Semantic understanding of natural language text using distant supervision

Sudha Rao (raosudha@cs.umd.edu)

1 Motivation

The beauty of natural language lies in the fact that it allows us to express the same thought in very different ways. But this makes it hard for machines to interpret natural languages since machines are good at memorizing data but pretty bad at generalizing from them. For e.g. a human can easily identify that the two sentences "The boy wants the girl to believe him" and "The boy has a desire to be believed by the girl" mean the same but it is difficult for a machine to do so. What can help machines in this case is if they can work with a kind of representation that abstracts away from various syntactic idiosyncrasies and captures the very meaning of the sentence. Such semantic level of understanding of text is becoming more and more necessary as the field of natural language processing (NLP) tries to solve problems that are closer to natural language understanding.

Most of these NLP systems learn from examples and therefore need annotated data to train on. However, manually annotating data with semantic information is expensive and time-consuming since it requires a linguistic expert in the loop. The availability of rich knowledge bases in recent times has given rise to an indirect way of gathering training data called distant supervision (Mintz et al., 2009). To elaborate, suppose we are interested in extracting relations between proteins from a text and we have database that contains information about protein interactions. Then in this technique, for each pair of proteins that appears in the database, we would find all sentences containing the two proteins in a large unlabeled corpus and use them as our training data.

My research work focuses on how we can use such distant supervision techniques for solving problems that require deeper semantic understanding of text. In subsequent sections I will discuss some of my work that fall under this category and also discuss how these ideas are relevant to work at Facebook Research.

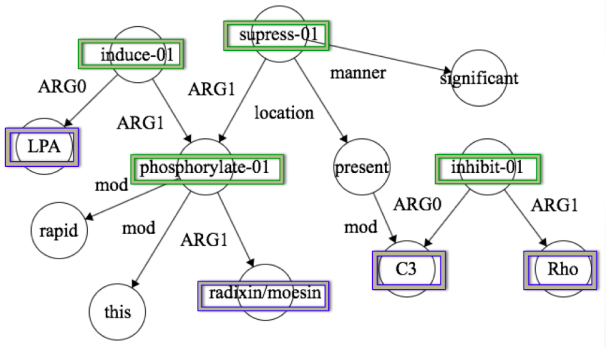

Figure 1: AMR for sentence *"This LPA-induced rapid phosphorylation of radixin/moesin was significantly suppressed in the presence of C3 toxin, a potent inhibitor of Rho"*

2 Relation extraction

An important aspect of natural language understanding is identifying relationships between different entities in a text. For e.g. consider the following sentence from biomedical literature: *"This LPA-induced rapid phosphorylation of radixin/moesin was significantly suppressed in the presence of C3 toxin, a potent inhibitor of Rho"*. Figure 1 shows its Abstract Meaning Representation (AMR) (Banarescu et al., 2013) which is a semantic representation that tries to capture the overall meaning of the sentence in the form of a graph structure. Identifying relations between proteins in this sentence becomes much easier for a machine given the AMR, as opposed to when given the original sentence. For e.g. the subtree rooted at induce-01 shows that *LPA induces phosphorylation*, subtree rooted at phosphorylate-01 shows *this phosphorylation occurs on radixin/moesin* and so on.

During my internship at ISI (Information Science Institute) this summer, I worked with Dr.

Daniel Marcu and Dr. Kevin Knight on developing a feedforward neural network model that made use of the AMR path information to extract relations from biomedical text. We trained our model on huge amounts of data that we labelled automatically using distant supervision from a biological knowledge base called BioPax. This model gave us an absolute improvement of 0.52 in recall (with 0.15 increase in F1) over an existing baseline model. This work is under preparation.

3 Semantic parsing

For semantic representations like AMR to be useful for NLP tasks, we need methods that can automatically parse sentences into these representations. There has been some work on AMR parsing (Flanigan et al., 2014), (Pust et al., 2015), etc including ours (Rao et al., 2015b), where we developed a novel technique to parse English sentences into AMR using SEARN (Daumé III et al., 2009), a learning to search approach to solving structured prediction task. Unlike our predecessors, we modeled the concept (nodes) and the relation (edges) learning in a unified framework and showed an improvement of 2-6% over the baseline on different datasets.

Though there has been some promising work on semantic parsing, accuracy of these parsers is much lower than those we can get for syntactic parsers (e.g. dependency parsers) making them less usable. Paucity of annotated data to learn from is one of the main reasons that makes semantic parsing challenging for machines. But humans do not need such huge amounts of data to efficiently parse a sentence for its meaning. This is because they learn to generalize using their common sense/prior knowledge. Following this insight, I plan on building a semantic parser that can be distantly supervised using knowledge bases like the Wikipedia, DBPedia, OpenCyc, Freebase, etc.

4 Zero pronoun resolution

I first came to appreciate the usefulness of distant supervision in my work on zero pronoun resolution for Chinese (Rao et al., 2015a) in a dialogue setting. Chinese being a "pro-drop" language tends to drop pronouns when they are implicit from the context. The corresponding task of resolving the dropped pronouns to one of the speakers (or to an outside entity in case of third person pronoun) is called zero pronoun resolution. In our work we developed a novel sequential model that explicitly tracked the conversation focus in a dialogue to resolve the zero pronouns.

To train our model, we needed annotated data and it was expensive to have humans annotate data for us. So we turned to distant supervision. English, unlike Chinese, is not a pro-drop language and mentions pronouns explicitly. Also there is huge amount of readily available Chinese-English parallel corpora. We developed a method to automatically label the zero pronouns in Chinese using the pronouns in its English translation, allowing us to have a large amount of labelled data to train on. This kind of a distant supervision, even though noisy, gave significant improvements over purely supervised approaches.

5 Applications to Microsoft Research

I believe that my current expertise in semantics puts me in a good position to contribute to the ongoing research in computational linguistics at Microsoft. The BioNLP project at Microsoft that aims at extracting structured information from biomedical text can make use of deeper semantic representation like AMR. Distant supervision from knowledge bases like the Freebase will enable semantic parsers to generalize beyond seen data and help them parse unseen sentences better. In general, designing models based on semantics will make machine understanding closer to human understanding and thus allow us to build applications that are more natural language friendly.

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.

Hal Daumé III, John Langford, and Daniel Marcu. 2009. Search-based structured prediction. *Machine learning*, 75(3):297–325.

Jeffrey Flanigan, Sam Thomson, Jaime Carbonell, Chris Dyer, and Noah A. Smith. 2014. A discriminative graph-based parser for the abstract meaning representation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1436, Baltimore, Maryland, June. Association for Computational Linguistics.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation ex-

traction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics.

Michael Pust, Ulf Hermjakob, Kevin Knight, Daniel Marcu, and Jonathan May. 2015. Parsing english into abstract meaning representation using syntax-based machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1143–1154.

Sudha Rao, Allyson Ettinger, Hal Daumé III, and Philip Resnik. 2015a. Dialogue focus tracking for zero pronoun resolution. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 494–503, Denver, Colorado, May–June. Association for Computational Linguistics.

Sudha Rao, Yogarshi Vyas, Hal Daumé III, and Philip Resnik. 2015b. Parser for abstract meaning representation using learning to search. *arXiv preprint arXiv:1510.07586*.