Semi-Supervised Frame-Semantic Parsing for Unknown Predicates: Supplementary Material

Dipanjan Das and Noah A. Smith
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{dipanjan,nasmith}@cs.cmu.edu

1 Introduction
This document includes supplementary material for the semi-supervised approach towards frame-semantic parsing for unknown predicates (Das and Smith, 2011). We include the names of the test documents used in the study, plot the results for frame-semantic parsing while varying the hyperparameter that is used to determine the number of top frames to be selected from the posterior distribution over each target of a constructed graph and argue why the semi-supervised self-training baseline did not perform well on the task.

2 Names of Test Documents
We include the names of the test documents in Table 1 to facilitate fair replication of our work. The test set contains a mix of several sources from which these documents were drawn.

3 Supplementary Results
For the graph-based models FullGraph and LinGraph, we tuned the hyperparameter $M$ (Das and Smith, 2011, Equation 6, §5.1) along with the other hyperparameters $\alpha$, $K$ and $\mu$ in our model using five fold cross validation. The configuration yielding the best average result for FullGraph was $\alpha = 0.2, K = 10, \mu = 1.0, M = 2$, while for LinGraph it was $K = 10, \mu = 0.1, M = 2$, given $\alpha = 0.0$. In this section, we include supplementary results, both for frame identification and full frame-semantic parsing. We take the label-propagated graph created using the first three hyperparameters set at the aforementioned selected values for the two graph-based models, but vary $M$ from 1 to 10, and observe the effects on frame identification accuracy, and full frame-semantic parsing precision, recall and $F_1$-score. We compare all four models: SEMAFOR, Self-training, LinGraph and FullGraph, for which we report results in the original paper. Figure 1 shows the comparison between the four variants.

For both the graph-based models, performance with respect to all the four evaluation criteria peaks

| Name                                      |
|-------------------------------------------|
| ANC__110CYL067                            |
| ANC__110CYL069                            |
| ANC__112C-L013                             |
| ANC__IntroHongKong                         |
| ANC__StephanopoulosCrimes                  |
| ANC__WhereToHongKong                       |
| KBEval__atm                                |
| KBEval__Brandeis                           |
| KBEval__cycorp                             |
| KBEval__parc                               |
| KBEval__Stanford                           |
| KBEval__utd-icsi                           |
| LUCorpus-v0.3__20000410_nyt-NEW            |
| LUCorpus-v0.3__AFGP-2002-602187-Trans      |
| LUCorpus-v0.3__enron-thread-159550         |
| LUCorpus-v0.3__IZ-060316-01-Trans-1        |
| LUCorpus-v0.3__SNO-525                     |
| LUCorpus-v0.3__sw2025-ms98-a-trans.ascii-1-NEW |
| Miscellaneous__Hound-Ch14                 |
| Miscellaneous__SadatAssassination          |
| NTI__NorthKorea_Introduction               |
| NTI__Syria_NuclearOverview                |
| PropBank__AetnaLifeAndCasualty             |

Table 1: Names of the documents in our test set, taken from the full-text section of the FrameNet 1.5 release.
Figure 1: Comparison of the four models for $M$ ranging from 1 to 10. All numbers correspond to evaluation using partial frame matching. Accuracy measured for frame identification is shown in (a). (b) shows the precision for full frame-semantic parsing, while (c) shows the recall. The $F_1$-score is plotted against $M$ in (d).

at $M = 2$, and then slowly falls off. For all four metrics, FullGraph performs better than LinGraph throughout the range of $M$, although the curves touch each other for $M = 7$ and $M = 8$ for frame-semantic parsing precision. Both these models perform better than SEMAFOR, which in turn performs better than Self-training. From this trend, we can conclude that lower values of $M$ result in better overall performance of the frame-semantic parser, which is supported by the empirical observation that the average frame ambiguity of targets is less than 2.

4 Note about Self-Training

For the frame identification task, Bejan (2009) found self-training to improve overall frame identification performance. Why does self-training not work in our setup? We conjecture that this may happen because of several reasons.
First, Bejan (2009) evaluated using five-fold cross-validation, used an older dataset, and reported a micro-averaged accuracy measure. Hence, there is a mismatch between our experimental setup and his. However, let us compare absolute frame identification accuracy values despite the differences in datasets used for model training. The best result reported in (Bejan, 2009) is 84.73%, which corresponds to self-training. This is an improvement over a supervised model, which is 76.1% accurate. Our purely supervised model SEMAFOR itself, achieves an accuracy of 90.51%, while the FullGraph model improves performance to 91.02%. Hence, our baseline supervised model is very powerful in comparison to the supervised counterpart of Bejan; therefore it is unclear whether self-training can improve over that powerful baseline.

Second, our self-training model is different from Bejan’s in one fundamental aspect. Our goal is to improve the coverage of our frame-semantic parser. Hence, we used a relaxed strategy to identify targets in unlabeled data, and then used the original supervised labels along with the automatic labels to self-train another model. However, Bejan only considers polysemous targets from supervised data, which he finds in unlabeled data, labels them with his supervised model, and uses them for self-training. Thus, we cannot directly compare our self-training setup to his.

Third, finally, our self-trained model does poorly on known targets, which the graph-based models never do, because they do not alter the prediction of SEMAFOR for such targets. We conjecture that this happens because SEMAFOR’s mistakes on even known targets during self-training introduce further noise, bringing about a kind of “semantic drift.”

References
C. A. Bejan. 2009. Learning Event Structures From Text. Ph.D. thesis, The University of Texas at Dallas.
D. Das and N. A. Smith. 2011. Semi-supervised frame-semantic parsing for unknown predicates. In Proc. of ACL-HLT.