ASER: A Large-scale Eventuality Knowledge Graph

Yangqiu Song
Department of CSE, HKUST, Hong Kong
Summer 2019

Contributors to related works: Hongming Zhang, Xin Liu, Haojie Pan, Cane Leung, Hantian Ding
Outline

• Motivation: NLP and commonsense knowledge

• Consideration: selectional preference

• New proposal: large-scale and higher-order selectional preference

• Evaluation and Applications
Understanding human’s language requires complex knowledge

• "Crucial to comprehension is the knowledge that the reader brings to the text. The construction of meaning depends on the reader's knowledge of the language, the structure of texts, a knowledge of the subject of the reading, and a broad-based background or world knowledge." (Day and Bamford, 1998)
  • Contexts and knowledge contributes to the meanings
Knowledge is Crucial to NLU

• Linguistic knowledge:
  • “The task is part-of-speech (POS) tagging with limited or no training data. Suppose we know that each sentence should have at least one verb and at least one noun, and would like our model to capture this constraint on the unlabeled sentences.” (Example from Posterior Regularization, Ganchev et al., 2010, JMLR)

• Contextual/background knowledge: conversational implicature

Example taking from VisDial (Das et al., 2017)
When you are asking Siri...

Interacting with human involves a lot of commonsense knowledge:
• Space
• Time
• Location
• State
• Causality
• Color
• Shape
• Physical interaction
• Theory of mind
• Human interactions

Judy Kegl, *The boundary between word knowledge and world knowledge*, TINLAP3, 1987
Ernie Davis, *Building AIs with Common Sense*, Princeton Chapter of the ACM, May 16, 2019
How to define commonsense knowledge? (Liu & Singh, 2004)

• “While to the average person the term ‘commonsense’ is regarded as synonymous with ‘good judgement’, ”
• “the AI community it is used in a technical sense to refer to the millions of basic facts and understandings possessed by most people.”
• “Such knowledge is typically omitted from social communications”, e.g.,
  • If you forget someone’s birthday, they may be unhappy with you.
How to collect commonsense knowledge?

• ConceptNet5 (Speer and Havasi, 2012)
  • Core is from Open Mind Common Sense (OMCS) (Liu & Singh, 2004)

• Essentially a crowdsourcing based approach + text mining
• Knowledge in ConceptNet
  • Things
  • Spatial
  • Location
  • Events
  • Causal
  • Affective
  • Functional
  • Agents
## Comparison

<table>
<thead>
<tr>
<th>Database content</th>
<th>Resource</th>
<th>Capabilities</th>
<th>Scales</th>
</tr>
</thead>
<tbody>
<tr>
<td>ConceptNet (2002-now)</td>
<td>Commonsense</td>
<td>OMCS (from the public) (automatic)</td>
<td>Contextual inference</td>
</tr>
<tr>
<td>Cyc (1984-now)</td>
<td>Commonsense</td>
<td>Expert (manual)</td>
<td>Formalized logical reasoning</td>
</tr>
</tbody>
</table>

Slides credit: Haixun Wang
The Scale

• “A founder of AI, Marvin Minsky, once estimated that ‘...commonsense is knowing maybe 30 or 60 million things about the world and having them represented so that when something happens, you can make analogies with others.’” (Liu & Singh, 2004)
What contribute to ConceptNet5.5 (21 million edges and over 8 million nodes)?

- Facts acquired from **Open Mind Common Sense** (OMCS) (Singh 2002) and sister projects in other languages (Anacleto et al. 2006)
- Information extracted from parsing **Wiktionary**, in multiple languages, with a custom parser (“Wikiparsec”)
- “Games with a purpose” designed to collect common knowledge (von Ahn, Kedia, and Blum 2006) (Nakahara and Yamada 2011) (Kuo et al. 2009)
- Open **Multilingual WordNet** (Bond and Foster 2013), a linked-data representation of WordNet (Miller et al. 1998) and its parallel projects in multiple languages
- JMDict (Breen 2004), a **Japanese-multilingual dictionary**
- OpenCyc, a **hierarchy of hypernyms** provided by Cyc (Lenat and Guha 1989), a system that represents commonsense knowledge in predicate logic
- A subset of DBPedia (Auer et al. 2007), a network of facts extracted from **Wikipedia infoboxes**

Most of them are entity-centric knowledge, there are only 74,989 relations among 116,097 edges about events.

