

# Computing Natural-Language Meaning for the Semantic Web

*Patrick Hanks*

Faculty of Informatics, Masaryk University, Brno

Riga: day 4

**29 November, 2007**

# Why are people so excited about the Semantic Web idea?

- It offers “unchecked exponential growth” of “data and information that can be processed automatically”
  - Berners-Lee et al., *Scientific American*, 2001
- Distributed, not centrally controlled
  - but with scientists as ‘guardians of truth’? -Wilks
- “... paradoxes and unanswerable questions are a price that must be paid to achieve versatility.”
  - Berners-Lee et al. 2001

# Aims of the Semantic Web

- “To enable computers to manipulate data meaningfully.”
- “Most of the Web's content today is designed for humans to read, not for computer programs to manipulate meaningfully.”  
– Berners-Lee et al., 2001

# The “Resource Description Framework”

- A strictly defined tagging language for classifying documents and parts of documents, and relating them to one another
- “a language for lightweight ontology descriptions” -- Hayes 2004
  - Current SW activities include classifying and relating documents, names, dates, addresses, etc.

# Meaning in unstructured text

- The Semantic Web is “the apotheosis of annotation” ...
  - “But what are its semantics?”
  - “Available information for science, business, and everyday life still exists overwhelmingly as text ... unstructured data.”
    - Y. Wilks, 2006
  - How can the meaning of natural language in texts be made available for SW inferencing?
  - In addition to annotation and mark-up

# Ontologies

- **SW ontologies** are, typically, interlinked networks of things like address lists, dates, events, and websites, with html mark-up showing attributes and values
- They differ from **philosophical ontologies**, which are theories about the nature of all the things in the universe that exist
- They also differ from **lexical ontologies** such as WordNet, which are networks of words with supposed conceptual relations

# Hypertext

- “The power of hypertext is that anything can link to anything.”
  - Berners-Lee et al., 2001
- Yes, but we need procedures for determining (automatically) what counts as a *relevant* link, e.g.
  - *Firing a person* is relevant to employment law.
  - *Firing a gun* is relevant to warfare and armed robbery.

# A paradox

- “Traditional KR systems typically have been centralized, requiring everyone to share exactly the same definition of common concepts such as 'parent' or 'vehicle'.”
  - Berners-Lee et al., 2001.
  - Implying that SW is more tolerant?
  - Apparently not:
- “Human languages thrive when using the same term to mean somewhat different things, but automation does not.” --*Ibid.*

# What is to be done?

- Process only the (strictly defined) mark-up of documents, not their linguistic content?
  - And so abandon the dream of enabling computers to manipulate linguistic content?
- Force humans to conform to formal requirements when writing documents?
  - Practical only in limited domains
- Teach computers to deal with natural language in all its fearful fuzziness?
  - Maybe this is what we need to do

# The Root of the Problem

- Word meaning in natural language is vague
- Scientists from Wilkins and Leibniz to the present day have wanted it to be precise
- See Umberto Eco, *The Search for the Perfect Language*.
- **Do not allow SW research to fall into this trap**

# The Paradox of Natural Language

- Word meaning is vague and fuzzy
- Yet people can use words to make very precise statements
  - Why? In part because text meaning is holistic, e.g.
  - “*fire*” in isolation is very ambiguous;
  - “*He fired the bullet that was recovered from the girl's body*” is not at all ambiguous.
  - “**Ithaca**” is ambiguous; “***Ithaca, NY***” is much less ambiguous
  - An inventory of phraseological patterns is needed.

# Precise definition does not help discover implicatures

- The meaning of the English noun *second* is vague: “a short unit of time” and “1/60 of a minute”.
  - *Wait a second.*
  - *He looked at her for a second.*
- It is also a very precisely defined technical term in certain scientific contexts, the basic SI unit of time:
  - “the duration of 9,192,631,770 cycles of radiation corresponding to the transition between two hyperfine levels of the ground state of an atom of caesium 133.”

# Precision and Vagueness

- Giving a precise definition to an ordinary word removes it from ordinary language.
- When it is given a precise, stipulative definition, an ordinary word *becomes* a technical term.
- “An adequate definition of a vague concept must aim not at precision but at vagueness; it must aim at precisely that level of vagueness which characterizes the concept itself.”
  - Wierzbicka 1985, pp.12-13

# Eliminating vagueness while preserving meaning

- Vagueness is reduced or eliminated when a word is used in context.
- We need to discover the **normal** contexts in which each word is used.
- This can be done by corpus pattern analysis.
- On this basis applications such as SW could build inferences about text meaning.
- But how can word meaning be linked to word use?

