Deep Language Understanding

James F. Allen

Dept. Computer Science, University of Rochester

Abstract

Deep language understanding involves mapping language to its intended meaning in context, using concepts and relations in an ontology that supports knowledge and reasoning. Currently, one would think there was consensus across the field of computational linguistics that deep understanding is not possible. As a result, current natural language work is mostly divorced from work in reasoning, planning and acting. In this talk I will argue that, contrary to current thought, an effective level of deep understanding is a very viable research area. I will present examples of recent work to support these claims. Interestingly, while I argue that the current statistical paradigm is unlikely to achieve deep understanding, it is also the case that deep understanding will likely only be possible by exploiting the advances in statistical approaches.

Myths about Deep Language Understanding

While most people would acknowledge that language cannot be understood without significant common-sense knowledge and reasoning processes, deep understanding research is quite unpopular these days. The field of natural language processing (NLP) is almost entirely focused on so-called shallow methods, mostly involving machine learning from corpora. This transformation has energized the field and brought significant advances and provided some principled techniques for dealing with the critical problem of resolving ambiguity in language. It is hard to argue about such successes. But in the process NLP ended up divorced from knowledge representation and reasoning, to the extent that the two are now considered virtually unrelated fields with no connection with each other.

This is very unfortunate, and results in the fields narrowing the set of problems that they will study. In NLP these days, the field mostly studies two types of problems: problems for which there are substantial labelled corpora, and problems that can be solved by machine learning techniques over raw text. There is an aversion to hand-engineered systems that require some human effort to encode the necessary knowledge for deep understanding. Human effort in producing corpora is still allowed, but limits the problems to those for which we can devise simple enough coding schemes to attain interannotator reliability. Thus, for instance, work on intention recognition from language is now almost unstudied, and what work there is limited to one small subproblem: identifying the core discourse act for each sentence out of a dozen or so predefined categories.

The attitude of the field towards deep understanding can be summarized in two myths, which are repeated whenever one wants to justify using shallow methods.

> **Myth #1**: Deep understanding systems will always be narrow and brittle, only shallow methods can produce robust performance.
>
> **Myth #2**: Deep understanding requires solving the complete AI problem, so is impractical.

People happily state these myths and move on to the shallow techniques, thinking the issue is resolved. Its a classic bait-and-switch argument. Since we have had great difficulty building deep understanding systems with robust semantic and contextual interpretation, we need a different approach. Statistical methods provide robust performance and utilize learning algorithms that can improve over time as more data becomes available. Thus, we should all work on statistical approaches.

All the claims above are true except for the conclusion. The catch is the field now can only work on tasks that are amenable to shallow methods. We started with problems involving interpretation of language in context, facing problems with identifying intended meanings underlying language, but we ended up tagging part of speech and building syntactic parse trees. Furthermore, problems that are so complex that we can't build an annotated corpus for are not considered worthy of study.

Some say I'm anti statistics, but this is not true. Statistical models and machine learning play a crucial role in the solution. All I claim is that they are not the entire solution. The field needs to acknowledge and continue to address problems that don't fit the statistical paradigm well, and that require knowledge-based reasoning to solve.

Dialogue Agents and Intention Recognition

The particular deep understanding problems I am most interested in are practical dialogue agents. These are systems that can interact in unconstrained natural conversation with humans to work with them to solve practical tasks. These problems require significant capabilities for knowledge representation and reasoning, and the key focus of language understanding is intention recognition. I would claim that human-level dialogue is primarily driven by intention recognition and only secondarily by structural properties of language that are the focus of statistical approaches.

As one simple example, consider an exchange that occurred at an information booth in the main train station

in Toronto. A passenger approaches the information booth and the following ensued:

Passenger: *The 4:50 to Windsor?*
Clerk: *Gate 7. Its running late.*

This is a perfectly normal everyday conversation, yet the structure of the sentences reveals little explanation of what just occurred. Only when we start to consider the agents' reasoning processes do we get a general account. The passenger mentioned a train. It is only by consulting the schedule and considering the passengers likely goals (e.g., are they taking a train, meeting a train?) can we explain why the clerks answer is coherent and meaningful.

Of course, if one had a large corpora of such conversations, one might be able to duplicate such interactions - though it would be complicated given that the appropriate answer depends on the current situation when the question was asked, not how it was answered in the past. But even getting around such problems, such a solution wouldn't give us a general theory for other interactions, such as this one in a grocery store:

Customer: *Black beans?*
Clerk: *Aisle five, on your left.*

or at the lunch counter

Customer: *Black beans?*
Server: *Do you want bread with that?*

or at home, when you are walking in with groceries

My wife: *Black beans?*
Me: *Ah sorry, I forgot them.*

It's fairly clear what is happening in these examples. Depending on the situation, the hearer identifies the most likely intention underlying the speakers utterance of *black beans*, and the response addresses that intention. This is common, everyday stuff in language, and I cannot see how it can be addressed with close integration with contextual knowledge and reasoning.

Making progress on deep understanding

To make progress on deep understanding systems, we need to address the concerns underlying in the two myths stated above. Like most good myths, they have an origin in truth, but have reached the wrong conclusion. In this final section, I'll discuss how we are approaching each.

