How to Reconcile Information Theory and Natural Language Semantics

“Information after Shannon”

Istituto Veneto

December 29-30, 2008

Victor Raskin

Purdue University

vraskin@purdue.edu

http://web.ics.purdue.edu/~vraskin/Raskin.html

Probability “vs.” Natural Language

- Personal memoir: information theory shock—not transmit infrequent signal?!
- Semantic events are infrequent: statistics works poorly
- The long tail in Internet search: where the Google beauty pageant fails and where all the pertinent information resides

Natural Language Information

- Most (?) information comes in natural language (NL)

- No computer application without understanding NL
  - Underlying production and comprehension rules
  - Users with low error tolerance
  - Observable output in principle irregular
- No Understanding NL without Semantics
  - Logic form conversion is not understanding
  - Surface co-occurence statistics is not understanding
  - Automatic semantic tagging presupposes understanding

Natural Language Information: Two Information Theory Points

- Compression: not frequency but focus
- Nature of information: Ontological Semantics (OntoSem)

Natural Language Information: Compression

- Where is the Workshop “Information after Shannon” taking place?
- The Workshop “Information after Shannon is taking place at the Istituto Veneto in Venice?
- (In) Venice.

Natural Language Information: Focus

- Wh-words: where, when, how, who, what, which
- General questions: compressed to yes or no
- Given-new = presupposition-focus

So, what is OntoSem?

- Databases and software that transform natural language text into a text meaning representation (TMR)
- TMR approximates human understanding of text
- Basis of multiple computer applications emulating human intellectual abilities

OntoSem Resources

- Language-Independent Ontology (conceptual hierarchy)
- Lexicons (one for each natural language, e.g., English)
- Onomasticons (lexicons of proper names, one for each natural language)
- Analyzer (text to meaning software)
- Generator (meaning to text software)

Ontology Top Level

ALL

- Objects
- Events
- Properties

Ontology Event Top Level

Events

- Mental events
- Social events
- Physical events

Objects: Two Top Levels

Objects

- Intangible object
  - force
  - energy
- Physical object
  - animate
  - inanimate
  - computer data
  - physical systems

Objects cont.

- Mental object
  - abstract object
  - representational object
- Social object
  - geopolitical entity
  - organization

Properties: Two Top Levels

Properties

- Case roles
  - agent
  - beneficiary
  - destination
  - experiencer
  - instrument
  - location
  - path
  - purpose
  - source
  - theme

Properties cont.

- Attributes
  - Literal attribute
    - object
      - physical
      - social
    - event
  - Scalar attribute
    - object
    - event

Examples of:

Ontological Concept

go

is-a motion-event

agent animal *instrument*
body-part, vehicle

source location

destination location

start-time temporal-unit

end-time temporal unit

Lexical Entry

drive-V1

[all but semantic information omitted]

sem-struc

go

agent human
& adult

instrument car

Simplified TMR

- Mary drove from Boston to New York on Wednesday
- *go*

agent	Mary	
	instrument	car
	source	Boston
	destination	New York
	start-time	Wednesday
	end-time	Wednesday

Caution to Workshop Participants

- The previous slide was the last slide of the Workshop presentation. You are done!

- The real “meat” of OntoSem starts with the next slide: proceed at your own risk and peril!

How does OntoSem Work?

- And now it’s going to get really complicated and detailed.
- Sorry, language actually is really complicated. We tend to forget that, because we are naturally so good at it.
- But a dumb machine has to be taught in minutest detail all that we humans do effortlessly and unconsciously.
- Here’s how OntoSem processes the meaning of:
  “Did Bush kill the last bill in the senate?”

