NLP for Medical Domain

Dan Moldovan

Human Language Technology Research Institute

University of Texas at Dallas

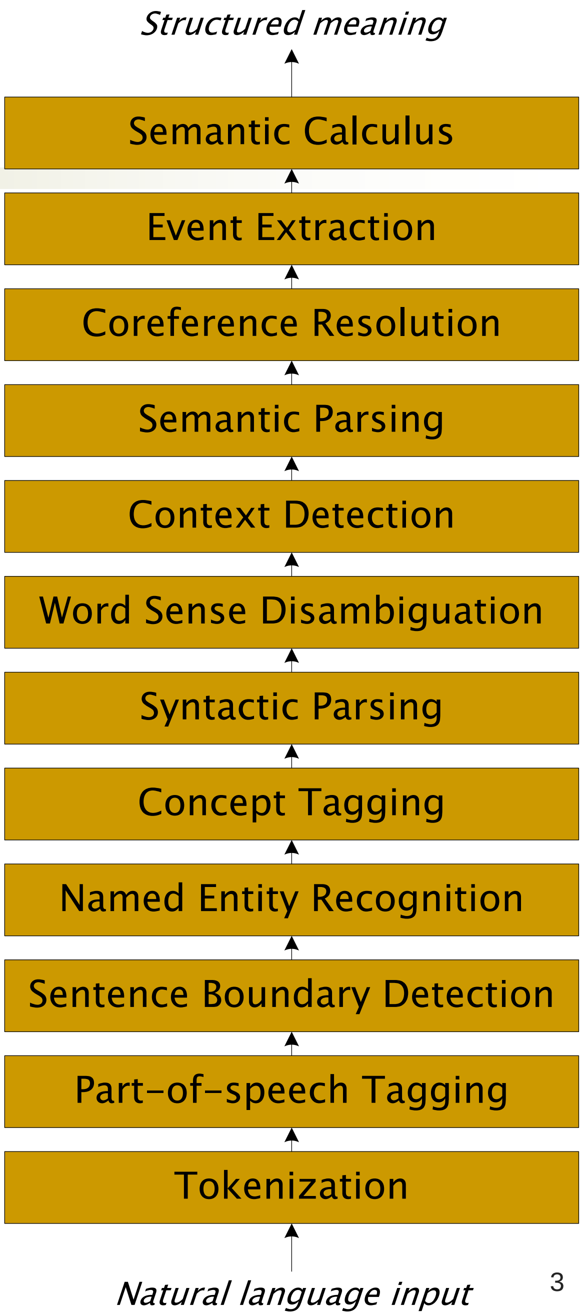
April 2012

Overview

- NLP Tool Suite
- Hierarchical Knowledge Representation
- Semantic Relations
- Semantic Calculus
- Automatic Ontology Building
- Document Similarity, Search
- Statistical vs Semantic Driven Approaches