Speer, Chin, and Havasi, ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. AAAI 2017.
Nowadays,

• Many large-scale knowledge graphs about entities and their attributes (property-of) and relations (thousands of different predicates) have been developed
  • Millions of entities and concepts
  • Billions of relationships

Google Knowledge Graph (2012)
570 million entities and 18 billion facts
However,

• Semantic meaning in our language can be described as ‘a finite set of mental primitives and a finite set of principles of mental combination (Jackendoff, 1990).’

• The primitive units of semantic meanings include
  • Thing (or Object),
  • Activity,
  • State,
  • Event,
  • Place,
  • Path,
  • Property,
  • Amount,
  • etc.

Outline

• Motivation: NLP and commonsense knowledge

• Consideration: selectional preference

• New proposal: large-scale and higher-order selectional preference

• Applications
Semantic Primitive Units

• Entities or concepts can be nouns or noun phrases
  • Concepts in Probase (2012):
    • Company,
    • IT company,
    • big company,
    • big IT company,
    • ...
  • Hierarchy is partially based on head+modifier composition

• Let’s think about verbs and verb phrases
  • How should we define semantic primitive unit for verbs?

Semantic Primitive Units Related to Verbs

• Too short?
  • v-o (verb + object): too general
    • have book
    • love reading
  • s-v-o (subject + verb + object): many overlaps with entity-centric KGs
    • Trump was born in 1946
    • Apple released iphone 8

• Too long?
  • Difficult to obtain frequent facts to reflect commonness of commonsense knowledge

Semantic Role Labeling (SRL) recovers the latent predicate argument structure of a sentence

https://demo.allennlp.org/semantic-role-labeling
Semantic Primitive Units

• Too fine grained, e.g., Event Extraction in ACE 2005?

After **Sept. 11, 2001**, Indonesia was quick to sign onto **U.S. President George W. Bush's**

Time-Starting
global **war** on terror.

**Trigger**

**Attack**

**Event Type: Conflict.Attack**

**Attacker**

State-of-the-art overall F1: around 40% (Ji, & Huang, 2013)

In **Baghdad**, a **cameraman** died when an **American tank** fired on the **Palestine Hotel**.

**Die.Place**

**Die.Victim**

**Trigger1**

**Die**

**Attack.Place**

**Attack.Victim**

**Attack.Instrument**

**Trigger2**

**Attack**

**Attack.Target**

**Event Type: Life.Die**

**Event Type: Conflict.Attack**

Qi Li, Heng Ji, Liang Huang: Joint Event Extraction via Structured Prediction with Global Features. ACL (1) 2013: 73-82
Commonsense Knowledge Construction

• The principle: a middle way of building primitive semantic units
  • Not too long
  • Not too short
  • Could be general
  • Better to be specific and semantically meaningful and complete

• Any linguistic foundation?
“Linguistic description – grammar = semantics”
The lower bound of a semantic theory (Katz and Fodor, 1963)

- Disambiguation needs both “the speaker's knowledge of his language and his knowledge” (Katz and Fodor, 1963)
  - The **bill** is large.
  - Some document demanding a sum of money to discharge a debt exceeds in size most such documents
  - The beak of a certain bird exceeds in bulk those of most similar birds
  - Syntactically unambiguous
  - Compare semantic meanings by fixing grammar

Selectional Preference (SP)

- The need of language inference based on ‘partial information’ (Wilks, 1975)
  - The soldiers fired at the woman, and we saw several of them fall.
  - The needed partial information: hurt things tending to fall down
    - “not invariably true”
    - “tend to be of a very high degree of generality indeed”
- Selectional preference (Resnik, 1993)
  - A relaxation of selectional restrictions (Katz and Fodor, 1963) and as syntactic features (Chomsky, 1965)
  - Applied to isA hierarchy in WordNet and verb-object relations

A Test of Commonsense Reasoning

- Proposed by Hector Levesque at U of Toronto
- An example taking from Winograd Schema Challenge
  
  • (A) The fish ate the worm. It was hungry.
  • (B) The fish ate the worm. It was tasty.

- On the surface, they simply require the resolution of anaphora
  - But Levesque argues that for Winograd Schemas, the task requires the use of knowledge and commonsense reasoning

http://commonsensereasoning.org/winograd.html
https://en.wikipedia.org/wiki/Winograd_Schema_Challenge
Why is it a challenge?

- Must also be carefully written not to betray their answers by selectional restrictions or statistical information about the words in the sentence

- Designed to be an improvement on the Turing test

The **soldiers** fired at the **woman**, and we saw several of **them** fall.

(A) The **fish** ate the worm. It was hungry.

(B) The fish ate the **worm**. It was tasty.
A Brief History of Datasets and Development

The first large dataset. Rahman and Ng: EMNLP-CoNLL 2012

Stanford: 55.19%
Their system: 73.05%

"Strictly speaking, we are addressing a relaxed version of the Challenge: while Levesque focuses solely on definite pronouns whose resolution requires background knowledge not expressed in the words of a sentence, we do not impose such a condition on a sentence."

Levesque. AAAI Spring Symposium 2011

Davis et al. "A Collection of Winograd Schemas" 2014

The first round of the challenge was a collection of 60 Pronoun Disambiguation Problems (PDPs). The highest score achieved was 58% correct, by Quan Liu, from University of Science and Technology, China.

Recent results

<table>
<thead>
<tr>
<th>Author/year</th>
<th>System</th>
<th>Fine-tuned</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emami et al. (2018)</td>
<td>Knowledge Hunter</td>
<td>No</td>
<td>54.58%</td>
</tr>
<tr>
<td>Trieu H. Trinh and Quoc V. Le (2018)</td>
<td>Language models (single)</td>
<td>No</td>
<td>54.58%</td>
</tr>
<tr>
<td></td>
<td>Language models (Ensemble)</td>
<td>No</td>
<td>63.74%</td>
</tr>
<tr>
<td>Alec Radford et al. (2019)</td>
<td>GPT-2</td>
<td>No details</td>
<td>70.70%</td>
</tr>
<tr>
<td>Ruan et al. (2019)</td>
<td>BERT-large + dependency</td>
<td>Rahman and Ng 2012 dataset</td>
<td>71.10%</td>
</tr>
<tr>
<td>Kocijan et al. (2019)</td>
<td>BERT-large</td>
<td>No</td>
<td>60.10%</td>
</tr>
<tr>
<td></td>
<td>GPT</td>
<td>No</td>
<td>55.30%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>72.20%</td>
</tr>
</tbody>
</table>

Human’s performance: 95% (Nangia and Bowman, 2018)
SP-10K: A Large-scale Evaluation Set

• Traditional evaluation
  • Small sets of one-hop direct dependency relations
    • McRae et al., 1998: 821 pairs of nsubj and dobj relations
    • Keller and Lapata, 2003: 540 pairs of dobj, noun-noun, and amod relations
    • Padó et al., 2006: 207 pairs of nsubj, dobj, and amod relations
    • Wang et al, 2018: 3062 (subject, verb, dobject) triplets
  • Pseudo-disambiguation (Ritter et al., 2010; de Cruys, 2014): corpus driven, no human annotation

• Ours:
  • 10K pairs of five relations, including two 2-hop relations
# Examples in SP-10K

