Open Extraction of General Knowledge through Compositional Semantics

Lenhart Schubert
Benjamin Van Durme
University of Rochester

Thanks: NSF (Grants IIS-0535105 and IIS-0328849),
& collaborators Ting Qian, Phil Michalak
Goal: Accumulation of General Knowledge

- A ROOM MAY HAVE WINDOWS.
- PEOPLE MAY WANT TO BE RID OF A DICTATOR.
- WHEN A PERSON ENTERS A ROOM, IT IS GENERALLY THROUGH A DOOR.
- WHEN A CAR CRASHES, THE OCCUPANTS MAY BE HURT OR KILLED.
Dreamed-of applications
Planned applications

• Improved guidance of a parser
  “He saw the bird [with binoculars]/
  [with yellow tail feathers]”

• Support for deep learning by reading
  “The mixture is ignited, and the resulting explosion
drives the piston downward”

• Building a large knowledge base for our EPILOG inference engine, to support
  commonsense reasoning (incl. self-aware agent)

Thus we need knowledge in formal, symbolic form
Episodic Logic & the EPILOG system

“A car crashed into a tree. …”

(∃ e: [e before Now34]
(∃ x: [x car] (∃ y: [y tree] [[x crash-into y] ** e])))

“The driver of car x may be hurt or killed as a result of crash e”
Knowledge Accumulation

The knowledge acquisition bottleneck: AI’s nemesis

- Is Cog/Kismet/Ripley/Toco... the answer? Will language behavior, reasoning, and KA “emerge”?
- Is CYC (hand-coding) the answer? A laudable effort; still much too small, idiosyncratic

- Our approach: General KNOWledge EXtraction from Text

A PERSON MAY BUY FOOD; A HOUSE MAY HAVE WINDOWS; A COMEDY MAY BE DELIGHTFUL; A BEHAVIOR CAN BE STRANGE; LEISURE MAY BE DEVOTED TO PLAY; ...
The KnexT project: General Knowledge Extraction from Text
(L. Schubert, M. Tong, J. Sinapov, B. Van Durme, T. Qian, …)

- Identify temporal phrases, etc.
- 80 regular phrase patterns, paired with semantic rules
- Proper name gazetteer; “of” knowledge, etc.

Sentence & phrase structure

Adjust phrase structure for interpretation

Adjusted input

Compute LFs

Sets of LFs

Extract & abstract propositions

Propositional LFs

Verbalize and filter propositions

Abstract LFs and English output

A PERSON MAY HAVE A FLU
A FLU CAN BE TERRIBLE

“Shallow” knowledge
Outputs from Treebank corpora (2002-3)
(Schubert HLT’02, Schubert & Tong, HLT Text Meaning’03)

**Brown Corpus:** 1 million words, with phrase structure  ----->
117,000 propositions (~ 60% “reasonable general claims about the world)

**Selected Brown examples:**
A PERSON MAY BELIEVE A PROPOSITION
BILLS MAY BE APPROVED BY COMMITTEES
A US STATE MAY HAVE HIGH SCHOOLS
CHILDREN MAY LIVE WITH RELATIVES
A COMEDY MAY BE DELIGHTFUL
A BOOK MAY BE WRITE-ED (i.e., written) BY AN AGENT
A FEMALE-INDIVIDUAL MAY HAVE A SPOUSE
AN ARTERY CAN BE THICKENED
A HOUSE MAY HAVE WINDOWS
PROTESTS CAN BE ADAMANT
A MALE-INDIVIDUAL MAY LEAD A FIGHT
A TEAM CAN BE WINLESS
LEGS MAY TWITCH
INDIVIDUALS MAY SHARE A BED
REVELATIONS MAY EMBARRASS TOWN OFFICIALS
A BRICK FAÇADE MAY BE SHEARED OFF BY A SHOCK OF A QUAKE
A TV-NETWORK MAY HAVE A SPOKESMAN
A BARREL MAY CONTAIN HEATING OIL
A LANGUAGE MAY BE MELLIFLUOUS
Recent work
(Van Durme & Schubert STEP’08, Van Durme, Qian & Schubert COLING’08)

• Use of state-of-the-art parsers: Collins, Charniak

• Improved filtering: “red rose” ----> A ROSE CAN BE RED
  “legal secretary” ---|--> A SECRETARY CAN BE LEGAL

• Evaluation as class-attribute extractor:
  Food may have: quality, taste, value,…
  An apple can be: red, juicy, fresh, bad, …

• Evaluation on 100 million word corpora

British National Corpus: 100 million words, analyzed with Collins’ parser ----> 6 million propositions

Sample sentences from Web:* 100 million words, analyzed with Charniak’s parser ----> 6 million propositions

About 2/3 judged “reasonable general claims about the world”

* Thanks to Michele Banko
Current & future work

- **Abstracting conditionals from sets of similar propositions:**
  
  A BIOLOGIST MAY WRITE A BOOK WHEN SOME X WRITES Y, 
  A CHILD MAY WRITE A POEM Y IS PROBABLY A COMMUNICATION ... ETC ...

- **Disambiguating light predicates** (**HAVE, FOR ...**)
  
  A DOCTOR MAY HAVE AN ACCIDENT ← EVENT NOMINAL 
  A DOCTOR MAY HAVE A PATIENT ← NON-EVENT NOMINAL

- **Identifying nominal predicates with missing arguments**
  
  LOTS [?] MAY HAPPEN 
  AN ATTACK MAY COME FROM THE BACK [?]

- **Guiding a parser with semantic patterns**
  
  (“He saw the bird with binoculars”) (cf. work at Boeing)

- **Learning conditionals from sentences with multiple verbs**
  
  (“A car crashed into a tree, injuring the driver”)

- **Deeper learning by reading**
  
  Reading WordNet glosses, Open Mind, Wikipedia, ...

- **Building self-motivated, self-aware agents that communicate, reason, and plan**
Conclusions

• Can readily derive millions of general “factoids” in symbolic logical form from electronic texts

• It seems possible to bootstrap this process by feeding the factoids back into parsing/LF process

• Derivation of stronger, conditional knowledge from sets of related factoids is feasible

• Disambiguation, filling missing arguments, acquiring reliable conditional knowledge remain difficult – but not mystifying -- challenges

LET’S GO FOR IT!!
The following references are available at http://www.cs.rochester.edu/~schubert/