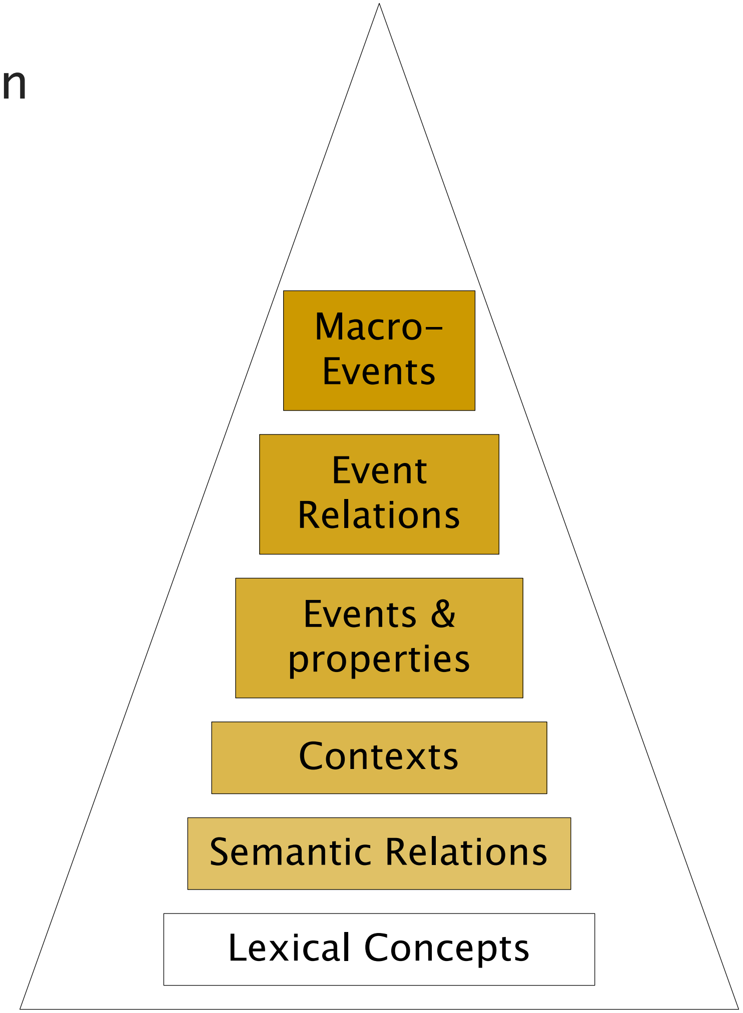
Suite of NLP Tools

- Statistical / rule-based approaches for low level modules
- Semantic-intense methods for high level applications
- Resources: WordNet, eXtended WordNet, EventNet, Lexical Chains, Domain Ontologies
- Powerful applications can be built on top of such a pipeline



Hierarchical Knowledge Representation

- Allows for various degrees of abstraction
  - Macro-events: discourse understanding
  - Event level: events + properties + event relations
  - Contextual level: concepts + semantic relations in contexts
  - Semantic level: concepts + semantic relations
  - Syntactic level: concepts + grammatical dependencies
    - Full syntactic parse tree
  - Lexical level: concepts only
    - Named entities, collocations, normalized temporal expressions



Macro-Events

Event Relations

Events & properties

Contexts

Semantic Relations

Lexical Concepts

Hierarchical Knowledge Representation

Example

The patient’s eye pain was associated with the surgical procedure and poly-L-lactic acid

Event relations: CAUSE(procedure,pain), BEFORE(procedure,pain)

Events: _state_event(pain), EXPERIENCER(patient,pain), _occurrence_event(procedure)

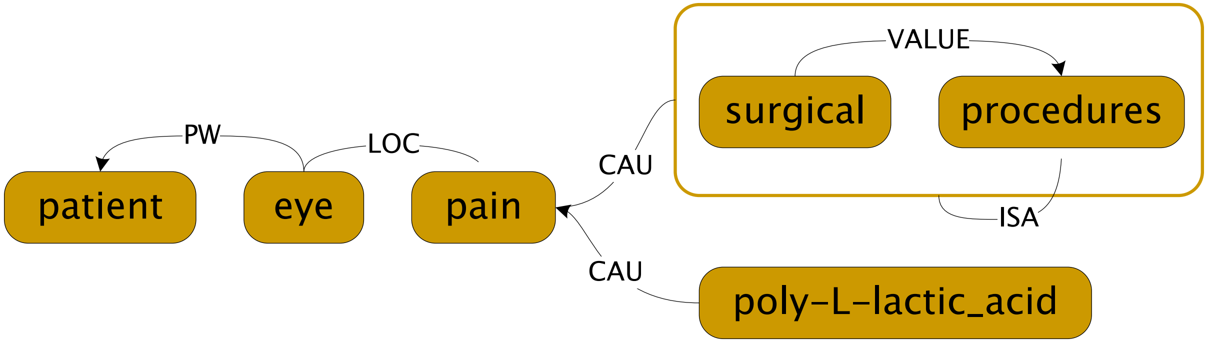
Semantic relations: PART-WHOLE(eye,patient), LOCATION(eye,pain), EXPERIENCER(patient,pain), CAUSE(surgical procedure,pain), ISA(surgical procedure, procedure), VALUE(surgical,procedure), CAUSE(poly-L-lactic acid,pain)