<table>
<thead>
<tr>
<th>dobj</th>
<th>Plausibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>(eat, meal)</td>
<td>10.00</td>
</tr>
<tr>
<td>(close, door)</td>
<td>8.50</td>
</tr>
<tr>
<td>(touch, food)</td>
<td>5.50</td>
</tr>
<tr>
<td>(hate, investment)</td>
<td>4.00</td>
</tr>
<tr>
<td>(eat, mail)</td>
<td>0.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>nsubj</th>
<th>Plausibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>(singer, sing)</td>
<td>10.00</td>
</tr>
<tr>
<td>(law, permit)</td>
<td>7.78</td>
</tr>
<tr>
<td>(women, pray)</td>
<td>5.83</td>
</tr>
<tr>
<td>(victim, contain)</td>
<td>2.22</td>
</tr>
<tr>
<td>(textbook, eat)</td>
<td>0.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>amod</th>
<th>Plausibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>(fresh, air)</td>
<td>9.77</td>
</tr>
<tr>
<td>(new, method)</td>
<td>8.89</td>
</tr>
<tr>
<td>(medium, number)</td>
<td>4.09</td>
</tr>
<tr>
<td>(immediate, food)</td>
<td>2.05</td>
</tr>
<tr>
<td>(secret, wind)</td>
<td>0.75</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>dobj_amod</th>
<th>Plausibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>(lift, heavy object)</td>
<td>9.17</td>
</tr>
<tr>
<td>(design, new object)</td>
<td>8.00</td>
</tr>
<tr>
<td>(attack, small object)</td>
<td>5.23</td>
</tr>
<tr>
<td>(inform, weird object)</td>
<td>3.64</td>
</tr>
<tr>
<td>(earn, rubber object)</td>
<td>0.63</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>nsubj_amod</th>
<th>Plausibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>(evil subject, attack)</td>
<td>9.00</td>
</tr>
<tr>
<td>(recent subject, demonstrate)</td>
<td>6.00</td>
</tr>
<tr>
<td>(random subject, bear)</td>
<td>4.00</td>
</tr>
<tr>
<td>(happy subject, steal)</td>
<td>2.25</td>
</tr>
<tr>
<td>(sunny subject, make)</td>
<td>0.56</td>
</tr>
</tbody>
</table>
Correlations with OMCS

(sing, song) (dobj, 9.25)
(song, UsedFor, sing)

(phone, ring) (nsubj, 8.75)
(phone, CapableOf, ring)

(cold, water) (amod, 8.86)
(water, HasProperty, cold)

(create, new) (dobj_amod, 8.25)
(create idea, UsedFor, invent new things)

(hungry, eat) (nsubj_amod, 10.00)
(eat, MotivatedByGoal, are hungry)
Performance on Winograd Schema

• 72 out of 273 questions satisfying nsubj_amod and dobj_amod relations
  • Jim yelled at Kevin because he was so upset.
  • We compare the scores
    • (yell, upset object) following nsubj_amod
    • (upset object, yell) following dobj_amod

• Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Correct</th>
<th>Wrong</th>
<th>NA</th>
<th>Accuracy (predicted)</th>
<th>Accuracy (overall)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stanford</td>
<td>33</td>
<td>35</td>
<td>4</td>
<td>48.5%</td>
<td>48.6%</td>
</tr>
<tr>
<td>End2end (Lee et al., 2018)</td>
<td>36</td>
<td>36</td>
<td>0</td>
<td>50.0%</td>
<td>50.0%</td>
</tr>
<tr>
<td>PP* (Resnik, 1997)</td>
<td>36</td>
<td>19</td>
<td>17</td>
<td>65.5%</td>
<td>61.8%</td>
</tr>
<tr>
<td>SP-10K</td>
<td>13</td>
<td>0</td>
<td>56</td>
<td>100%</td>
<td>59.0%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>dobj_amod</th>
<th>Plausibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>(lift, heavy object)</td>
<td>9.17</td>
</tr>
<tr>
<td>(design, new object)</td>
<td>8.00</td>
</tr>
<tr>
<td>(attack, small object)</td>
<td>5.23</td>
</tr>
<tr>
<td>(inform, weird object)</td>
<td>3.64</td>
</tr>
<tr>
<td>(earn, rubber object)</td>
<td>0.63</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>nsubj_amod</th>
<th>Plausibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>(evil subject, attack)</td>
<td>9.00</td>
</tr>
<tr>
<td>(recent subject, demonstrate)</td>
<td>6.00</td>
</tr>
<tr>
<td>(random subject, bear)</td>
<td>4.00</td>
</tr>
<tr>
<td>(happy subject, steal)</td>
<td>2.25</td>
</tr>
<tr>
<td>(sunny subject, make)</td>
<td>0.56</td>
</tr>
</tbody>
</table>