It is true that attempts to hand-build grammars that can produce deep logical forms from natural language, with ambiguities such as word senses, semantic roles and scoping resolved have generally been brittle systems that only work in narrow domains. On the other hand, statistical approaches typically only address a single issue or two (e.g., part of speech tagging, syntactic structure, word sense disambiguation, semantic role labeling) and do not focus on producing a deep logical form suitable as input for reasoning. We have been approaching this by working on hybrid systems, that combines our existing deep parser, the TRIPS parser (Allen et al, 2008), with a range of state-of-the-art statistical methods. The statistical models produce advice about different aspects of the sentence, and TRIPS integrates this information to identify the most likely intended logical form. In addition, to deal with gaps in lexical coverage, TRIPS accesses Wordnet and maps the Wordnet hypernym hierarchy to the TRIPS ontology to produce semantic definitions of each item.

We have been evaluating progress on building a generic deep parsing system in several different ways. In Allen et al (2008), we evaluated the system against a gold standard logical form graphs. We defined precision and recall measures based on graph matching, and found an improvement in F-score from 64% with the standalone TRIPS parser to 81% with the same system running using advice from a part-of-speech tagger, named entity recognizer and statistical parser (all from the Stanford group) plus Wordnet lookup for unknown words. The LF graphs that were evaluated required word sense disambiguation of all words, identification of arguments and semantic roles for verbs, nouns, and modifiers, but not quantifier scoping, which we view as a post-parsing issue.

On a different front, we constructed another system to extract events and temporal information (UzZaman and Allen 2010) using a hybrid system combining the augmented TRIPS parser, as described above, plus post-processing using Markov Logic Networks. Using the Timebank Corpus as a test set (Pustejovsky et al, 2003), our system identified events with an accuracy that was equivalent to the interannotator agreement of human judges. In the Tempeval-2 competition (Pustejovsky & Verhagen, 2010), our system was only one of three efforts that attempted all six tasks, and was the best performing of these systems in four of the six tasks (Uzzaman and Allen, 2010b). What is noteworthy here is that our system was the only one that performed all six tasks solely from the raw text. All other systems used annotated data to perform the tasks, except for the tasks of event and temporal expression identification.

Regarding the second myth, that deep understanding requires solving the complete AI problem, this is simply a false claim. Deep understanding is not some magical ideal that is either achieved or not. Even humans vary greatly in the depth to which they understand what they are reading or what they hear. The real question is whether we can produce systems that can function effectively in specific, realistic, application domains that require an integration of language understanding and reasoning. Furthermore, can we accomplish this using mostly generic techniques, allowing new applications to be constructed fairly simply using the same infrastructure. This is an ambitious goal, but I would argue that we have demonstrated significant progress. Starting with the original TRIPS system, which engaged with a human in mixed-initiative planning (Ferguson et al, 1998), we have generalized the framework and developed a generic architecture for dialogue systems based on models of collaborative problem solving (Allen et al, 2002). This has been applied and evaluated in a number of different domains. Most recently we have been building systems that can learn how to perform tasks for their users from simple dialogue-based demonstrations. We have

shown effective learning of procedures from a single demonstration (Allen et al, 2007; Blaylock et al, 2010). In another application, we have shown that a TRIPS-based system can successfully engage actual medical patients in interviews about their general health status (Ferguson et al, 2009).

Each of these systems shows enough depth of understanding in its domain of application. While each has its own specialized knowledge base and reasoning systems, the language understanding in each is the same, all based on the generic TRIPS architecture (Allen et al, 2001).

Concluding Remarks

My goal is this talk is not to argue that these problems are solved. Far from it, many difficult and challenging problems remain. My goal is to argue that such problems are worthy of study, and that significant progress can be made. I encourage people to think of these deeper problems and look at all options for how we might solve them.

References

Allen, J., G. Ferguson, et al. (2001). An Architecture for More Realistic Conversational Systems. Intelligent User Interfaces (IUI-01), Santa Fe, NM, ACM Press.

Allen, J., N. Blaylock, et al. (2002). A Problem Solving Model for Collaborative Agents. AAMAS-02, Italy, ACM Press.

Allen, J. F., N. Chambers, et al. (2007). PLOW: A collaborative task learning agent. Best Paper, AAAI. Vancouver, BC.

Allen, J.F. et al. (2008) "Deep semantic analysis of text," Symposium on Semantics in Systems for Text Processing (STEP), 2008.

Blaylock, N., W. de Beaumont, et al. (2010). Learning Collaborative Tasks on Textual User Interfaces. FLAIRS-23. Daytona Beach, FL.

Ferguson, G. and J. Allen (1998). TRIPS: An Integrated Intelligent Problem-Solving Assistant. National Conference on Artificial Intelligence (AAAI), Madison, WI, MIT Press.

Ferguson, G., J. Allen, et al. (2009). CARDIAC: An Intelligent Conversational Assistant for Chronic Heart Failure Patient Health Monitoring. AAAI Workshop on Virtual Healthcare Interaction, Arlington, VA.

Pustejovsky, J. et al. (2003) "The TIMEBANK corpus".

Pustejovsky. J. and M. Verhagen (2010) "SemEval-2010 task 13: evaluating events, time expressions, and temporal relations (TempEval-2)," Workshop on Semantic Evaluations: Recent Achievements and Future Directions.

UzZaman. N. and J. F. Allen (2010b), "TRIPS and TRIOS System for TempEval-2: Extracting Temporal Information from Text," presented at the International Workshop on Semantic Evaluations (SemEval-2010), ACL, 2010.