Interpreting the meaning of text and discourse

Tatu Ylönen

University of Helsinki

October 31, 2019

Contents

My research objective

- Represent meaning of text and discourse using an interpretable representation that can be integrated with knowledge bases and symbolic control systems
- Develop a parser that interprets text in a discourse situation into that representation
- Develop a generator from that representation into text given a discourse situation and an objective
- Produce a collection of background knowledge that helps in these
- Develop discourse and pragmatics oriented evaluation metrics
- Focus on English for now (availability of resources) with an eye on Finnish

Language as communication

- I firmly think language should be viewed as a communicative process that aims to instill beliefs, emotions, and desires in the recipient
  - I guess this puts me in the functional linguistic camp (Dik, Van Valin, Halliday etc – though I’ve not yet read that much of their work)
- I view clauses and sentences as invoking and constructing mental representations of situations, not as truth-functional
  - Where do I find the people who think the same way about semantics?
- Linguistic expressions are interpreted in context - their semantics are thus unlikely to be compositional
  - In contrast, much prior work is on compositional semantics based on logic.
- Semantics and pragmatics are largely inseparable
  - Why would I care about the literal meaning of sentences?

Classical knowledge representation ideas

- Code (i.e., knowledge is implicit in the program)
- Rules (e.g., expert systems 1970s-1980s)
- Frames (Minsky 1974)
- Scripts, plans, goals (Schank 1977)
- Ontologies (formal system of concepts and their relationships, often for knowledge sharing)
- Conceptual Dependency theory (Schank 1975)
- Semantic web (recent work: Helbig 2006)
- Predicate logic, Lambda calculus, Logics of knowledge and belief, other modal logics
- Description Logics (easier inference with adequate expressive power)
- Defeasible logics (tolerant to falsehoods, conflicts, etc) and other non-monotonic logics
- Uncertainty handling: Dempster-Shafer theory, Bayesian inference, Naive Bayes, probabilistic networks, ad hoc weights
- Model-theoretic semantics (Tarski?), closed world assumption (Reiter 1978), circumscription (McCarthy 1980)

Neural/numeric knowledge representations

- Connectionism (Rumelhart 1986, though Perceptrons 1950s and similar cognitive models much earlier), Backpropagation learning (Rumelhart et al 1986)
  - Word embeddings, sentence embeddings, document embeddings, event embeddings
  - Knowledge encoded as weights of a neural network and input/output tensors
- Dimensionality reduction (PCA, ICA, etc)
  - Self-organizing map (Kohonen 1982): non-linear dimensionality reduction
- Interpretable sequence representations (e.g., Gröndahl&Asokan 2019)
- Conceptual spaces (Gärdefors 2000): high-dim feature vectors
- Latent semantic analysis (LSA)

A side note on compressed representations

- We can uniquely encode any knowledge structure with a 128 bit random value (or hash). Per the birthday problem, 128 bit random vectors encode 10^16 values with 10^-6 probability of a conflict (100M objects/sec for 100 years). 192 bits increases this to 10^26 objects with p=10^-6 or 10^23 objects with p=10^-12 (cf. typical hard drive uncorrected read error rate 10^-14 of bits read).
  - However, decoding such random bit vectors takes exponential time.
- Neural representations use 100-10000x more bits, but support similarity by structuring of the multidimensional space. The structure also makes approximate decoding practical and enables inferences using the encoded representation. Space used for the network weights is on the order of 10^10 to 10^14 bits.
- Can we find other constructions with similar compression properties as provided by deep learning? Can similarity hashes or invertible hashes be generalized somehow? Is there a way to understand the approximation errors?

Modern graph-based representations of knowledge

- MultiNet (Helbig 2006)
- Abstract Meaning Representation (AMR) (Banarescu et al 2013)
- Lexical semantic databases (WordNet, FrameNet, VerbNet, PropBank)
- Some relevant corpora (Groeningen Meaning Bank, OntoNotes, FrameNet)
- Discourse annotated corpora (Penn Discourse Treebank, Discourse graphbank, RST Discourse graphbank)

Abstract Meaning Representation (AMR) (Banarescu 2013)

- Fairly recent, but in my opinion at least 10 years behind Helbig
- Various curious design choices, for example no distinction between concepts and individuals
- Uses OntoNotes word senses for concepts and many relations
- Several automated parsers and annotated data sets available
- Active research area currently

MultiNet (Helbig 2006)

- Semantic network, with high-level ontology and over 100 relation types and dozens of functions defined
- Rich representation of nodes with information on multiple layers: intensional level, pre-extensional level
- Clear distinction of types of entities: concepts, individuals, groups, groups of groups, etc
- Defines top-level ontology but otherwise no definition of concepts (they had Hagenlex dictionary with defined them for their use)
- In my view the best available fully thought out semantic network framework, has not gotten very much attention

Semantics vs. pragmatics

- Traditionally, semantics has been about literal meaning of a sentence.
- Pragmatics studies how context contributes to meaning.
  - Meaning, with focus on what the speaker's intentions and beliefs are
  - Meaning in context (time, place, participants, prior discussion)
  - What is implied, even if not overtly expressed
  - Influence of social and physical distance between speaker and recipient
  - Information structure
- Information structure
  - What is being talked about, and what new information is conveyed about it (topic, focus)
  - What entities are new to the discussion (given vs. new)

Discourse concepts

- Cohesion – grammatical means in which sentences and paragraphs are linked
- Coherence – links in meaning between sentences; consistency in ideas
- Structure – document, chapter, section, paragraph, sentence, clause
- Relations – elaboration, concession, contrast, etc
- Topic, focus
- Information structure – given, new
- Foreground, background information
- Types of information – events, participants, setting, explanation, evaluation, discourse irrealis, performative

Discourse representation formalisms

- Discourse representation theory (DRT) (Kamp 1981)
  - Primary deals with reference resolution
  - Discourse representation structure (DRS): entities under discussion (referents), information known about those entities (conditions)
- Rethorical structure theory (RST) (Mann & Thompson 1988)
  - Expresses relations between parts of the text
  - Nuclei – most important parts of text
  - Satellites - additional information about the nucleus
- Situation semantics

Potentially questionable common assumptions/practices in computational linguistics

- Model-theoretic semantics, compositionality
- Processing isolated sentences
- Semantics as independent from pragmatics
- Syntax as independent from semantics/pragmatics
- Interpretation of sentences outside their context
- Staged pipeline for processing – greedy decisions at each level
- ”x BLEU better in narrowly defined problem Y”
- Insufficient treatment of layered rule systems and conventions in punctuation, capitalization, compounds, hyphenation, commas, italicization, even morphology and syntax (prescribed correct word forms and expressions for writing)