Did Bush kill the last bill in the senate?

```
(bush
     (bush-n1
       (cat n)
       (anno(def "")(ex "")(comments ""))
       (syn-struc((root $var0)(cat n)))
       (sem-struc(bush))
     )
     (bush-v1
       (anno(def "")(comment(transitive)))
       (cat v)
       (syn-struc
              ((root $var0)(cat v)(subject((root $var1)
              (cat n)))(directobject((root $var2)(cat n)))))
(sem-struc
              (protect
                     (agent(value ^$var1))
                     (theme(value ^$var2)(sem tree))
                     (instrument(sem bush))))
     )
     (bush-v2
       (anno(def "")(comment(transitive)))
       ...
(sem-struc
              (supply
                     (agent(value ^$var1))
                     (beneficiary(value ^$var2))
                     (theme(sem bush))))
     )
     ...
)
```

what happens:

- lexicon lookup for “Bush”
- 1 entry with 10 senses found
- 2 verbs, 8 nouns
- part-of-speech tagging disambiguates to noun senses

we see:

- one example for a noun sense
- two examples for verb senses, one full, the other abbreviated
- most importantly **sem-strucs** that give the sense in terms of ontological concepts, here BUSH, PROTECT, and SUPPLY as the head concepts of the sem-strucs

Did Bush kill the last bill in the senate?

```
(Bush
    (Bush-n1
        (pos n)
        (anno(def ”43rd U.S. president"))
        (syn-struc
            (((root $var3)(value President)(cat n)(opt +))
            ((root $var2)(value George)(cat n)(opt +))
            ((root $var1)(value W)(cat n)(opt +))
            ((root $var4)(value .)(cat period)(opt +))
            ((root $var1)(value Walker)(cat n)(opt +))
            (root $var0)(cat n)))
        (sem-struc
            (president
                (has-first-name(value "George"))
                (has-middle-name(value (or "W” “W.” “Walker”)))
                (has-last-name(value "Bush"))
                (has-political-party(value "Republican"))
                (location
                    (sem country(has-name(value "United-States"))))))
    )
    (Bush-n2 ... “George Herbert Walker” ...)
    (Bush-n3 ... “Jeb” ...)
    (Bush-n4 ... "Laura” ...)
    (Bush-n5 ... “Barbara 1” ...)
    (Bush-n6 ... ”Jenna” ...)
    (Bush-n7 ... ”Barbara 2” ...)
    )
)
```

what happens:

- capitalization disambiguates to onomasticon entries, i.e., names
- semantic priming (ordering of senses) prefers Bush-n1 (but doesn’t exclude the other senses) in case no further constraints are found
- syn-struc contains further surface clues, surrounding words, in particular for multi-word entries

we see:

- other senses from onomasticon
- only the first shown completely

Did Bush kill the last bill in the senate?

```
(kill
     (kill-n1 ... “a kill event” ...)
     (kill-n2 ... “the theme of a kill event, esp. animal” ...)
     (kill-v1
          (cat v)
          (synonyms "murder-v1")
          (morph)
          (anno
               (def "to cause to die; with an agent")
               (ex "the intruder killed him"))
          (syn-struc
               ((subject((root $var2)(cat np)))
               (root $var0)(cat v)
               (directobject((root $var3)(cat np)))
               (pp-adjunct((root (or by with through))(cat prep)(obj((root $var4)(cat np))))(opt +))))
          (sem-struc(kill
                         (agent(value ^$var2))
                         (theme(value ^$var3))
                         (instrument(value ^$var4))))
     )
     (kill-v2
          (cat v)
          (morph)
               (anno(def "to cause to die”)
               (ex "that disease killed him"))
          (syn-struc
               ((subject((root $var1)(cat np)))
               (root $var0)(cat v)
               (directobject((root $var2)(cat np)))))
          (sem-struc
               (die
                    (theme(value ^$var2))
                    (caused-by(value ^$var1)(sem event))))
     )
     ...
)
```

what happens:

- lexicon lookup for “kill”
- 1 entry with 2 noun senses and 5 verb senses are found
- the part-of-speech tag excludes the noun senses

we see:

- a summary of the noun senses
- the first two verb senses

Did Bush kill the last bill in the senate?

```
(
        ...
        (kill-v3
                (cat v)
                (morph)
                (anno
                        (def "to cause to die; with the subject being an instrument”)
                        (ex "the bullet killed him"))
                (syn-struc
                        ((subject((root $var1)(cat np)))
                        (root $var0)(cat v)
                        (directobject((root $var2)(cat np)))))
                (sem-struc
                        (kill
                                (theme(value ^$var2))
                                (instrument(value ^$var1)(sem object))))
        )
        (kill-v4
                (cat v)
                (morph)
                (anno
                        (def "to cause to cease operating; of a device")
                        (ex "he killed the motor"))
                (syn-struc
                        ((subject((root $var1)(cat np)))
                        (root $var0)(cat v)
                        (directobject((root $var2)(cat np)))))
                (sem-struc
                        (operate-device
                                (phase end)
                                (agent(value ^$var1))
                                (theme(value ^$var2)(default (or device vehicle))(sem artifact))))
        ...
        )
)
```

we see:

- the next 2 verb senses
- note in particular **additional constraints defined in the sem-struc** that further specify constraints of the head concepts, KILL and OPERATE-DEVICE

Did Bush kill the last bill in the senate?

```
(
    ...
    (kill-v5
        (cat v)
        (morph)
        (anno
            (def     ”to end the debate about a document’s
                      acceptance by the legislative body” )
            (ex "they killed this bill")(comments ""))
        (syn-struc
            ((subject((root $var1)(cat np)))
            (root $var0)(cat v)
            (directobject((root $var2)(cat np)))))
        (sem-struc
            (veto
                (agent(value ^$var1)(sem political-role))
                (theme(value ^$var2)(sem bill-legislative)))))
)
```

we see:

- the last verb sense, which will be identified as correct

Did Bush kill the last bill in the senate?

```
(bill
     (bill-n1
          (cat n)
          (anno(def "itemized statement of fees, charges”))
          (syn-struc((root $var0)(cat n)))
          (sem-struc(bill))
     )
     (bill-n2
          (cat n)
          (anno(def “a list of legal statements ...”))
          (syn-struc((root $var0)(cat n)))
          (sem-struc(bill-legislative))
     )
     (bill-n3
          (cat n)
          (anno(def "phrasal: bill of exchange")(ex "")(comments ""))
          (syn-struc
               ((root $var0)(cat n)
               (pp-adjunct((root of)(root $var1)(cat prep)(obj((root $var2)(cat n)(root exchange))))))
          (sem-struc(bill-of-exchange))
     )
     (bill-n4
          (cat n)
          (anno(def "phrasal: bill of rights")(ex "")(comments ""))
          (syn-struc
               ((root $var0)(cat n)
               (pp-adjunct((root of)(root $var1)(cat prep)
                    (obj((root $var2)(cat n)(root right)(number pl)))))))
          (sem-struc(bill-of-rights))
     )
     ...
)
```

what happens:

- lexicon lookup for “bill”
- retrieves 1 entry with 7 senses, 5 noun and 2 verb senses
- the part-of-speech tag discards the verb senses

we see:

- the first 4 noun senses of “bill”

Did Bush kill the last bill in the senate?

```
(
     ...
     (bill-n5
          (cat n)
          (anno(def "a beak")(ex "")(comments ""))
          (syn-struc((root $var0)(cat n)))
          (sem-struc(beak))
     )
     (bill-v1
          (cat v)
          (morph)
          (anno
               (def "send s.o. a bill for a service or item")
               (ex "he billed me for the job"))
          (syn-struc
               ((subject((root $var1)(cat np)))
               (root $var0)(cat v)
               (directobject((root $var2)(cat np)))
               (pp-adjunct(opt +)((root for)(cat prep)(obj((root $var4)(cat np)))))))
          (sem-struc
               (send
                    (agent(value ^$var1))(theme(value refsem1))
                    (beneficiary(value ^$var2))
                    (refsem1(bill))
                    (refsem2(relation(domain(value refsem1))(range(value ^$var4))))
     )
     (bill-v2
          (anno(def "put up an advertizing bill")(comment(transitive)))
          (cat v)
          (syn-struc
               ((subject((root $var1)(cat np)))
               (root $var0)(cat v)
               (directobject((root $var2)(cat np)))))
          (sem-struc
               (advertise
                    (agent(value ^$var1))
                    (theme(value ^$var2)(sem human))))
     )
)
```

we see:

- the last noun sense and the 2 verb senses of “bill”

Did Bush kill the last bill in the senate?

```
(senate
     (senate-n1
          (pos n)
          (syn-struc((root $var0)(cat n)))
          (sem-struc(senate))
     )
)
```

what happens:

- lookup retrieves the only sense of "senate"

we see:

- the one noun sense

Taking Stock

- 7 noun senses for “Bush” (out of 10)
- 5 verb senses for “kill” (out of 7)
- 5 noun senses for “bill” (out of 7)
- 1 noun sense for “senate” (out of 1)
- after lexicon lookup 10 x 7 x 7 x 1 = 490 meanings
- after syntactic analysis: 7 x 5 x 5 x 1 = 175 meanings

TMR selection and filling

- Text meaning representations (TMRs) per clause are built on EVENTs
- the sentence has only one clause
- “kill” is the only supplier of EVENT senses
- setting up potential TMRs based on EVENT senses of kill
- maximizing the filled case roles for these TMRs (AGENT, THEME, INSTRUMENT,...)

noun senses for “Bush”

- PRESIDENT “George W. Bush”
- PRESIDENT “George H.W. Bush”
- SOCIAL-ROLE
- SOCIAL-ROLE
- SOCIAL-ROLE
- SOCIAL-ROLE
- SOCIAL-ROLE

Did Bush kill the last bill in the senate?

```
president
     definition
         "the chief executive of a republic"
     is-a
        elected-governmental-role
              is-a
                  governmental-role
                       is-a
                           social-role
                                ...
     head-of
         multiparty-presidential-regime
     beneficiary-of
         elect
     agent-of
         run-for-office
     ...
```

we see:

- the ontological concept for the correct sense PRESIDENT and several of its properties, in particular
- its location in the ontology as a grandchild of GOVERNMENTAL-ROLE
- thus, PRESIDENT meets the constraint of VETO to have a GOVERNMENTAL-ROLE as AGENT

verb senses for “kill”

- KILL
- DIE
- KILL
- OPERATE-DEVICE
- VETO

Did Bush kill the last bill in the senate?

```
veto
    definition
        "to prohibit action or legislation"
    is-a
        political-event
            is-a
                social-event
    agent
        governmental-role (default)
        political-role (relaxable-to)
    theme
        legal-object
    has-event-as-part
        prohibit
    ...
```

we see:

- the ontological concept for the correct sense VETO and several of its properties, in particular
- the AGENT that is a GOVERNMENTAL-ROLE by default, but can be relaxed to the more general POLITICAL-ROLE, and
- the THEME that is a LEGAL-OBJECT

noun senses for “bill”

- BILL
- BILL-OF-EXCHANGE
- BILL-OF-RIGHTS
- BILL-LEGISLATIVE
- BEAK

Did Bush kill the last bill in the senate?

```
bill-legislative
     definition
         ”a bill that comes up before a legislative institution"
     is-a
        legal-object
             is-a
                  representational-object
                       ...
     has-object-as-part
         law
     represented-by
         language-related-object
     theme-of
         veto
         vote
         approve
         ...
     ...
```

we see:

- the ontological concept for the correct sense BILL-LEGISLATIVE and several of its properties, in particular
- that it is the THEME-OF events like VETO, VOTE, etc.

noun senses for “senate”

- SENATE

Did Bush kill the last bill in the senate?

```
senate
    definition
        "the upper branch of a two-branch legislature"
    is-a
        legislative-branch
            is-a
                government-branch
                    ...
    agent-of
        approve
        revoke
        vote
        ...
    part-of-object
        governmental-parliament
    object-involved
        law (default)
    member-type
        senator (default)
        governmental-role (relaxable-to)
    ...
```

we see:

- the ontological concept for the correct sense SENATE and several of its properties, in particular
- that it’s the AGENT-OF for many of the very EVENTs for which BILL-LEGISLATIVE as a LEGAL-OBJECT is the THEME-OF

selecting the right sense of “kill” and the fillers for its case roles

EVENT	CASE-ROLE	CONSTRAINT(1)	POTENTIAL FILLER SENSES(2)	
kill				
	agent	animate	president, ...	
	theme	animate	*none*	
	location	place	senate	
	...			leftover: “bill”
die				
	agent	*none*	*none*	
	caused-by	event	*none*	
	location	place	senate	
	...			leftover: “bill” “Bush”
operate-device				
	agent	human	president, ...	
	theme	device	*none*	
	location	place	senate	
	...			leftover: “bill”
veto				
	agent	**governmental-role**	**president, ...**	
	theme	**legal-object**	**bill-of-rights, bill-legal**	
	location	**place**	**senate**	
	...			**leftover: *none***

(1) both from the ontological head concept and as (further) specified in the sense entry
(2) meets the semantic constraints as well as the syntactic ones

Did Bush kill the bill in the senate?

what happens:

- the VETO sense of “kill” has both its constraints for the AGENT to be a GOVERNMENTAL-ROLE and for the THEME to be a LEGAL-OBJECT met by the sentence and is chosen as the EVENT that can accommodate the highest number of senses of the nouns.
- DIE is excluded, as there is no EVENT sense among the noun senses to fill the CAUSED-BY and because it can’t accommodate any senses of two nouns, “Bush” and “bill”.
- KILL is excluded as there is no second ANIMATE to fill the THEME slot.
- OPERATE-DEVICE is excluded because of the constraint that the THEME is an ARTIFACT, usually a DEVICE or VEHICLE, which no sense of “bill” meets.
- BILL-LEGAL is chosen as the more generic filler for the THEME of VETO.
- among the two senses of PRESIDENT, the 41st and the 43rd presidents of the U.S., the one that is primed by order is chosen: George W. Bush.

Final Result

- after lexicon lookup 10 x 7 x 7 x 1 = 490 sentence meanings
- after syntactic analysis: 7 x 5 x 5 x 1 = 175 sentence meanings
- after OntoSem analysis: 1 correct sentence meaning

Where Keywords can’t Go

keyword and enhanced keyword approaches fail

1. false positives:
   - this new pesticide kills moss even under a bush
   - the cheetah hid its fresh kill in the bushes
2. misses:
   - the potus vetoed the proposal
   - another amendment was shot down by the white house
   - we heard the swan song for senator kennedy’s motion

Where Markup/Semantic Web can’t Go

- Users won’t do It! So there will be no semantic web.
- Want simplicity, generality, uniformity, low cost, and ease?
- Sure, automate!
- Go where you can find it—not where the street light is and you can continue to use your favorite methods: playing with with formalisms
- Or go to meaning processing system: OntoSem
- But then you don’t need the Semantic Web anymore.

Where OntoSem Can Go

- Relates a text to a much larger number of texts on semantic, meaningful connections and associations, like a human
- Pursues inferences, entailments, presuppositions, etc.
- Catches relevant web pages even if the actual words in a query do not appear there
- Rejects irrelevant pages even if some actual words in a query appear in it
- Improves the quality of the search beyond anything attainable by a bag-of-words method
- Introduces a new era of human-caliber searches: Can you imagine a search engine that understands the language of your query like a human does, but can also understand the meaning of all webpages as well as remember them all to find the best answers for you?

Applications of OntoSem in Search and Beyond

- Relates a text to a much larger number of texts on semantic, meaningful connections and associations, like a human
- Pursues inferences, entailments, presuppositions, etc.
- Catches relevant web pages even if the actual words in a query do not appear there
- Rejects irrelevant pages even if some actual words in a query appear in it
- Improves the quality of the search beyond anything attainable by a bag-of-words method
- Introduces a new era of human-caliber searches: Can you imagine a search engine that understands the language of your query like a human does, but can also understand the meaning of all webpages as well as remember them all to find the best answers for you?