*PP: posterior probability for SP acquisition using Wikipedia data
Outline

• Motivation: NLP and commonsense knowledge

• Consideration: selectional preference

• New proposal: large-scale and higher-order selectional preference

• Applications
Higher-order Selectional Preference

• The need of language inference based on ‘partial information’ (Wilks, 1975)
  • The soldiers fired at the woman, and we saw several of them fall.
  • The needed partial information: hurt things tending to fall down
  • Many ways to represent it, e.g.,

  (hurt, X) connection (X, fall)

• How to scale up the knowledge acquisition and inference?
ATOMIC

• Crowdsourcing 9 Types of IF-THEN relations

• All entity information has been removed to reduce ambiguity

KnowlyWood

• Perform information extraction from free text
  • Mostly movie scripts and novel books

• Four relations: previous, next, parent, similarity

• No subject information
  • Only verb+object

Scales of Verb Related Knowledge Graphs

FrameNet (Baker et al., 1998)
ACE (Aguilar et al., 2014)
PropBank (Palmer et al., 2005)
TimeBank (Pustejovsky et al., 2003)
ConceptNet (Liu & Singh, 2004)
Event2Mind (Smith et al., 2018)
Propora (Dalvi et al., 2018)
ATOMIC (Sap et al., 2018)
Knowlywood (Tandon et al., 2015)
ASER (core)
ASER (full)

#Eventualities    #Relations
ASER

Activities, States, Events, and their Relations

Mourelatos’ taxonomy (1978)

- State: The air smells of jasmine.
- Process: It’s snowing.
- Development: The sun went down.
- Punctual occurrence: The cable snapped. He blinked. The pebble hit the water.

Bach’s taxonomy (1986)

- Static states: be in New York, love (one’s cat);
- Dynamic states: sit, stand, drunk, present, sick;
- Processes: walk, push a cart, sleep;
- Protracted events: build (a cabin), eat a sandwich, polish a shoe, walk to Boston;
- Culminations: take off; arrive, leave, depart;
- Happenings: blink, flash, knock, kick, hit, pat, wink;

Many other and subtle definitions

• According to Alexander Mourelatos (1978),
  • “Event, can be sharply differentiated by ... the contrast between perfective and imperfective aspect in verbs corresponds to the count/mass distinction in the domain of nouns.”
    • “Cardinal count” adverbials versus frequency
    • adverbials occurrence versus associated occasion
  • “Mary capsized the boat” is an event predication because (a) it is equivalent to “There was at least one capsizing of the boat by Mary,” or (b) because it admits cardinal count adverbials, e.g., “at least once,” “twice,” “three times.”

• Learning based classification
  • English: state vs. non-state; ~93.9% accuracy; culminated/nonculminated ~74.0% (Siegel and McKeown, 2000)
  • Chinese: state, activity, change; ~73.6% accuracy (Liu et al., 2018)
Our Approach

- Use verb-centric patterns from dependency parsing
  - Principle #1: For comparing semantics by fixing syntax (Katz and Fodor, 1963)
- Maintain a set of key tags and a set of auxiliary tags
  - Principle #2: For obtaining frequent ‘partial information’ (Wilks, 1975)

A hybrid graph of
- Each eventuality is a hyper-edge of words
- Heterogeneous edges among eventualities
Eventualities

- Using patterns to collect partial information
  - Six relations are also kept but treated as auxiliary edges
    - advmod,
    - amod,
    - nummod,
    - aux,
    - compound,
    - neg