Concepts: patient/NN/1, eye/NN/1, pain/NN/1, associate/VBN/1, surgical/JJ/1, procedure/NN/1, poly-L-lactic_acid/NN/1, _chemical_element(poly-L-lactic_acid)

Semantic Parser

- Fixed set of 26 semantic relations
  - Verb-argument relations: AGENT, THEME, RECIPIENT, MANNER, INSTRUMENT, LOCATION, TIME, etc.
  - Complex nominal relations: ISA, PART-WHOLE, POSSESSION, KINSHIP, SYNONYMY, QUANTIFICATION, PROPERTY, etc.

The patient's eye pain was associated with the surgical procedure and poly-L-lactic acid



Semantic Parser

- Dan Moldovan and Eduardo Blanco, LREC-2012
- Various syntactic patterns: verb-argument, complex nominals, genitives, adjectival phrases/clauses, etc.
- Semantic restrictions on relation arguments
  - Domain and range restrictions defined using an ontology of sorts

    Ex: KINSHIP: [AnimateConcreteObject] → [AnimateConcreteObject]
  - Filter relations that cannot exist between certain arguments
- Performance
  - Precision: 0.321
  - Recall: 0.731

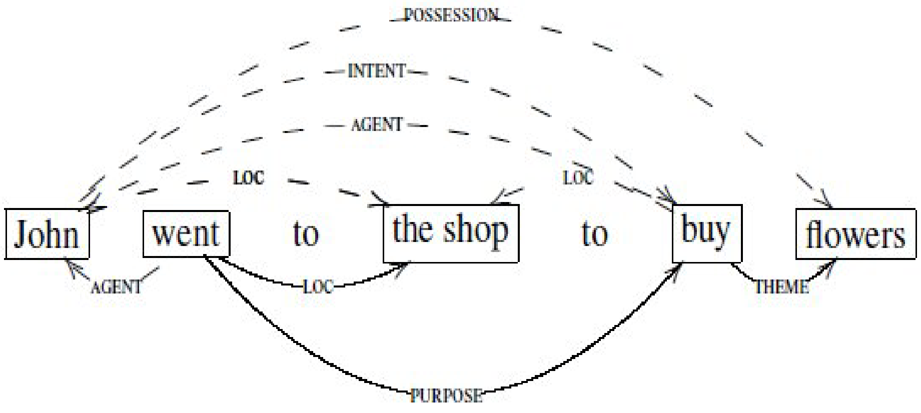
Semantic Parser

- Bracketer – determine semantic dependencies between compound nouns with three or more nouns
  - Sugar industry analyst vs. Female industry analyst
- Argument detection – identify argument pairs likely to encode a semantic relation based on lexico-syntactic patterns
- Domain and range filtering – filter candidate arguments based on their semantic classes and relation definitions
- Feature extraction – extract features corresponding to each pattern
  - Lexical, syntactic path, voice, semantic classes, etc.
- Machine learning classifiers – per-relation and per-pattern approaches
  - Support vector machines, Decision trees, Naïve Bayes, Semantic Scattering
- Conflict resolution – resolve relation conflicts between classifiers

