Automatic Semantic Acquisition Using the Web

Jesse English

Institute for Language and Information Technologies

University of Maryland, Baltimore County

5/3/2007

Overview

- Introduction to Natural Language Processing
- Ontological Semantics (gracefully)
- “Learning By Reading”

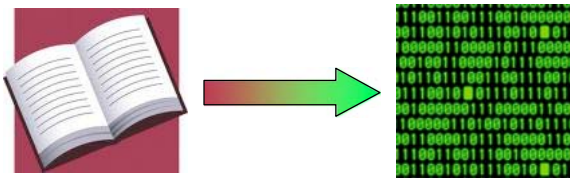
NLP

What is Natural Language Processing (NLP)?

- Text processing
- Meaning Extraction

NLP

- The goal of any NLP system is to take some (raw) text, and produce a machine-understandable “translation”.
- Ideally, there will be no loss of understanding.



NLP - What is it good for?

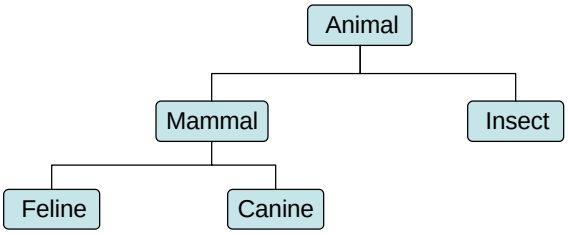
- Machine Translation
  - "Serves you right!"
    - AltaVista Babel Fish: "Services vous droit!"
    - Actual French Speakers: "Et toc! C'est bien fait pour toi!"
- Question/Answering
  - Who insulted Bush recently?
    - Sean Penn called the president our country's most devastating enemy.
    - Martin Sheen slammed the president yesterday for failing to tackle climate change.
    - A bush was recently insulted by an angry drunk.

Ontological Semantics

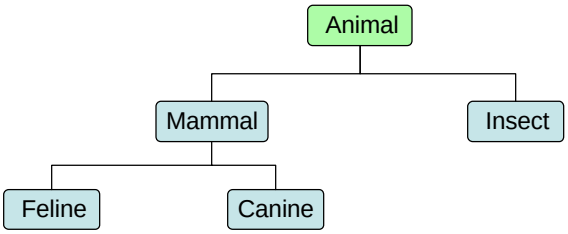
- A theory that supports NLP through the use of a semantically rich ontology and lexicon.
  - Semantically rich?
    - human-n1: a person
    - human-n1: a person, has-object-as-part=head

What is an Ontology?

- A tree-like inheritance structure (convenient for describing hyponym and hypernym relations)

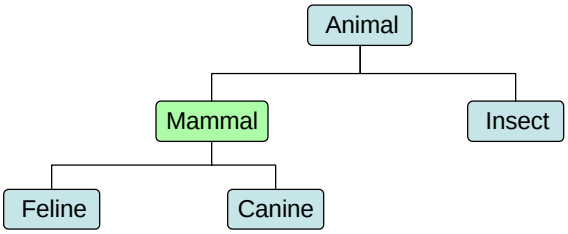


What is an Ontology?



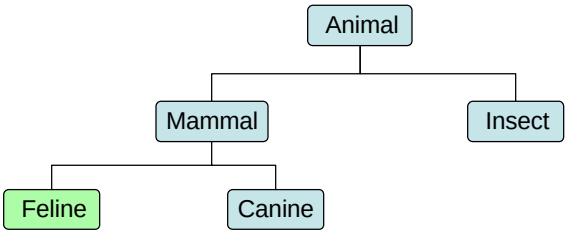
LIFESPAN = (<> 1 100)

What is an Ontology?



LIFESPAN = (<> 1 100)
BIRTH-METHOD = LIVE-BIRTH
HAS-OBJECT-AS-PART = HAIR

What is an Ontology?



LIFESPAN = (<> 1 12)
BIRTH-METHOD = LIVE-BIRTH
HAS-OBJECT-AS-PART = HAIR
DEFAULT = FUR
NUM-LIVES = 9

What is a Lexicon?

- A dictionary of lexical terms, mapped (with constraints) to ontological entries.
- tabby-n1
  - A FELINE WHERE COLORATION=STRIPED
  - Usage: ((ROOT v1) (noun))

OntoSem

- An ontological semantic NLP system developed by the ILIT lab at UMBC.
- Over 8000 concepts (some 90000 unique property/values)
- Over 16000 unique lexical senses
- Vast onomasticon, growing fact repository

“Learning By Reading”

- A hybrid branch of the NLP and ML communities
- Using some form of NLP, can a machine take raw text, and annotate it sufficiently in order to use the knowledge in the future?

Our Method: Overview

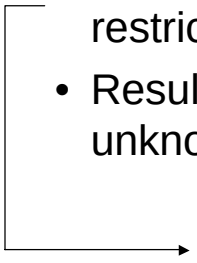
- Unknown word
  - Raw text from the web
  - Syntactic filtering
- Filtered Sentences
  - Semantic analysis
  - Property/value extraction
- Property/value pairs
  - Similarity comparison
  - Concept insertion

Our Method: Step 1

- An unknown word (or phrase) is discovered or selected manually.
- Google is queried for results on the unknown term.
- Returned web pages are stripped of HTML, and split into sentences.
- Syntactic filtering is performed to remove “web junk”.

Our Method: Step 2

- The remaining sentences are semantically analyzed using OntoSem.
- Unidirectional application of selectional restrictions… huh?
- Resulting property/value pairs for the unknown terms are extracted.



BAKE-EVENT: AGENT = BAKER
THEME = PASTRY

The baker cooked up a foobar.

Our Method: Step 3

- The property/value pairs are accumulated, and are filtered.
- A similarity metric is used to find a “hook” in the ontology for the candidate.
- The candidate is appended to the ontology, and a one-to-one lexical entry is created.

Our Method: Assumptions, and Future Work

- Candidates are children, not parents or siblings.
- Mostly working with nominal word senses now, however: 1) verbal senses are being tested along with some additional statistical enhancements 2) adjectival senses are being learned in a “knowledge-lean” manner.

The Big Picture

- Spiral method of learning (infinite bootstrapping)
  - Find an unknown word, use the current knowledge to learn what it is
  - Add the unknown word to the current knowledge
  - Repeat

Questions?

"It is no coincidence that in no known language does the phrase 'As pretty as an Airport' appear."

Douglas Adams (1952 - 2001)

"Learning without thought is labor lost; thought without learning is perilous."

Confucius (551 BC - 479 BC), The Confucian Analects