(R|C)$, each defined below:

1. **Connections Prior (Sec. 3.1)**: $P(C) = \prod_i P(d_i|d_1, \ldots, d_{i-1})$
   Define $d_i$ as the list of connections with destination index $i$. Let $c_p = (o, i, j, k, t^{syn}_k, t^{sem}_k) \in d_i$. Then,
   - $P(d_i|d_1, \ldots, d_{i-1}) = P(vs(d_i)) \prod_{c_p \in d_i} P(\mathbf{1}(o \rightarrow s_{ij}^k)|vs(d_i), d_1, \ldots, d_{i-1}, c_1, \ldots, c_{p-1})$
     (a) $P(vs(d_i))$: multinomial verb signature model (Sec. 3.1.1)
     (b) $P(\mathbf{1}(o \rightarrow s_{ij}^k)|vs(d_i), d_1, \ldots, d_{i-1}, c_1, \ldots, c_{p-1})$: multinomial connection origin model, conditioned on the verb signature of $d_i$ and all previous connections (Sec. 3.1.2)

2. **Recipe Model (Sec. 3.2)**: $P(R|C) = \prod_i P(e_i|C, e_1, \ldots, e_{i-1})$
   For brevity, define $h_i = (e_1, \ldots, e_{i-1})$.
   - $P(e_i|C, h_i) = P(v_i|C, h_i)P(a_{ij}|C, h_i)$ (Sec. 3.2)
     Define argument $a_{ij}$ by its types and spans, $a_{ij} = (t^{syn}_{ij}, t^{sem}_{ij}, S_{ij})$.
     (a) $P(v_i|C, h_i) = P(v_i|g_i)$: multinomial verb distribution conditioned on verb signature (Sec. 3.2)
     (b) $P(a_{ij}|C, h_i) = P(t^{syn}_{ij}, t^{sem}_{ij}|C, h_i) \prod_{s_{ij} \in S_{ij}} P(s_{ij}^k|t^{syn}_{ij}, t^{sem}_{ij}, C, h_i)$
       i. $P(t^{syn}_{ij}, t^{sem}_{ij}|C, h_i)$: deterministic argument types model given connections (Sec. 3.2.1)
       ii. $P(s_{ij}^k|t^{syn}_{ij}, t^{sem}_{ij}, C, h_i)$: string span model computed by case (Sec. 3.2.2):
          A. $t^{sem}_{ij} = \text{food}$ and $\text{origin}(s_{ij}^k) \neq 0$: IBM Model 1 generating composites (Part-composite model)
          B. $t^{sem}_{ij} = \text{food}$ and $\text{origin}(s_{ij}^k) = 0$: naïve Bayes model generating raw food references (Raw food model)
          C. $t^{sem}_{ij} = \text{location}$: model for generating location referring expressions (Location model)

Figure 2: Summary of the joint probabilistic model $P(C, R)$ over connection set $C$ and recipe $R$. 
Sample Knowledge in the Model

- **Part-composite model**: how likely it is to generate a composite word given the incoming ingredients/raw materials
  - $P(\text{"dressing" | "oil" "vinegar"}) > P(\text{"batter" | "oil" "vinegar"})$

- **Location model**: how likely a location is given the action verb
  - $P(\text{"stove" | "bake"}) < P(\text{"oven" | "bake"})$
Current Thoughts

• Can overcome report bias by
  – Combining evidence from multiple modalities
  – Multi-step inference

• However, image / video processing is extremely expensive

• Not all physical knowledge is visual or perceptually attainable

• Language-only approaches

• Crowdsourcing might help

Our recent attempts on “reverse engineering” knowledge:
EMNLP ’15, AAAI ’16 ICCV ’16, ACL ’16
Thanks!