# Computing meaning

- An alternative to check-list theories of meaning
  - Fillmore, 1975
    - Compute closeness to a prototype, rather than looking for satisfaction of necessary and sufficient conditions
    - But an inventory of prototypical patterns of word use still does not exist!
    - In Brno we are building such an inventory -- *the Pattern Dictionary of English Verbs*

# Corpus Pattern Analysis (CPA)

1. Identify usage patterns for each word
  - **Patterns include semantic types and lexical sets of arguments (valencies)**
2. Associate a meaning (“implicature”) with each pattern (**NOT** with each word)
3. Match occurrences of the target word in unseen texts to the nearest pattern (“norm”)
4. If 2 matches are found, choose the most frequent
5. If no match is found, it is not normal usage -- it is an exploitation of a norm (or a mistake).

# How useful are standard dictionaries for SW inferencing?

- Dictionaries show very little semantic structure.
- Dictionaries don't show syntagmatic preferences.

# Taxonomy

- Ontologies such as WordNet and the Brandeis Semantic Ontology show words linked to a taxonomy of semantic types, e.g.

- *a gun, pistol, revolver, rifle, cannon, mortar, Kalashnikov, ... is a:*

weapon

artefact

physical object (or ‘material entity’)

entity

- Can such taxonomies be used for SW? If so, how?

# Ontological reasoning

## EXAMPLE:

If it's a *gun*, it must be a *weapon*, an *artefact*, a *physical object*, and an *entity*, and it is used for *attacking* people and things.

- Otherwise known as 'semantic inheritance'
- So far, so good
- **How useful is ontological information such as this as a basis for verbal reasoning?**
- Not as useful as we would like for NLP applications such as word sense disambiguation, semantic web, text summarization, etc.

# Semantics and Usage (1)

- *He was pointing a **gun** at me*  
-- is a Weapon < Physical Object.

***BUT***

## 2. *A child's toy **gun***

-- is a Toy (“Entertainment Artifact”), not a Weapon

## 3. *The fastest **gun** in the west*

-- is a Human < Animate Entity, not a Weapon

- “must be a weapon” on the previous slide is too strong; should be “is probably a weapon”
- probabilities can be measured, using corpus data
- The normal semantics of terms are constantly exploited to make new concepts (as in 2 and 3)

# Semantics and Usage (2)

- Knowing the exact place of a word in a semantic ontology is not enough
- To compute meaning, we need more info....
- Another major source of semantic information (potentially) is usage:
  - how words go together (normally | unusually | never)
- How do patterns of usage (syntagmatic) mesh with the information in an ontology?

# The Semantics of Norms

- Dennis closed his eyes and **fired** the gun
  - [[Human]] fire [[Firearm]]
- He **fired** a single round at the soldiers
  - [[Human]] fire [[Projectile]] {at [[PhysObj = Target]]}
    - BOTH MEAN: [[Human]] causes [[Firearm]] to discharge [[Projectile]] towards [[Target]]
- Rumsfeld **fires** anyone who stands up to him.
  - [[Human 1 = Employer]] fire [[Human 2 = Employee]]
    - MEANS discharge from employment
    - The **roles** Employer and Employee are assigned by context -- not part of the type structure

# Complications and Distractions

Minor senses:

- reading this new book **fired** me with fresh enthusiasm to visit this town
  - [[Event]] fire [[Human]] {with [[Attitude = Good]]}
- Mr. Walker **fired** questions at me.
  - [[Human 1]] fire [[Speech Act]] {at [[Human 2]]}

Named inanimate entity:

- I ... got back on Mabel and **fired** *her* up.
  - Mabel is [[Artifact]] (a motorbike, actually)
  - [[Human]] fire [[Artifact > Energy Production Device]] {up}

# Ontology-based reasoning

- If it's a *gun*, it's a *physical object*, so whatever you can do with a physical object, you can do with a gun:
  - you can *touch* it
  - you can *see* it
  - it has to be somewhere (has *physical extension*)

# Collocations and Types don't match

From Word Sketch Engine:

freq. of '*weapon*': in BNC 5,858; in OEC 115, 836

Collocate (verb with <i>weapon</i> as direct object)	Frequency of the collocation		Salience	
	in BNC	in OEC	BNC (rank)	OEC (rank)
carry	107	1021	32.08 (1)	43.85 (6)
surrender	23	95	29.87 (2)	34.86 (23)
possess	35	771	28.75 (3)	56.86 (1)
use	167	2993	28.08 (4)	45.50 (5)
deploy	18	179	26.03 (5)	39.69 (15)
fire	21	599	23.44 (6)	47.91 (4)
acquire	15	601	16.24 (17)	50.72 (2)

