Selective Sampling for Information Extraction with a Committee of Classifiers

Ben Hachey, Markus Becker, Claire Grover & Ewan Klein
University of Edinburgh
{bhachey,s0235256,grover,ewan}@inf.ed.ac.uk

Introduction

We describe the Edinburgh-Stanford approach to task 2 of the Pascal Challenge *Evaluating Machine Learning for Information Extraction*,[1] the active learning task. Active learning promises to reduce the cost of supervised training by requesting the most informative data points for human annotation. The literature contains a number of approaches to selecting informative data points. Example informativity can be estimated by the degree of uncertainty of a single learner as to the correct label of a data point (Cohn et al., 1995) or in terms of the disagreement of a committee of learners (Seung et al., 1992). Active learning has been successfully applied to a variety of tasks such as document classification (McCallum and Nigam, 1998), part-of-speech tagging (Argamon-Engelson and Dagan, 1999), and parsing (Thompson et al., 1999).

We employ a committee-based method where the degree of deviation of different classifiers with respect to their analysis can tell us if an example is potentially useful. Trained classifiers can be caused to be different by bagging (Abe and Mamitsuka, 1998), by randomly perturbing event counts (Argamon-Engelson and Dagan, 1999), or by employing different feature sets for the same classifiers (Baldridge and Osborne, 2004). In this paper, we present active learning experiments for information extraction following the last approach.

Our approach gives the highest average improvement over random sampling for micro-averaged and macro-averaged f-scores.

System Description

We use a conditional Markov model tagger (Klein et al., 2003; Finkel et al., 2005) to train two different models on the same data by splitting the feature set. The tagger incorporates a maximum entropy model classifier and models tag sequences following McCallum et al. (2000). There are a large number of metrics that could be used to quantify the degree of deviation between classifiers in a committee (e.g. KL-divergence, information radius, f-complement). The work reported here uses two sentence-level metrics based on KL-divergence and one based on f-measure.

[1] http://nlp.shef.ac.uk/pascal/

KL-divergence has been used for active learning to quantify the disagreement of classifiers over the probability distribution of output labels (McCallum and Nigam, 1998). It measures the divergence between two probability distributions p and q over the same event space:

$$D(p \,||\, q) = \sum_{x \in \chi} p(x) \log \frac{p(x)}{q(x)}$$

KL-divergence is a non-negative metric. It is zero for identical distributions; the more different the two distributions, the higher the KL-divergence. Intuitively, a high KL-divergence score indicates an informative data point. However, in the current formulation, KL-divergence only relates to individual tokens. In order to turn this into a document score, we need to combine the individual KL-divergences for the tokens within a document into one single score. We employed mean and max.

The *f-complement* has been suggested for active learning in the context of NP chunking as a structural comparison between the different analyses of a committee (Ngai and Yarowsky, 2000). It is the pairwise f-measure comparison between the multiple analyses for a given sentence:

$$f_{comp}^{\mathrm{M}} = \frac{1}{2} \sum_{M, M' \in \mathrm{M}} (1 - F_1(M(t), M'(t)))$$

where F_1 is the balanced f-measure of $M(t)$ and $M'(t)$, the preferred analyses of data point t according to different members M,M' of ensemble M. The definition assumes that in the comparison between two analyses, one may arbitrarily assign one analysis as the gold standard and the other one as a test case. Intuitively, examples with a high f-complement score are likely to be informative.

Results & Discussion

We tested 3 feature splits using the features described in the Stanford submission to tasks 1 and 3 of the challenge. None of the feature splits are completely independent of each other. We considered independence important, but we considered it more important to have classifiers that perform about equally well. Experimental results back this up. The feature split we used for the test set evaluation divides the features into word features—such as the word token and the word shape string—and non-word features—such as part-of-speech and whether or not a name has been recognised previously.

We also ran development experiments to try and determine the best selection metric. We chose to use averaged KL-divergence for the test set evaluation. It performed better than maximum KL-divergence and f-complement in the development experiments for this task. Furthermore, we have found it to be the most cost-efficient selection metric in previous work on named entity recognition. Figure 1 contains results on the test set for plotting against number of documents.

The results are generally good: the selective sampling curve shows significant improvement over the baseline random sampling curve. In fact, our active learning system achieves the highest average improvement over baseline (0.013 and 0.014 points respectively for micro- and macro-averaged f-scores). We did not, however, achieve the highest absolute scores.

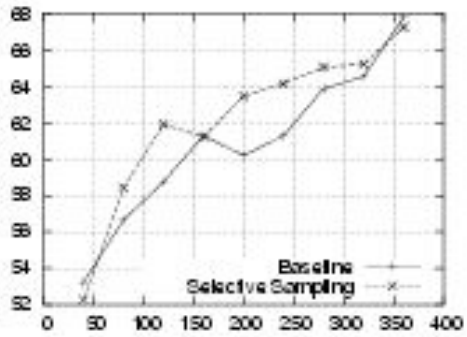

Figure 1: Learning curve on the test set.

Related work suggested that feature sets do not necessarily have to be perfectly independent to build a functional committee (Osborne & Baldridge, 2004). The fact that the selective sampling curve is consistently higher than the baseline random sampling curve lends further support to this conclusion.

Conclusion

We have presented the approach to active learning that we used for the Pascal Challenge *Evaluating Machine Learning for Information Extraction*. This involves forming a committee of classifiers by splitting the features into two feature sets that perform approximately equally. We used averaged KL-divergence as a document-level measure of the difference between the analyses of the classifiers. We also discussed the evaluation results where our system achieved the best average f-score improvement. In the workshop talk, we will also present an analysis of sources of errors and discuss the types of difficult examples our approach is able to identify.

References

Shlomo Argamon-Engelson and Ido Dagan. 1999. Committee-based sample selection for probabilistic classifiers. *Journal of Artificial Intelligence Research*, 11:335–360.

Jason Baldridge and Miles Osborne. 2004. Ensemble based active learning for parse selection. In *Proceedings of the 5th Conference of the North American Chapter of the Association for Computational Linguistics*.

David. A. Cohn, Zoubin. Ghahramani, and Michael. I. Jordan. 1995. Active learning with statistical models. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7, pages 705–712. The MIT Press.

Jenny Finkel, Shipra Dingare, Christopher Manning, Beatrice Alex, Malvina Nissim, and Claire Grover. 2005. Exploring the boundaries: Gene and protein identification in biomedical text. *BMC Bioinformatics*. In press.

Dan Klein, Joseph Smarr, Huy Nguyen, and Christopher D. Manning. 2003. Named entity recognition with character-level models. In *Proceedings the 7th Conference on Natural Language Learning*.

Andrew McCallum and Kamal Nigam. 1998. Employing EM and pool-based active learning for text classification. In *Proceedings of the 15th International Conference on Machine Learning*.

Andrew McCallum, Dayne Freitag, Fernando Pereira. 2000. Maximum Entropy Markov Models for Information Extraction and Segmentation. In *Proceedings of the 17th International Conference on Machine Learning*.

Grace Ngai and David Yarowsky. 2000. Rule writing or annotation: Cost-efficient resource usage for base noun phrase chunking. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*.

H. Sebastian Seung, Manfred Opper, and Haim Sompolinsky. 1992. Query by committee. In *Computational Learning Theory*.

Stephanie Strassel, Alexis Mitchell, and Shudong Huang. 2003. Multilingual resources for entity extraction. In *Proceedings of the ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition*.

Cynthia A. Thompson, Mary Elaine Califf, and Raymond J. Mooney. 1999. Active learning for natural language parsing and information extraction. In *Proceedings of the 16th International Conference on Machine Learning*.