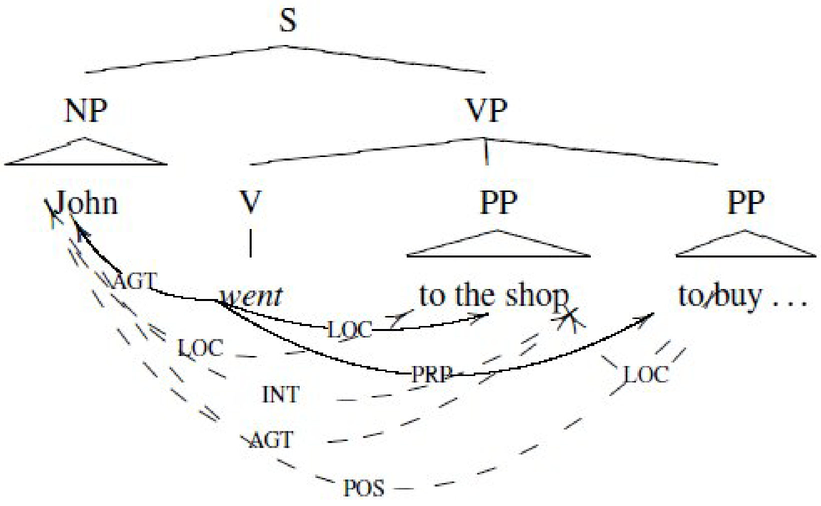
Semantic Calculus

Idea: More semantic relations can be extracted from text in addition to what Semantic Role Labelers or Semantic Parsers can provide

Moldovan and Blanco LREC 2012



(a) Semantic representation.



(b) Parse tree.

Semantic Calculus

- Axioms Obtained Automatically

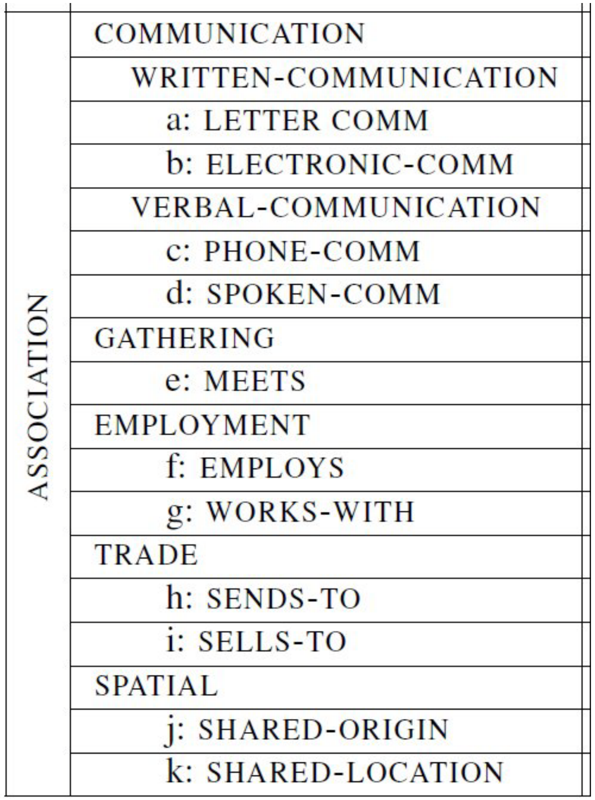
  Using 26 generic relations, we obtained 216 unique inference axioms that do not require constraints and infer relations from the set of 26

  - Relation inventory: CAUSE, JUSTIFICATION, INFLUENCE, INTENT, PURPOSE, VALUE, SOURCE, AGENT, EXPERIENCER, INSTRUMENT, THEME, TOPIC, STIMULUS, ASSOCIATION, KINSHIP, IS-A, PART-WHOLE, MAKE, POSSESSION, MANNER, RECIPIENT, SYNONYMY, LOCATION, TIME, PROPERTY, and QUANTIFICATION
  - Examples of axioms
    - AGT(x,y) o PRP(y,z) → AGT(x,z)
    - CAU(x,y) o JST(y,z) → JST(x,z)
    - AGT(x,y) o ISA(y,z) → AGT(x,z)
    - AGT(x,y) & CAU(y,z) → AGT(x,z)
  - Evaluation
    - We sampled eight axioms and retrieved the first 1,000 instantiations from Treebank
      - They add 36.55% more relations
      - Average accuracy is 0.86, similar or even superior than the performance of semantic parsers

