Unsupervised Prediction Aggregation

Alexandre Klementiev
Johns Hopkins University
Baltimore, MD 21218
aklement@gmail.com

Dan Roth
University of Illinois
Urbana, IL 61801
danr@illinois.edu

Kevin Small
Tufts University
Medford, MA 02155
kevin.small@tufts.edu

Ivan Titov
University of Saarland
66041 Saarbrücken, Germany
titov@coli.uni-sb.de

Consider the scenario where votes from multiple experts utilizing different data modalities or modeling assumptions are available for a given prediction task. The task of combining these signals with the goal of obtaining a better prediction is ubiquitous in Information Retrieval (IR), Natural Language Processing (NLP) and many other areas. In IR, for instance, meta-search aims to combine the outputs of multiple search engines to produce a better ranking. In NLP, aggregation of the outputs of computer systems generating natural language translations [7], syntactic dependency parses [8], identifying intended meanings of words [1], and others has received considerable recent attention. Most existing learning approaches to aggregation address the supervised setting. However, for complex prediction tasks such as these, data annotation is a very labor intensive and time consuming process.

In this line of work, we first derive a mathematical and algorithmic framework for learning to combine predictions from multiple signals *without supervision*. In particular, we use the extended Mallows formalism (e.g. [5, 4]) for modeling aggregation, and derive an unsupervised learning procedure for estimating the model parameters [2]. While direct application of the learning framework can be computationally expensive in general, we propose alternatives to keep learning and inference tractable. The intuition behind our approach is that the agreement between signals can serve to estimate their relative quality, which can in turn be used to induce aggregation. Indeed, higher quality signals are better at generating labels close (defined in terms of a distance function) to correct prediction and thus will tend to agree with one another, whereas the poor ones will not. The key assumption we make is that predictions induced by signals are conditionally independent given the true prediction. We demonstrate the effectiveness of our framework on the tasks of aggregating *permutations* and aggregating *top-k* lists.

In many practical applications, the relative quality of the constituent signals is unlikely to remain the same across different domains. Consider, for example, the meta-search task we mentioned earlier. The relative quality of the search engines is likely to depend on the type of the query issued: one may specialize on ranking product reviews while others on ranking scientific documents. Therefore, we extend our aggregation formalism to explicitly model such latent variability in the quality of the constituent signals [3]. We again instantiate the extended framework on aggregating *permutations* and *top-k* lists and experimentally demonstrate (Figure 1, left) that it is capable of learning a better aggregation than our type agnostic model if the variability is indeed present in the data.

The original and the extended distance-based formalisms we used to model aggregation were originally introduced in the context of rank data. However, we propose that they can be generalized to arbitrary types of predictions as long as an appropriate distance function is defined for the output space. We instantiate the framework again for aggregating the outputs of dependency parsers, and experimentally demonstrate on the CoNLL-2007 shared task [6] data that we can induce an effective aggregation (Figure 1, right) particularly when the number of systems we combine is small.

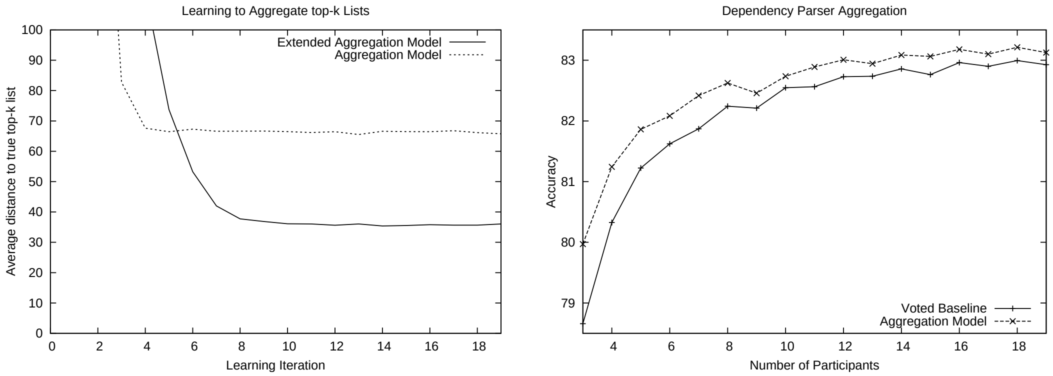

Figure 1: *Left*: learning to aggregate several experts generating top-k lists and exhibiting variability in relative quality. The extended model significantly outperforms the type agnostic counterpart (i.e. achieves lower average extended Kendall distance to true predictions, see [3]) trained on the same data indicating that latent expertise variability can indeed be exploited to produce a better aggregation model. *Right*: learning to aggregate dependency parsers from CoNLL-2007 shared task [6] with varying number of systems (averaged over ten languages). The learned aggregation consistently outperforms the majority vote baseline.

References

[1] S. Brody, R. Navigli, and M. Lapata. Ensemble methods for unsupervised WSD. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 97–104, 2006.

[2] A. Klementiev, D. Roth, and K. Small. Unsupervised rank aggregation with distance-based models. In *Proc. of the International Conference on Machine Learning (ICML)*, 2008.

[3] A. Klementiev, D. Roth, K. Small, and I. Titov. Unsupervised rank aggregation with domain-specific expertise. In *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2009.

[4] G. Lebanon and J. Lafferty. Cranking: Combining rankings using conditional probability models on permutations. In *Proc. of the International Conference on Machine Learning (ICML)*, 2002.

[5] C. L. Mallows. Non-null ranking models. *Biometrika*, 44:114–130, 1957.

[6] J. Nivre, J. Hall, S. Kübler, R. McDonald, J. Nilsson, S. Riedel, and D. Yuret. The CoNLL 2007 shared task on dependency parsing. In *Proc. of the CoNLL Shared Task Session of EMNLP-CoNLL*, pages 915–932, 2007.

[7] A.-V. I. Rosti, N. F. Ayan, B. Xiang, S. Matsoukas, R. Schwartz, and B. J. Dorr. Combining outputs from multiple machine translation systems. In *Proc. of the Annual Meeting of the North American Association of Computational Linguistics (NAACL)*, pages 228–235, 2007.

[8] K. Sagae and A. Lavie. Parser combination by reparsing. In *Proc. of the Annual Meeting of the North American Association of Computational Linguistics (NAACL)*, pages 129–132, 2006.