Research Overview

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 deal with natural languages since machines are good at memorizing data but pretty bad at generalizing from them. What can help machines deal with this problem 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 computational linguistics tries to solve more challenging problems. My current research focuses on automatically learning a kind of semantic representation called *Abstract Meaning Representation (AMR)* [] given a text and using it for solving various tasks. Since machines learn from examples, one crucial element of my work is to be able to get semantically annotated data that machine learning algorithms can be trained on. However, annotating data with semantic information is a hard task since it requires a linguistic expert in the loop. The availability of rich knowledge bases in recent times has given rise to a learning approach called *distant supervision* where instead of annotating a sentence with information, given an information stored in the database, we gather sentences that may contain evidence for that information. Such an indirect way of gathering training data obviates the need for expensive linguistic annotations. In my research, I also focuses on exploiting this learning paradigm for solving various semantically challenging tasks.

this hints that your thesis/ research is about collecting data

do you want to say time-consuming and expensive. Hard sounds like computationally hard + industry is more interested in time aspect

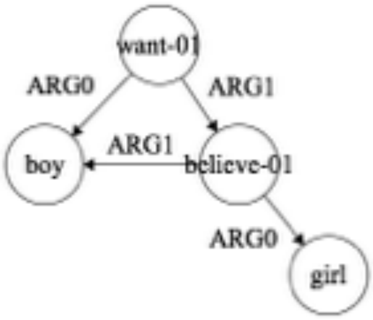
very long sentence

Zero pronoun resolution: In our work on zero pronoun resolution for Chinese [], I first came to appreciate the usefulness of a certain form of distant supervision. Chinese being a "pro-drop" language tends to drop pronouns when they are implicit from the context. This phenomena is more apparent in a dialogue setting since the speakers tend to refer to each other a lot. 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 and used that to resolve the zero pronouns. To train our model, we needed annotated data. One way is to have humans annotate data for us. But this is expensive. English unlike Chinese is not a pro-drop language and hence mentions the pronouns explicitly. Also there is huge amount of readily available Chinese-English parallel corpora. So in our work we developed a method to automatically annotate the training data using the pronouns in the English translation of the Chinese sentence, thus allowing us to have a large amount of data to train on. This kind of a distant supervision, though ~~quite~~ noisy, turned out to be quite useful for our task and gave significant improvements over purely supervised approaches.

don't use vague expressions like a certain form

pronouns are mentioned explicitly

Semantic Parsing: AMR is a form of semantic representation that captures the notion of "who did what to whom" in a sentence in a way that allows several different syntactic variations of a sentence to have the same representation (see figure below).



AMR on the left 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.

As a first step towards using AMR for various natural language processing tasks, we developed a novel technique to parse English sentences into AMR using SEARN [], a learning to search approach to solving structured prediction task. We modeled the concept (nodes) and the relation (edges) learning in a unified framework and showed an improvement of 2-6% over the state-of-the-art on different datasets [].
Relation extraction: In summer 2015, I worked with Dr. Daniel Marcu and Dr. Kevin Knight at ISI on a recent DARPA project called Big Mechanism which aims to develop technology to read research papers and assemble the information into causal models. The group at ISI approached this problem by first parsing the sentence from the paper into an AMR and then developing supervised methods to extract relations from AMR. This baseline system gave an F1 of 0.32 with a low recall of 0.23. We hypothesized that the key reason for low recall was the low amount of training data. Hence I worked on developing a model based on distant supervision that leveraged the information contained in BioPax, a knowledge base of protein-protein interactions, to gather a large amount of additional training data. Further, we labeled them automatically using a path heuristic in its AMR. This work lead to an F1 of 0.49 with a high recall of 0.75. Currently I am working on building deep neural network that can further exploit the AMR graph structure.
Connection to Adobe: I think that my current expertise in semantics puts me in a good position to contribute to the ongoing research in text analytics at Adobe Research. In recent times the combination of vision and natural language processing has been successfully applied to solve many interesting problems. One specific example of work at Adobe would be the project "Editing images with natural language" []. Interpreting user commands can be challenging when they go from simple ones like "make it bright" to more challenging ones like "the image needs to be bright" or "the image is too bright" which clearly have different interpretations. Making use of semantic representation to parse the phrases can ease the task of interpretation and allow the users to express themselves more naturally. At Adobe, I would also be interested in some of the works on deeper semantic analysis of text documents []. Contextualized advertising system based on text can be benefited by deeper semantics. Developing methods to relate the content of the text to existing ontologies (o

The received source data statements are semantically analyzed, which includes matching elements in the received source data statements to respective one or more entries in an ontology associated with the selected domain.

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.
Laput, Gierad P., et al. "PixelTone: a multimodal interface for image editing." Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2013.
USPTO #7,873,640 - Method for Semantic Analysis of Documents to Select and Rank Contextual Advertising Keywords, W. Chang, issued: Jan. 18, 2011.