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Code</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>n1-nsubj-v1</td>
<td>s-v</td>
<td>'The dog barks'</td>
</tr>
<tr>
<td>n1-nsubj-v1-dobj-n2</td>
<td>s-v-o</td>
<td>'I love you'</td>
</tr>
<tr>
<td>n1-nsubj-v1-xcomp-a</td>
<td>s-v</td>
<td>'He felt ill'</td>
</tr>
<tr>
<td>n1-nsubj-(v1-iobj-n2)-dobj-n3</td>
<td>s-v-o-o</td>
<td>'You give me the book'</td>
</tr>
<tr>
<td>n1-nsubj-a1-cop-be</td>
<td>s-be-a</td>
<td>'The dog is cute'</td>
</tr>
<tr>
<td>n1-nsubj-v1-xcomp-a1-cop-be</td>
<td>s-v-be-a</td>
<td>'I want to be slim'</td>
</tr>
<tr>
<td>n1-nsubj-v1-xcomp-n2-cop-be</td>
<td>s-v-be-o</td>
<td>'I want to be a hero'</td>
</tr>
<tr>
<td>n1-nsubj-v1-xcomp-v2-dobj-n2</td>
<td>s-v-v-o</td>
<td>'I want to eat the apple'</td>
</tr>
<tr>
<td>n1-nsubj-v1-xcomp-v2</td>
<td>s-v-v</td>
<td>'I want to go'</td>
</tr>
<tr>
<td>(n1-nsubj-a1-cop-be)-nmod-n2-case-p1</td>
<td>s-be-a-p-o</td>
<td>'It' cheap for the quality'</td>
</tr>
<tr>
<td>n1-nsubj-v1-nmod-n2-case-p1</td>
<td>s-v-p-o</td>
<td>'He walks into the room'</td>
</tr>
<tr>
<td>(n1-nsubj-v1-dobj-n2)-nmod-n3-case-p1</td>
<td>s-v-o-p-o</td>
<td>'He plays football with me'</td>
</tr>
<tr>
<td>n1-nsubjpass-v1</td>
<td>spass-v</td>
<td>'The bill is paid'</td>
</tr>
<tr>
<td>n1-nsubjpass-v1-nmod-n2-case-p1</td>
<td>spass-v-p-o</td>
<td>'The bill is paid by me'</td>
</tr>
</tbody>
</table>
Extraction Results

• Extract examples from 11-billion tokens from Yelp, NYT, Wiki, Reddit, Subtitles, E-books

• Evaluate about 200 examples in each pattern using Amazon Turk
Distribution

- Frequency characterizes selectional preference, e.g.,
- ‘The dog is chasing the cat, it barks loudly’
  - ‘dog barks’ appears 12,247
  - ‘cat barks’ never appears
Eventuality Relations: Pattern Matching + Bootstrapping

- Seeds from Penn Discourse Treebank (PDTB) (Prasad et al., 2007)
- 14 relations taking from CoNLL shared task
  - More frequent relations
- Less ambiguous connectives
  - ‘so that’ 31 times only in ‘Result’ relations
- Some are ambiguous
  - ‘while’: Conjunction 39 times, Contrast 111 times, Expectation 79 times, and Concession 85 times

<table>
<thead>
<tr>
<th>Relation Type</th>
<th>Seed Patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precedence</td>
<td>E1 before E2; E1, then E2; E1 till E2; E1 until E2</td>
</tr>
<tr>
<td>Succession</td>
<td>E1 after E2; E1 once E2</td>
</tr>
<tr>
<td>Synchronous</td>
<td>E1, meanwhile E2; E1 meantime E2; E1, at the same time E2</td>
</tr>
<tr>
<td>Reason</td>
<td>E1, because E2</td>
</tr>
<tr>
<td>Result</td>
<td>E1, so E2; E1, thus E2; E1, therefore E2; E1, so that E2</td>
</tr>
<tr>
<td>Condition</td>
<td>E1, if E2; E1, as long as E2</td>
</tr>
<tr>
<td>Contrast</td>
<td>E1, but E2; E1, however E2; E1, by contrast E2; E1, in contrast E2; E1, on the contrary, E2</td>
</tr>
<tr>
<td>Concession</td>
<td>E1, although E2</td>
</tr>
<tr>
<td>Conjunction</td>
<td>E1 and E2; E1, also E2</td>
</tr>
<tr>
<td>Instantiation</td>
<td>E1, for example E2; E1, for instance E2</td>
</tr>
<tr>
<td>Restatement</td>
<td>E1, in other words E2</td>
</tr>
<tr>
<td>Alternative</td>
<td>E1 or E2; E1, unless E2; E1, as an alternative E2; E1, otherwise E2</td>
</tr>
<tr>
<td>ChosenAlternative</td>
<td>E1, E2 instead</td>
</tr>
<tr>
<td>Exception</td>
<td>E1, except E2</td>
</tr>
</tbody>
</table>