# What do you do with a gun?

Word Sketch Engine: freq. of *gun*: BNC 5,269; OEC 91,781

Collocate (verb with <i>gun</i> as object)	Frequency of collocation		Salience (rank)	
	BNC	OEC	BNC	OEC
fire	104	1132	45.39 (1)	60.96 (2)
point	59	1639	30.80 (2)	61.37 (1)
carry	85	974	28.42 (3)	44.87 (10)
<b>jump</b>	31	434	27.77 (4)	46.35 (8)
brandish	11	98	25.86 (5)	42.55 (14)
wave	20	249	20.58 (6)	---
hold	70	1504	20.38 (7)	44.79 (11)
aim	-	663	-	54.96 (4)
load	-	330	-	48.70 (7)

# Shimmering Lexical Sets (1)

- ***weapon***: carry, surrender, possess, use, deploy, fire, acquire, conceal, seize, ...

- 
- ***gun***: fire, carry, point, jump, brandish, wave, hold, cock, spike, load, reload, ...
  - ***rifle***: fire, carry, sling (over one's shoulder), load, reload, aim, drop, clean, ...
  - ***pistol***: fire, load, level, hold, brandish, point, carry, wave, ...
  - ***revolver***: empty, draw, hold, carry, take, ...

# Shimmering Lexical Sets (2)

- *spear*: thrust, hoist, carry, throw, brandish
- *sword*: wield, draw, cross, brandish, swing, sheathe, carry, ...
- *dagger*: sheathe, draw, plunge, hold
- *sabre*: wield, **rattle**, draw
- *knife*: brandish, plunge, twist, wield
- *bayonet*: fix

# Shimmering Lexical Sets (3)

- *missile*: fire, deploy, launch
- *bullet*: bite, fire, spray, shoot, put
- *shell*: fire, lob; crack, ...
- *round*: fire, shoot; ...
- *arrow*: fire, shoot, aim; paint, follow

# Shimmering Lexical Sets (4)

- **fire:** shot, gun, bullet, rocket, missile, salvo ...  
[[Projectile]] or [[Firearm]]
- **carry:** passenger, weight, bag, load, burden, tray, weapon, gun, cargo ... [polysemous]
- **aim:** kick, measure, programme, campaign, blow, mischief, policy, rifle ... [polysemous]
- **point:** finger, gun, way, camera, toe, pistol ...  
[polysemous?]
- **brandish:** knife, sword, gun, shotgun, razor, stick, weapon, pistol ... [[Weapon]]
- **shoot:** glance, bolt, Palestinian, rapid, policeman;  
– **shoot ... with:** pistol, bow, bullet, gun

# Triangulation

- Words in isolation don't have meaning, they have **meaning potential**
- Meanings attach to **patterns**, not words
- A typical pattern consists of a verb and its arguments (with semantic values), thus:  
[[Human]] **fire** [[Projectile]] {from [[Firearm]]}  
{PREP [[PhysObj]]}
- Pattern elements are often omitted in actual usage.  
(See FrameNet)

# Semantic Type vs. Semantic Role

[[Human]] fire [[Firearm]] {at [[PhysObj]]}

[[Human]] fire [[Projectile]] {at [[PhysObj]]}

*Bond walks into our sights and fires his pistol **at the audience***

*The soldier fired **a single shot at me***

*The Italian authorities claim that three US soldiers fired **at the car**.*

- ‘**audience**’, ‘**me**’, and ‘**car**’ have the semantic type [[Human]] and [[Vehicle]] (< [[PhysObj]])
- The context (pattern) assigns the semantic role Target

# Lexical sets don't map neatly onto semantic types

- *calm* as a transitive (causative) verb:
- What do you calm? 1 lexical set, 5 semantic types:
  - *him, her, me, everyone*: [[Human]]
  - *fear, anger, temper, rage*: [[Negative Feeling]]
  - *mind*: [[Psychological Entity]]
  - *nerves, heart*: [[Body Part]] but not *toes, chest, kidney*)
  - *breathing, breath*: [[Living Entity Relational Process]]  
(but not *defecation, urination*)
    - 3 criterial types here, and 2 peripheral?

# Two Different Problems

- Ontologies such as Roget and WordNet attempt to organize the lexicon as a representation of 2,500 years of Aristotelian scientific conceptualization of the universe.
- This is not the same as investigating how people use words to make meanings.
- Why ever did we think it would be?

# Conclusions

- Word meaning is vague, but the vagueness can be captured -- and measured
- In context, word meaning often becomes precise
  - But it can also be creative
- We must distinguish precision from creativity
- To do reliable inferencing on ordinary language texts for SW applications, we need to compare actual usage with patterned norms, and chose the best match
- Therefore, we need inventories of patterned norms such as the *Pattern Dictionary of English Verbs*