Tree Automata Techniques and the Learning of Semantic Grammars

Michael Minock
Umeå University
Umeå, Sweden 90187
mjm@cs.umu.se

Johan Granberg
Umeå University
Umeå, Sweden 90187
johang@cs.umu.se

1 Our current work

We focus on learning statistically weighted semantic grammars for natural language interfaces. Like earlier work (Wong and Mooney, 2007), our system learns over a training set of natural language (NL), meaning representation language (MRL) expression pairs. Also like earlier work our system gives as a result a synchronous context free grammar enriched with λ expressions (λ-SCFG) [1]

As an example, consider that from the input NL/MRL pair "Give me the cities in Ohio" # `(X CITY (= X STATE "Ohio"))` we learn a semantic grammar $G = (N, T_e, T_f, L, S)$ where S is the start symbol, N, T_e and T_f are nonterminals, NL and MRL trminals respectivly. L is the following productions $\langle S \rightarrow GAP^3 \cdot cities \cdot GAP^1 \cdot C_1, (X\ CITY\ C_1\ X)\rangle$ $\langle C \rightarrow Ohio, \lambda x(=\ x\ STATE\ "OHIO")\rangle$ $\langle GAP^3 \rightarrow Give \cdot GAP^2, \rangle$ $\langle GAP^2 \rightarrow me \cdot GAP^1, \rangle$ $\langle GAP^1 \rightarrow the, \rangle$ $\langle GAP^1 \rightarrow in, \rangle$

A key step in the learning algorithm is the extraction of λ-SCFG rules. We model this as tree transduction. A tree in this case is the parse tree of the MRL expression with words from the corresponding NL expression attached [2] In our example from above this tree has the following words aligned with the following MRL productions $\langle cities \leftrightarrow S \rightarrow (X\ CITY\ C)\rangle$ $\langle Ohio \leftrightarrow C \rightarrow (= x\ STATE\ "OHIO")\rangle$.

Given such a tree, a bottom up tree transducer consumes a set of leaf nodes at each step and side effect a λ-SCFG rule. In our example this is the λ-SCFG above. We seek to rigorously formalize this process and leverage more tree automata techniques (**?**). In the short term this promises to give more insight into the rule extraction process and enable us to systematically specify various adaptations and refinements. In the longer term this should give us insight into systematically limiting the hypothesis space over which learning occurs. As is well known, the more one can restrict the hypothesis space, the more rapidly one can learn from examples. Because each tree is of questionable quality due to alignment errors, our idea is to incorporate background linguistic theories encoded as regular tree grammars to limit permissible alignment so as to remove the offending trees in a declarative manner. The closure properties of regular tree grammars seem to be especially promising in this context.

[1] A statistical model θ is also induced, but this is beyond the scope of our presentation.

[2] An alignment process determines these attachment points based on statistical regularities over the entire training set.

2 Our quickfire presentation

We will quickly present the example above and then show a quick sketch of our formalization to date. Then we will pose some questions to the group and subsequently enter into a scientific discussion with some of the researchers present. The work promises to be of interest for the other attendees of the workshop because it touches on a case study of how tree transducers, synchronous grammars and related devices are applied to the NLP problem of learning semantic grammars from NL/MRL corpora.

References

Y. Wong and R. Mooney. 2007. Learning synchronous grammars for semantic parsing with lambda calculus. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-2007)*, pages 960–967.