Eventuality Relations: Pattern matching + Bootstrapping

- Bootstrapping: incrementally self-supervised learning
- For each instance $x = (E1; E2; \text{sentence})$
  - Use three bidirectional LSTMs
- Reduce the confident rate by iterations to reduce error propagation
Extraction Results

• Left: number of relations and overall accuracy
• Right: accuracy of each relations for the last iteration
• Each point is annotated with 200 examples by Amazon Turk
An Example of Inference over ASER

- We can support both eventuality-based and relation based inference
- We also do higher-order relation inference

<table>
<thead>
<tr>
<th>Eventuality Retrieval</th>
<th>Relation Retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>One-hop</strong></td>
<td></td>
</tr>
<tr>
<td>P('I have lunch'</td>
<td>‘I am hungry’, <em>Result</em>) = 1</td>
</tr>
<tr>
<td>P('I go’</td>
<td>‘I make a call’, <em>Precedence</em>) = 0.6</td>
</tr>
<tr>
<td>P('I depart away’</td>
<td>‘I make a call’, <em>Precedence</em>) = 0.4</td>
</tr>
<tr>
<td><strong>Two-hop</strong></td>
<td></td>
</tr>
<tr>
<td>P('I rest one a bench’</td>
<td>‘I sleep’, <em>Reason</em>, <em>Result</em>) = 1</td>
</tr>
<tr>
<td>P('I am hungry’</td>
<td>‘I sleep’, <em>Reason</em>, <em>Conjunction</em>) = 0.91</td>
</tr>
<tr>
<td>P('I rest one a bench’</td>
<td>‘I sleep’, <em>Reason</em>, <em>Conjunction</em>) = 0.09</td>
</tr>
</tbody>
</table>
Outline

- Motivation: NLP and commonsense knowledge
- Consideration: selectional preference
- New proposal: large-scale and higher-order selectional preference
- Applications
Inference for Winograd Schema Challenge

**Question**

97. *The fish* ate *the worm*. It was hungry.

98. *The fish* ate *the worm*. It was tasty.

**Extracted Eventualities**

- The fish: (‘X ate Y’, ‘X was hungry’)
- The worm: (‘X ate Y’, ‘Y was hungry’)
- The fish: (‘X ate Y’, ‘X was tasty’)
- The worm: (‘X ate Y’, ‘Y was tasty’)

**ASER Knowledge**

- ASER(‘X ate Y’, ‘X was hungry’) = 18
- ASER(‘X ate Y’, ‘Y was hungry’) = 1
- ASER(‘X ate Y’, ‘X was tasty’) = 0
- ASER(‘X ate Y’, ‘Y was tasty’) = 7

**Prediction**

- The fish
- The worm
Results

- We selected a subset of 165 questions
  - The sentence does not have a subordinate clause
  - The targeting pronoun is covered by a pattern we used

