Using Distributed NLP to Bootstrap Semantic Representations from Web Resources

Núria Bertomeu
Harry Halpin
Yi Zhang
Motivation

- Large domain ontologies are being created in the Web.
- Bootstraping the content of Web text and binding it to existing ontologies allows for complex inference tasks on Web content.
- Previous approaches to this task only have used shallow methods such as NER (Popov, 2003) and n-gram tagging (Guha, 2003)
- Is it possible to weigh the logical results using probabilities of extraction from documents?
OWL and Description Logics (DL)

- The OWL (Web Ontology Language) is designed as common syntax and semantics for ontology classification of Web pages and their content.
- SHOIN(D) DL is the formal foundation of OWL DL.
- OWL uses URIs to disambiguate and store ontologies.
- OWL uses an open world assumption.
Description Logics (DL)

• DL supports inference patterns such as classification and satisfiability checks.
• Decidability of inference procedures.
• Basic syntax:
  • atomic concepts (unary predicates),
  • atomic roles (binary predicates),
  • individuals (constants).
• Small set of constructors for building complex concepts and roles: union, full existential quantification, number restrictions, negation.
Description Logic vs. First Order Logic

- DL is a subset of First Order Logic (FOL).
- FOL is more suitable for the representation of NL statements.
- However, inference procedures with FOL tend to be undecidable.
- Correspondences between DL and FOL:
  - atomic concepts → unary predicates;
  - atomic roles → binary predicates;
  - intersection → conjunction;
  - union → disjunction;
  - negation → negation;
Semantic Representations

- MRS – a framework for flat semantic representation that allows for underspecification.
  - predicate calculus with generalized quantifiers;
  - an MRS structure is a tuple $\langle GT, LT, R, C \rangle$, where $GT$ is the global top handle; $LT$, the local top; $R$, elementary predicates and their arguments, and $C$, constraints on handles.
  - rMRSs (Copestake et al., 2003) obtained from the HoG (Callemeier et al., 2003).
Semantic Representations

- **DRS** - semantic representations within the framework of DRT.
  - a notational variant of standard predicate logic.
  - a DRS is a pair \( <U, C> \), where \( U \), the universe is a set of variables, and \( C \), a set of conditions. Simple Conditions are predicates applied to variables. Complex conditions are recursive and include quantifiers, negation, and conditionals.
  - DRSs (Kamp, 1980) obtained from a CCG Parser (Bos et al., 2004).
Representing Natural Language in DL

- Predicates:
  - each predicate belongs to a class and instantiates an individual of that class.
  - basic relations: AGENT, THEME, PATIENT, RECIPIENT, MODIFEE.
    - Domain: event individual;
    - Range: individual;
Algorithm (RMRS -> DL)

• General Algorithm
  • Given RMRS \(<T,L,R,G,C>\)
    • For each \(ep\) in \(L\)
      - \(h:pred(a) \Rightarrow a:pred\)
      - \(map(h) := a\)
    • For each \(rarg\) in \(R\)
      - \(role(h, c) \Rightarrow (map(h), c):role\)
      - \(role(h, v) \Rightarrow (map(h), v):role\)
    • For each variable equality \(u=v\) constraint in \(C\)
      - Merge the individuals created for \(u\) and \(v\) to a single individual

• Language Specific Tuning (English)
  • Don’t create individual or class for general predications like \(message\_rel\)
  • Ignore all the generalized quantifiers (See future work)
  • For preposition, create individual of class \(Event\)
  • For pronoun, create individual of class \(Thing\)
  • For \(named\_rel\), crate individual of \(Named\_Thing\)
Algorithm (DRS -> DL)

- Given a DRS
  - Flattening the DRS to a list of unary or binary predicates.
  - For each unary predicate:
    - pred(\(v\)) \(\Rightarrow\) \(v:pred\)
  - For each binary predicate:
    - role(\(u, v\)) \(\Rightarrow\) \((u, v):role\)
  - For each equality condition \(u=v\):
    - merge the two individuals created for \(u\) and \(v\) to a single individual
  - Implies condition \(X \rightarrow Y\) and disjunction \(X \lor Y\) are difficult, but can be interpreted as union and inheritance relationships. Unclear if this preserves anything resembling semantics.
Example

- Pat argues with Tim.
Finished Work

- Automatically collected and semantically annotated 100 web-pages from an „Eiffel“ Google search.
- Manually classified for ontology disambiguation experiments.
- Processed real scuba diving accident reports from a Web database with an ontology and inference test suite from IBM.
- DRS WS: http://axon.inf.ed.ac.uk/ws/sem2owldrs
- CCG2SEM works on real web text!
- HoG (MRS) processes text, but has more technical problems. No web service, considering it.
- Probabilities stored as a RDF tag per “word” with corpus frequency
Future research

- Disambiguation experiments with Eiffel corpus.
- What exactly do probabilistic weights “mean” once the concepts have been extracted?
- Better handling of quantification and plurals.
- Need a large scale SemWeb gold standard corpus which has an ontology and inference test suite that stretches to the level of natural language semantics.
- Need a set of tasks (like Q&A) on this corpus.
- Evaluation against n-gram and NER
- Standard transformations from NL DL to more standard OWL triples.
- Formalize concepts of „minimizing power lost“ in logic.