

Fall 2011 LTI Colloquium Report

Beyond Shallow Semantics

Ni Lao

This is a response to Eduard Hovy's talk "Toward Merging Propositional and Distributional Semantics" and Martha Palmer's talk "Beyond Shallow Semantics".

Any **theory of semantics** should either help us understand how human brains work or build systems that can understand. *Propositional semantic theories* provide an important idea that there should exist an entity correspond to each phrase or sentence. It is not only biologically plausible, but also serves as a potential way to construct systems to deal with semantics. As for the representation of such entities, symbolic systems are attractive, because they have the cleanness and beauty of mathematically systems. However, they are inadequate in representing the **fuzziness** of meaning, and have no obvious explanations for psychological phenomena such as confusions, forgetting, and the degree of processing complexity, etc.

A new trend in today's NLP research is distributional 'semantics', which represents a concept as the distribution of words that appears together with the concept. However, it is not **compositional**—cannot combine two concepts to form a new concept such as negations or modalities. Edward described one attempt to combine propositional and distributional semantics (Penas & Hovy, 2010; Hovy et al., 2011). A concept is a list of triples $C = \{(r_1 w_1 s_1) (r_2 w_2 s_2) \dots (r_n w_n s_n)\}$ where r_i is a relation, w_i is a word, and s_i is a weight. Therefore, each concept can be represented as a set of word distributions one for each relation.

Martha mainly talked about using syntactic and shallow semantic features for WSD tasks (Chen et al., 2007, Brown et al., 2011). I would say that Edward and Martha's approaches are limited in two senses

1) The representations are specific to text. It is true that most semantic parsing systems are dealing with text, and word is a good median for communication. However, using word as a representation of semantics and inference is not ideal. First, words are just one type of observations that a system can make. For example we might want to develop a system which only has visual and sensual information. Second, word is not a deep enough representation for semantics. They are ambiguous, and by themselves do not easily represent background knowledge of a domain.

Current semantic parsers produce shallow structures such as Syntactic Structure (e.g. parse trees), Semantic types (e.g. person, location, organization), Semantic roles (e.g. agents, instrument), Sense distinctions – (e.g. WordNet, OntoNotes groups), and coreference. However, the applications down the stream—such as information extraction, question answering, and machine translation -- need deeper semantics to bring out implicit knowledge. For example it might be useful to know the class a noun in the sentence belongs to (e.g. Young is a baseball player, and a receiver); it might also be useful to know the indication of an action (e.g. stabbing is a way of hurting); it might also be useful to know possible relations between concepts (e.g. to parse the phrase “Young touchdown pass”).

Entity relation graph is a common way to represent knowledge bases and also a common way to represent shallow semantics parsed from sentences. Therefore, a natural way to represent deep semantics is to jointly represent information extracted from sentences and a knowledge base as a graph. Once the nodes in a sentence are connected to a knowledge base, the depth of semantics is basically unbounded---as far as one traverses into the knowledge base through typed edges. How to find the path that leads to useful signal is not a problem of semantic representation but rather a problem of learning.

Furthermore, Martha’s approach assumes gold standard semantic labels (e.g. VerbNet) which are not desirable in practice. The internal representation of a system does not necessarily have anything to do with the system’s performance---e.g. as long as a robot goes and fetches a cup of tea, one should not care less about how it internally represents the task of “fetching a cup of tea”. Not only does a gold standard representation unnecessary, it might also be unattainable. When creating semantic labels, even human judges have disagreements. In general, it should be OK to allow each agent to have its own interpretation of what it observes, as long as it can perform tasks correctly.

2) Defining a suitable semantic representation is not enough--a practical theory should also cover how the structures are generated and applied. For feature generation, Eduard uses statistics of dependency tree fragments, while Martha uses several existing shallow parsers plus syntactic patterns. Both approaches are easy to implement. Whereas for application, Eduard doesn’t have any yet, while Martha uses WSD accuracy to evaluate the enriched semantic tags. Overall, I find that the focus on semantic representations miss some real challenge of the research on semantics—how do we know a particular representation such as certain background knowledge is useful? How do we link certain observations to such representations? These questions are intrinsic to any use of semantics, therefore should be part of any semantic theory or algorithm. A

semantic theory or algorithm should be end-to-end --- starting from observations (e.g. sentences) to the task (e.g. classification or control). From this point of view Socher et al.(2011)'s work "Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions" is very good. The algorithm is end-to-end --- from observations (sentences) to actions (sentiment analysis). However, the semantic representation is too simple--- just a state vector. We can imagine an entity-relation representation would better support inference tasks.

Proposed new theory:

Now we put together the two elements discussed above, and come up with a new plan:

- 1) Represent semantics as an entity relation graph which contains a knowledge base
- 2) Develop a theory which includes how we get from observations to semantics and from semantics to performing tasks.

For example, when dealing with syntactical parsing task we can imagine a type of shift-reduce parser where the states are not CFG categories but snapshots of a knowledge base. At each step of shift-reduce action, the knowledge base can be modified by adding or removing nodes or edges. The task is to predict whether the shift or reduce action should be taken at each step.

For a more general theory which subsumes non-textual observations and tasks we can formulate it as a control problem. At each step an agent can receive certain observations (e.g. seeing a word, or feeling cold), and also take several actions as long as these actions do not conflict with each other. Example actions can be looking at the next word, or making a certain change to its internal representation. The objective of the agent is to predict future observations or performing correct action. How the representations are modified based on past history of observations, and how predictions are made based on the representation should also be part of the theory.