Semantic Calculus

Can help with extracting new high lever relations from text

- Case study: associations between people



ASSOCIATION	COMMUNICATION
	WRITTEN-COMMUNICATION
	a: LETTER COMM
	b: ELECTRONIC-COMM
	VERBAL-COMMUNICATION
	c: PHONE-COMM
	d: SPOKEN-COMM
	GATHERING
	e: MEETS
	EMPLOYMENT
	f: EMPLOYS
	g: WORKS-WITH
	TRADE
	h: SENDS-TO
	i: SELLS-TO
	SPATIAL
	j: SHARED-ORIGIN
	k: SHARED-LOCATION

Semantic Calculus

- Example axiom:

  $AGT(x,y) \circ THM^{-1}(y,z) \circ RCP^{-1}(z,w)$ & [y is *write/produce, z* is *writing*]

  → COMMUNICATE(x,w) → ASSOCIATED_WITH(x,w)

  - Axiom has three premises: AGENT, THEME and RECIPIENT and restrictions on arguments *y* and *z*
  - Interpretation: *People who write to others , communicate with the recipients and furthermore are associated with the recipients.*
  - Instantiation:

    $AGT(Bob, wrote) \circ THM^{-1}(wrote, letter) \circ RCP^{-1}(letter, Mary)$

    → COMMUNICATE(Bob,Mary) → ASSOCIATED(Bob, Mary)

Automatic Ontology Creation ?

- A major bottleneck today is :
  - Manual creation of ontologies
  - Yet, ontologies are very useful

- Having transformed text into semantic triples, it becomes possible to automatically create domain ontologies.

Automatic Ontology Generation

Automatic ontology builder

- Input: document collection; (optional) seed concepts, or existing ontologies

1. Process documents using the NLP pipeline
   - Identify large set of semantic relations within document content
2. Start with some seed concepts and extract related concepts from text
   - Use ISA, PART-WHOLE, CAUSE, PURPOSE, ATTRIBUTE and other semantic relations between concepts identified in text
3. Classify derived concepts
   - Use structural subsumption

- Output: semantically-rich domain ontology
  - Human edits ontology generated by machine

Document Similarity

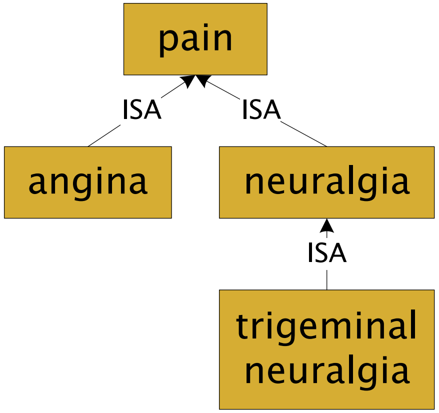
- Possible applications in medical domain
  - For diagnosis – patient data vs medical knowledge
  - For research – text snippet vs Medline
  - Match decision rules to KB
  - Others
- Approaches
  - Statistical approaches: Latent Dirichlet Allocation, Pachinko Allocation, others
  - Semantic approaches:
    - Event based
    - Ontology based – outlined here
  - Others

Sample Search

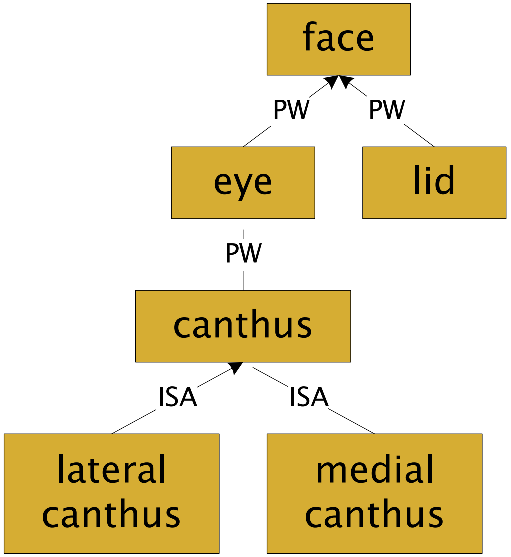
- Search: The patient’s eye pain was associated with the surgical procedure and poly-L-lactic acid
- Result: She describes this area as looking like a "bug bite" & was located "on top of" (above) gortex implant, near the lateral canthus. Its shape is round about one-fourth inch in diameter w/a rise w/a peak "maybe" one-eighth of an inch in height total. She said her phys has treated the "bug bite" area w/an unknown type of steroid injection, w/o effect. He now wants to remove this surgically, however, she is not certain if she wants this done. She noted that she did not massage for first week, as had no instruction to do so; she also had lid lift surgery at the time (of the face lift,) & surgeon did not want any pressure on surgical site. She reported her concomitant medications as estradiol, gabapentin (neurontin), for trigeminal neuralgia & facial non-specific neuralgia; also a multivitamin. Add'l medical history included trigeminal neuralgia & facial non-specific neuralgia both following the accident. No further medical info reported. Add'l info for sculptra from ptc report case (b)(4) dated (b)(6)2008, received by (b)(6) on 25mar08: b/c no lot # is available, an investigation has been performed on the documentation of all potentially involved manufactured batches. The review of the device history reports & of the analytical results of these batches did not show any anomaly that could be related to the event which occurred.
- Repository: Manufacturer and User Facility Device Experience (MAUDE)

Sample Search – Supporting Ontologies

- Medical Subject Headings (MeSH) controlled vocabulary



- Encyclopedic knowledge



Statistical vs Semantic Driven Approaches

Statistical methods

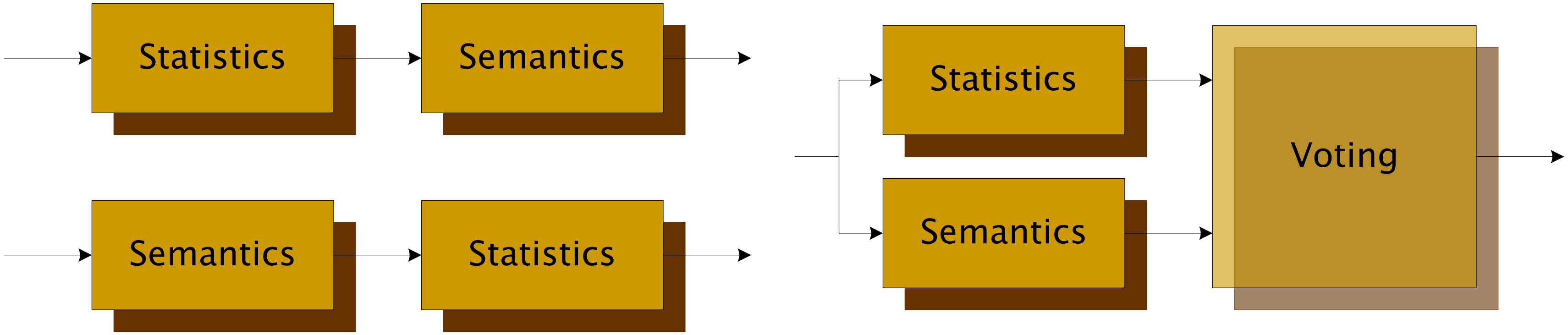
- Pros
  - Handle larger volumes of data
  - More robust to noise
- Cons
  - Course grained decisions
  - Make more errors
  - Require a lot of training data

Semantic driven approaches

- Pros
  - Finer grained, higher precision results
  - Leverages machine learning with knowledge driven features (ontologies, lexicons, semantic relations)
  - Algorithms and ML results tuned with knowledge sources and rules that can be learned (semantic calculus)
  - Easier to customize
- Cons
  - Knowledge sources can be labor intensive to build
  - More computationally intensive

Statistical vs Semantic Driven Approaches

- The two approaches can be married together in several ways
  1. Most common way is to intertwine the two approaches at module level
  2. Use the statistical methods to filter results and then do higher precision analysis with the knowledge driven approaches
  3. Use knowledge driven results as features to the statistical algorithms
  4. Use both approaches and perform voting at the end


Statistics
Semantics
Semantics
Statistics
Statistics
Semantics
Voting