<table>
<thead>
<tr>
<th>Methods</th>
<th>Correct</th>
<th>Wrong</th>
<th>NA</th>
<th>Predicted</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Guess</td>
<td>83</td>
<td>82</td>
<td>0</td>
<td>50.30%</td>
<td>50.30%</td>
</tr>
<tr>
<td>Deterministic (Raghunathan et al., 2010)</td>
<td>75</td>
<td>71</td>
<td>19</td>
<td>51.40%</td>
<td>51.20%</td>
</tr>
<tr>
<td>Statistical (Clark &amp; Manning, 2015)</td>
<td>75</td>
<td>78</td>
<td>12</td>
<td>49.00%</td>
<td>49.10%</td>
</tr>
<tr>
<td>Deep-RL (Clark &amp; Manning, 2016)</td>
<td>80</td>
<td>76</td>
<td>9</td>
<td>51.30%</td>
<td>51.20%</td>
</tr>
<tr>
<td>End2end (Lee et al., 2018)</td>
<td>79</td>
<td>84</td>
<td>2</td>
<td>48.50%</td>
<td>48.50%</td>
</tr>
<tr>
<td>Knowledge Hunting (Emami et al., 2018)</td>
<td>94</td>
<td>71</td>
<td>0</td>
<td>56.90%</td>
<td>56.90%</td>
</tr>
<tr>
<td>LM (single) (Trinh &amp; Le, 2018)</td>
<td>90</td>
<td>75</td>
<td>0</td>
<td>54.50%</td>
<td>54.50%</td>
</tr>
<tr>
<td>SP (human) (Zhang et al., 2019)</td>
<td>15</td>
<td>0</td>
<td>150</td>
<td>100%</td>
<td>54.50%</td>
</tr>
<tr>
<td>SP (PP) (Zhang et al., 2019)</td>
<td>50</td>
<td>26</td>
<td>89</td>
<td>65.80%</td>
<td>57.30%</td>
</tr>
<tr>
<td>ASER</td>
<td>63</td>
<td>27</td>
<td>75</td>
<td>70.00%</td>
<td>60.90%</td>
</tr>
</tbody>
</table>
## Dialogue Generation

**• DailyDialog dataset (Li et al., 2017)**

<table>
<thead>
<tr>
<th>Post</th>
<th>I should eat some food.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Response</strong></td>
<td>Yeah, you must be hungry. Do you like to eat some beaf?</td>
</tr>
<tr>
<td><strong>ConceptNet</strong></td>
<td><code>eat food', MotivatedByGoal, </code>you are hungry' <code>eat food', HasPrerequisite, </code>open your mouth'</td>
</tr>
<tr>
<td><strong>KnowlyWood</strong></td>
<td>(eat,food), next, (keep, eating) (eat,food), next, (enjoy, taste) (eat,food), next, (stick, wasp) ...</td>
</tr>
<tr>
<td><strong>ASER</strong></td>
<td>i eat food [s-v-o], Conjunction, beef is good [s-be-a] i eat food [s-v-o], Condition, i am hungry [s-be-a] i eat food [s-v-o], Concession, i take picture [s-v-o] ...</td>
</tr>
</tbody>
</table>
Coverage

- We select all pairs that at least one KG can cover
  - 30,145 of 49,188 conversation pairs are selected

<table>
<thead>
<tr>
<th>KG</th>
<th># Covered Pairs</th>
<th>Coverage Rate</th>
<th># Unique matched events</th>
</tr>
</thead>
<tbody>
<tr>
<td>ConceptNet</td>
<td>7,246</td>
<td>24.04%</td>
<td>1,195</td>
</tr>
<tr>
<td>Knowlywood</td>
<td>17,183</td>
<td>57.00%</td>
<td>30,036</td>
</tr>
<tr>
<td>ASER</td>
<td>20,494</td>
<td>67.98%</td>
<td>9,511</td>
</tr>
</tbody>
</table>
Dialogue Generation Model

• A typical seq2seq model with memories
Results

• Metric: BLEU score of generated responses

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU-1</th>
<th>BLEU-2</th>
<th>BLEU-3</th>
<th>BLEU-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>30.16</td>
<td>5.75</td>
<td>2.28</td>
<td>0.98</td>
</tr>
<tr>
<td>+ConceptNet</td>
<td>30.89</td>
<td>6.14</td>
<td>2.60</td>
<td>1.21</td>
</tr>
<tr>
<td>+KnowlyWood</td>
<td>30.72</td>
<td>6.26</td>
<td>2.68</td>
<td>1.29</td>
</tr>
<tr>
<td>+ASER</td>
<td><strong>32.10</strong></td>
<td><strong>7.14</strong></td>
<td><strong>3.54</strong></td>
<td><strong>2.07</strong></td>
</tr>
</tbody>
</table>
Conclusions and Future Work

• We extended the concept of selectional preference for commonsense knowledge extraction

• A very preliminary work with many potential extension
  • More patterns to cover
  • More links in the KG
  • More types of relations
  • More applications

• Code and data
  • https://github.com/HKUST-KnowComp/ASER

• Project Homepage
  • https://hkust-knowcomp.github.io/ASER/

Thank you 😊