FRAUNHOFER INSTITUTE FOR ALGORITHMS AND SCIENTIFIC COMPUTING SCAI

Committee-based selection of weakly labeled instances for learning relation extraction

Tamara Bobić[1,2], Roman Klinger[1,3]

[1] Fraunhofer SCAI, Schloss Birlinghoven, 53754 St. Augustin, Germany [2] B-IT, Uni Bonn, Dahlmannstr. 2, 53113 Bonn, Germany [3] Semantic Computing, CITEC, Uni Bielefeld, 33615 Bielefeld, Germany

Introduction:

Manual annotation is a tedious and time consuming process, usually needed for generating training corpora to be used in a machine learning scenario. The distant supervision paradigm aims at automatically generating such corpora from structured data, while active learning aims at reducing the effort needed for manual annotation. We explore active and distant learning approaches jointly to limit the amount of automatically generated data needed for the use case of relation extraction by increasing the quality of the annotations.

Distant supervision:

- Automatically annotated data set using structured knowledge bases
- Typically noisy, filtering approaches using different heuristics applied
- Silver standard corpora derived from Medline

Relation extraction:

- Machine learning based classification of co-occurring entities in a sentence
- Linear SVM classifier with a rich feature vector

Committee-based selection of instances:

- Small manually annotated seed set
- Each committee member selected by sampling with replacement
- The agreement of the committee used to rank and select preferred instances (high agreement - high confidence regarding a label)

Selection strategies:

(1) Most similar to the seed training set (high confidence, low information gain)
(2) Dissimilar to the seed training set (low confidence, high information gain)
(3) Similar to the seed set, with a chance of having novel aspects

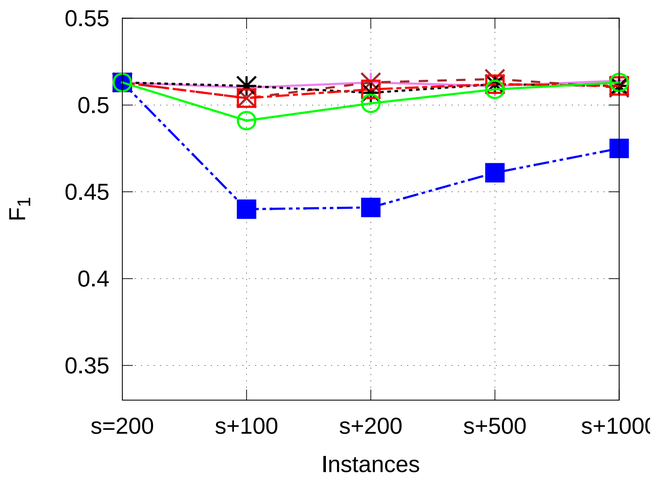


Fig. 1(a): Train on LLL, test on IEPA

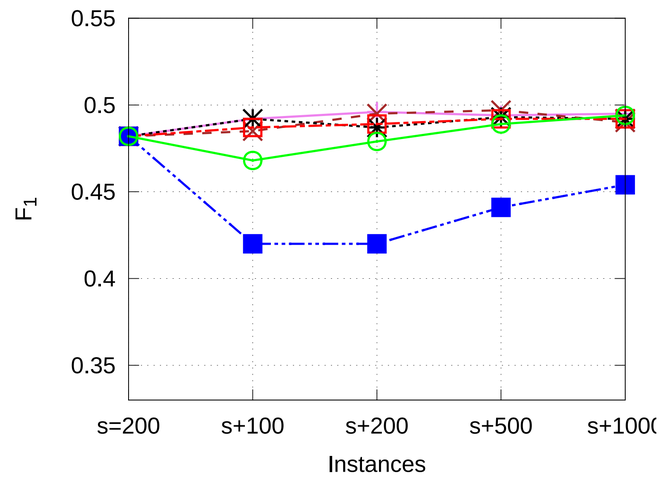


Fig. 1(b): Train on LLL, test on BioInfer

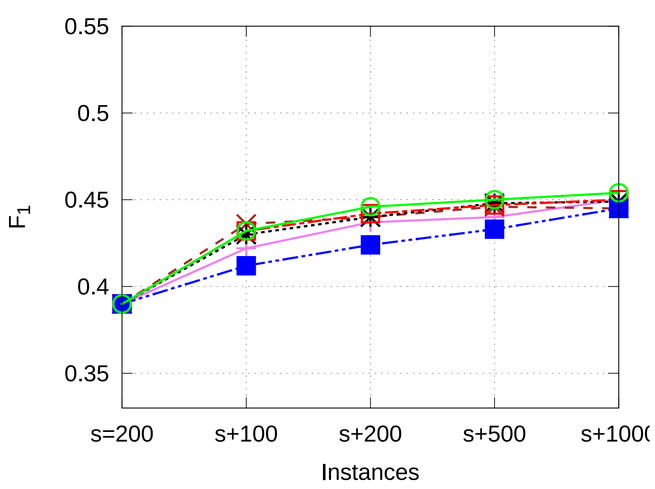


Fig. 2(a): Train on HPRD50, test on IEPA

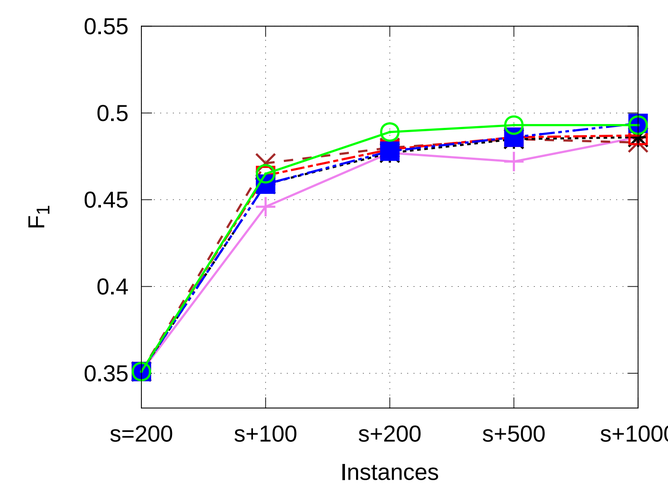
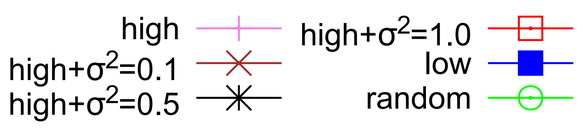


Fig. 2(a): Train on HPRD50, test on IEPA



Results / Discussion:

- Training on LLL: all strategies except (2) outperform the random baseline significantly
- Training on HPRD: all selection methods have a positive impact (these differences are not significant)
- Difference in performance probably due to internal structure of the corpora
- Committee-based selection of 100 additional instances with strategy (1) comparable to 500-1000 instances chosen randomly
- Surprisingly, „safe“ instances are most favourable
- Results are motivating, however, future work should evaluate additional parameters

References:

Thomas, P., Bobic, T., Hofmann-Apitius, M., Leser, U., Klinger, R.: Weakly labeled corpora as silver standard for drug-drug and protein-protein interaction. In: Workshop on Building and Evaluating Resources for Biomedical Text Mining, LREC. (2012)
Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: ACL/AFNLP. (2009)
Freund, Y., Seung, S., Shamir, E., Tishby, N.: Selective sampling using the query by committee algorithm. Machine Learning (1997)

Contact:

tamara.bobic@scai.fraunhofer.de
rklinger@cit-ec.uni-bielefeld.de