Leveraging Heterogeneous Data Sources for Relational Semantic Parsing

Meghana Kshirsagar, Nathan Schneider, Chris Dyer
Language Technologies Institute, Carnegie Mellon University

Summary

Semantic parsing of natural language sentences consists of identifying semantic concepts and labeling their arguments. FrameNet and PropBank are both popular linguistic resources for semantic annotations, and were a result of substantial annotation efforts. Semantic parsing systems have so far used one of these resources to train models. Leveraging the knowledge from multiple resources will improve a parser’s coverage of the semantic space. In this work, we present a preliminary exploration of the opportunities and challenges of learning semantic parsers from heterogeneous semantic annotation sources.

The goal of this work is to improve the performance of the frame-semantic parsing system called SEMAFOR[1] by tapping into PropBank-style annotations.

Towards this, we present:
• An analysis of the differences in the semantic coverage of two resources: FrameNet (FN) and PropBank (PB)
• An analysis of the mappings between FN and PB provided by another independent resource called SemLink[2]
• Candidate models for jointly learning a parser on the two resources

The main challenges that any joint model will need to address are:
• The annotations provided by each resource use a different schema: i.e. the label-spaces of the relations and the arguments differ
• Most concepts do not have a one-to-one mapping between the two resources despite being semantically related. This makes it difficult to transfer the annotations from one schema to the other
• The sentence-level annotation densities of the two resources is different and will influence learning
• A mechanism to incorporate the available noisy SemLink mappings

Another related challenge is that of evaluating such a joint model.

Background

Frame semantic parse from the SEMAFOR system for an example sentence:

The semantic parse errors seen in the above sentence are for the following reasons:
• appeals is annotated with the wrong frame label, the correct being EXPERIENCER_OBJ. This frame currently has several predicates without annotations and “appeal” is one of them.
• abolishing is not associated with any frame label, because it is absent from the FN lexicon. It should be recognized as evoking the PROHIBITING frame, which contains synonymous verbs.
• taxing is not identified as a target. It is absent from the FN lexicon; further, none of the existing frames can accommodate it.

Coverage difference: FN and PB

Verb coverage:
FrameNet 1.5  
– 3218 verb types  
PropBank 1.7  
– 5992 verb types  

PB has ~6000 rolesets.

The red portion of each bar shows the additional annotations obtained for that frame upon processing the SemLink mappings. • FN has a total of ~1100 frames. 173 frames get additional annotations. STATEMENT frame gets the highest new annotations.

SemLink data analysis


○ Sentence-level mappings: available on PB-WSJ section • PB-WSJ section has ~75,000 annotations • 50% have SemLink mappings • Of these 20% are usable due to noise and inconsistencies

○ Concept-level mappings:

Statistics of PB-WSJ sentence-level mappings

<table>
<thead>
<tr>
<th>FN frame annotation</th>
<th>PB verb tokens</th>
<th>% of all</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frame label = NF</td>
<td>14,624</td>
<td>20%</td>
</tr>
<tr>
<td>Frame label = IN</td>
<td>22,982</td>
<td>31%</td>
</tr>
<tr>
<td>Frame with no arguments</td>
<td>15,533</td>
<td>21%</td>
</tr>
<tr>
<td>Frame with at least 1 mappable argument</td>
<td>15,323</td>
<td>20%</td>
</tr>
<tr>
<td>Instances not mapped due to other issues</td>
<td>6,516</td>
<td>9%</td>
</tr>
<tr>
<td>Total</td>
<td>74,977</td>
<td>100%</td>
</tr>
</tbody>
</table>

51% of the frame labels are NF (no frame) suggesting there isn’t an equivalent frame or IN (indefinite) suggesting ambiguity in mapping to an appropriate frame

Models

The goal of this work is to improve the performance of the SEMAFOR system. The current system works in two key phases. Frame Identification: selecting a frame for each target  
Argument Identification: finding arguments and labeling them

Guided Parsing based [3]:

Latent variable based:

P(y|x) = P(z|x,y)P(y|x)

Multi-task learning based:

L_{fn}(\theta) + L_{au}(\theta) + \lambda \| \theta \|^2_2

L_{fn}: likelihood over FN data  
L_{au}: likelihood over SemLink

References