Integrated Research Plan

In the field of computational linguistics, solving many problems requires transforming natural language text into a representation that is closer to how machines interpret input. Syntactic representations like constituent parse trees and dependency parse trees have been successfully used for solving many problems in this field for a long time now. But an equivalent sentence level semantic representation has been less explored. In recent few years, the field is seeing an increasing need for more semantic level understanding of text for solving more challenging problems. Currently most of the approaches to learning any kind of representation lay their foundation on techniques that need huge amounts of annotated data to work well. However obtaining such huge amounts of semantically annotated data is difficult and costly since it requires a linguistic expert in the loop. But humans do not need such huge amounts of data to be able to efficiently parse the meaning of a sentence. They can generalize much better. So how do they do that? There has been lot of work in the linguistic field that tries to answer this question. But the field of computational linguistics has not fully made use of this research. In this work I make an attempt to bridge this gap. In particular I am interested in developing linguistically motivated methods to automatically parse a sentence into one kind of semantic representation called Abstract Meaning Representation (Banarescu et al., 2013) (AMR). The two areas of linguistics that I wish to borrow ideas from are semantics and pragmatics. Using semantics I hope to understand how humans learn to compose meaning of a sentence using meaning of its parts and thus learn to identify the right relations between the parts. This understanding will be helpful for AMR parsing where one of the key challenges right now is getting the relations right between the concepts. Pragmatics on the other hand will help me understand how humans make use context and common sense knowledge while parsing a sentence. I plan to incorporate these ideas into AMR parsing using the distant supervision paradigm (described later). I believe that my prior experience in AMR parsing puts me in a good starting position to explore the computational aspect of this work. For gaining more insights into the linguistic aspect, I plan to take relevant courses and also collaborate with people working in Semantics and Pragmatics in UMD's strong linguistics group.

What is AMR?

The key motivation behind developing AMR was to have a comprehensive and broad-coverage semantic formalism that puts together the best insights from a variety of semantic annotations (like named entities, co-reference, semantic relations, discourse connectives, temporal entities, etc.) in a way that would enable it to have the same kind of impact that syntactic treebanks (e.g. Penn treebank) have on natural language processing tasks. AMR tries to capture the notion of 'who did what to whom' in a sentence. It abstracts away from syntactic idiosyncrasies yet preserves the meaning of the sentence. It is composed of concepts and relations, not nouns and verbs. In fact the creators go as far as to say that there are no nouns, verbs, adjectives, adverbs, affixes or zero pronouns in AMR. This allows AMR to have the same representation for a huge number of variations of a sentence (see fig 1)

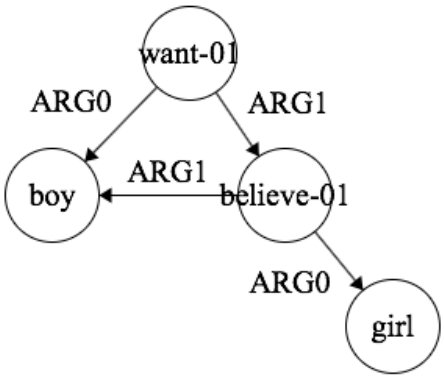
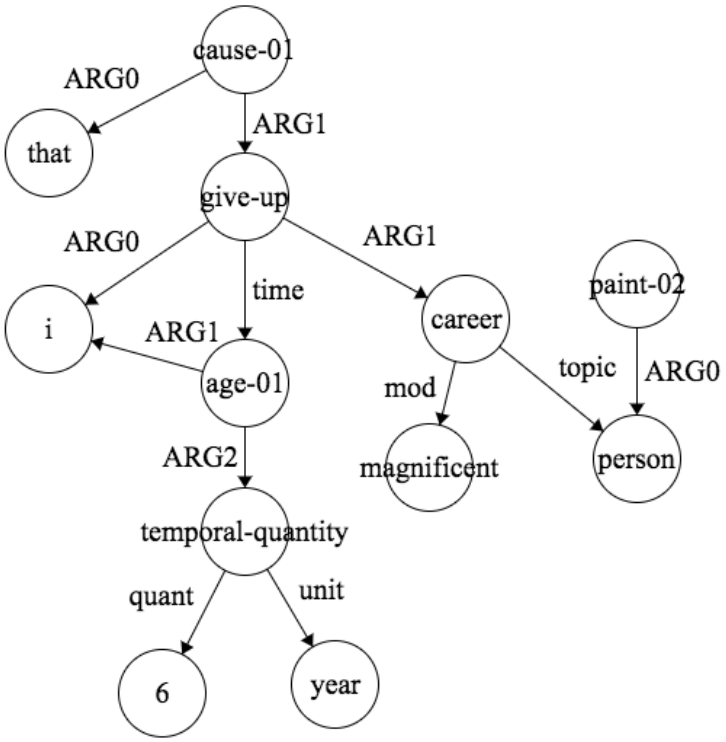


Figure 1: AMR above can be expressed variously in English as:
The boy wants the girl to believe him.
The boy wants to be believed by the girl.
The boy has a desire to be believed by the girl.
The boy's desire is for the girl to believe him.
The boy is desirous of the girl believing him.

Using Semantics

Currently one of the key challenges that we are facing while developing computational models for parsing text into AMR is identifying the right relations between concepts, especially the relations that are closer to the root of the tree. For e.g. below is a sentence and its AMR.

"That is why, at the age of six, I gave up what might have been a magnificent career as a painter."



For a computational model, it is more difficult to identify the relation "ARG1" between the concepts "give-up" and "career" than the relation "ARG0" between "give-up" and "i" because the former is a long-range dependency. Humans are faced with the same kind of challenge when trying to parse the meaning of a complex sentence, but they are good at it. Humans are able to compute the meaning of a sentence from the meaning of their parts. The way they learn to generalize is as Donald Davidson puts it:

> "The work of the theory is in relating the known truth conditions of each sentence to those aspects ("words") of the sentence that recur in other sentences, and can be assigned identical roles in other sentences."

In the field of linguistics, there has been a lot of work in developing a theory of meaning composition. According to Frege's Conjecture, semantic composition is a functional application on the parts of the sentence. In this work, I plan to borrow some of these principles and incorporate them into our computational models for improving AMR parsing.

Distant supervision and the use Pragmatics

Another key challenge in AMR parsing is parsing sentences that contain concepts or relations (between concepts) that were not seen in the training data used to train the parser. Since annotating sentences with their semantic representation is a hard task (requiring humans with linguistic expertise), we may not be able to get huge amounts of annotated data to train a parser. Even though annotating sentences is a hard task, there is a lot of structured data available out there in the form of knowledge bases like OpenCyc, DBPedia, Freebase, etc In this work I plan to use some of these to automatically get annotated sentences that we can train our AMR parser on. This technique of using external knowledge resources to get annotated data is called 'distant supervision' in computational land and has been successfully used for getting state-of-the-art results for relation extraction task. Since identifying relations is a key component in AMR parsing, this approach of distant supervision seems promising for AMR parsing as well. Another obvious justification for such an approach is the fact that humans definitely make use of their common sense knowledge a lot when trying to parse a sentence for its meaning. Linguists study this very phenomenon under the area of Pragmatics. Hence, in this work, I plan to first read through some of the existing work in Pragmatics that looks into how humans achieve this and try to apply them to AMR parsing.

References

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking.