Unsupervised Aggregation for Classification Problems with Large Numbers of Categories

Ivan Titov
Saarland University†

Alexandre Klementiev
Johns Hopkins University†

Kevin Small
Tufts University†

Dan Roth
University of Illinois

Abstract

Classification problems with a very large or unbounded set of output categories are common in many areas such as natural language and image processing. In order to improve accuracy on these tasks, it is natural for a decision-maker to combine predictions from various sources. However, supervised data needed to fit an aggregation model is often difficult to obtain, especially if needed for multiple domains. Therefore, we propose a generative model for unsupervised aggregation which exploits the agreement signal to estimate the expertise of individual judges. Due to the large output space size, this aggregation model cannot encode expertise of constituent judges with respect to every category for all problems. Consequently, we extend it by incorporating the notion of category types to account for variability of the judge expertise depending on the type. The viability of our approach is demonstrated both on synthetic experiments and on a practical task of syntactic parser aggregation.

1 INTRODUCTION

Multiple experts utilizing a range of data modalities or modeling assumptions are often available for a given classification task. Aggregating their predictions, or votes, has been shown to yield more accurate and robust predictions on a variety of classification problems. However, previous work on expert aggregation, also known as ensemble methods (Dietterich, 2000), tends to make at least one of the following assumptions which often do not hold in practice.

Firstly, most approaches constrain the classification target to have a small number of categories (e.g. (Genest and Zidek, 1986; Kahn, 2004)), making them of limited utility for increasingly complex prediction tasks where decisions need to be made over a large, possibly unbounded, output space. Examples of such problems are omnipresent in vision, natural language processing, information retrieval, and numerous other fields. Prototypical tasks include text categorization, tagging text with Wikipedia-based sense tags, structured prediction problems in natural language parsing and image segmentation amongst a plethora of others. Naturally, lifting this assumption makes the aggregation task harder as a large number of categories coupled with finite training data size make parameter estimation significantly more difficult.

Secondly, most existing learning approaches to aggregation address the supervised setting (e.g. (Lebanon and Lafferty, 2003; Liu et al., 2007)). However, we cannot generally assume the availability of annotation for complex prediction tasks we consider in this paper. Indeed, for these problems labeled data is typically sparse and is very expensive to obtain (e.g. (Marcus et al., 1993)). Moreover, relative expert performance depends on the distribution from which the test data is sampled, which implies the need for additional annotation such that a new aggregation model can be estimated for each potential test distribution.

Finally, the votes of constituent experts are often assumed to be independent conditioned on the true category (e.g. (Hinton, 1999; Lebanon and Lafferty, 2003)), an aggregation strategy known as the product rule. While providing a suitable approximation for some applications, it simply may not suffice for others. Normally, even for problems with a very large set of categories for every example there exists a relatively small set of categories (*confusion set*) such that any reasonable expert predicts a category from this set. This property is not explained by the product rule model, and, as we will discuss further, results in aggregate

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9.

†This research was partially done while the authors were at the University of Illinois at Urbana-Champaign.

predictions virtually equivalent to majority votes.

We start by proposing a generative model for *unsupervised* aggregation of experts with a *large (possibly infinite)* number of output categories by *relaxing the conditional independence assumption* on their votes. The model exploits the agreement between constituent judges to estimate their relative expertise. Unlike the product rule model, it makes the exchangeability assumption on the distribution of categories predicted by experts conditioned on these experts being incorrect for a given example. This assumption is weaker than the independence assumption and the resulting model can explain preference towards smaller confusion sets and thereby potentially perform significantly better than the majority vote.

In a realistic aggregation scenario, one would often expect a constituent judge's expertise to depend on the *type* of decision it is being asked to make. For example, a judge may rely on input data characteristics particularly informative for discriminating among a specific subset of the output space, while making poor predictions for other kinds of decisions. We therefore extend this proposed model to also account for variability of the judge expertise depending on the *type* of the predicted decision. Agreement between constituent experts for particular types of predictions is again exploited for model parameter estimation.

We demonstrate the effectiveness of our approach on simulated experiments and the task of dependency parser aggregation. Our aggregation method significantly outperforms the majority vote baseline on both sets of experiments.

The rest of the paper is organized as follows. Section 2 introduces the aggregation model and extends it with the notion of decision types. In Section 3 we describe model learning and inference. The model and its multitype extension are then evaluated on synthetic and dependency parser aggregation experiments in Section 4. Section 5 reviews related work, and we conclude and provide ideas for future work in Section 6.

2 UNSUPERVISED AGGREGATION

As discussed in the preceding section, our goal is to produce a method for aggregating predictions of multiple experts without any supervision requirements. Therefore, it is natural to use generative models of classifier predictions as a basis for our approach. In this section, we formalize the problem and proceed by examining the standard approach of assuming conditional independence of the constituent expert votes given the true category. This analysis assists us in recognizing the limitations of this assumption in the context of our task and motivates our contributions discussed later.

We consider votes of experts selecting a categorical target $y \in \mathcal{Y}$, where $\mathcal{Y}$ is a large (possibly unbounded) set $\mathcal{Y} \subset \mathbb{Z}$. Each expert k predicts a category $y^{(k)}$ to be aggregated into a joint prediction. Formally, our goal is to produce a forecast y^* given the set of expert votes $\mathbf{y} = (y^{(1)}, \ldots, y^{(K)})$. Note that, unlike much of the previous work on aggregating experts' judgements (e.g., (Cooke, 1991; Wallsten et al., 1999; Kahn, 2004)), we do not assume access to the probability distributions over the events defined by each expert k, since it is not realistic for problems with a large set of possible categories. Furthermore, we do not assume that we have access to any supervised data (i.e. examples labeled with the true categories y^*), but only the unlabeled dataset $\{x_n\}_{n=1}^N$ and the associated expert predictions $\{\mathbf{y}_n = (y_n^{(1)}, \ldots, y_n^{(K)})\}_{n=1}^N$. In the subsequent sections, we analyze various models which define joint distributions of $\mathbf{y}$ and y^*. It is straightforward to extend these models to incorporate information about the input examples x, but this extension is largely orthogonal to the general goal of this paper.

2.1 CONDITIONALLY INDEPENDENT JUDGES

The simplest generative model would assume that each experts' vote $y^{(k)}$ is conditionally independent given the underlying true category y^* (Kahn, 2004). This assumption results in the following conditional distribution over the categories:

$$P(y^*|\mathbf{y}) = \frac{P_0(y^*)\prod_{k=1}^K P(y^{(k)}|y^*)}{\sum_{y'} P_0(y')\prod_{k=1}^K P(y^{(k)}|y')} \qquad (1)$$

$P_0(y^*)$ is the prior probability of the category y^* and $P(y^{(k)}|y^*)$ is the probability of predicting $y^{(k)}$ instead of y^* by expert k. Note that given a large set of categories $\mathcal{Y}$, an aggregation model cannot reliably estimate the parameters of these distributions and additional independence assumptions are necessary. Similar to distance-based models (Mallows, 1957; Klementiev et al., 2008), we can assume that the probability of making a mistake θ_k is independent of y^* and that the probability of each category $y^{(k)}$ given that expert k is incorrect is proportional to the prior probability $P_0(y^{(k)})$:

$$P(y^{(k)}|y^*) = \begin{cases} 1-\theta_k & \text{if } y^{(k)} = y^*; \\ \frac{\theta_k}{1-P_0(y^*)} P_0(y^{(k)}) & \text{otherwise.} \end{cases}$$

We can then rewrite (1) as

$$\hat{P}(y^*|\mathbf{y}) \propto$$
$$P_0(y^*)\prod_{k=1}^{K}(1-\theta_k)^{[\![y^{(k)}=y^*]\!]}\left(\frac{\theta_k P_0(y^{(k)})}{1-P_0(y^*)}\right)^{1-[\![y^{(k)}=y^*]\!]}$$

where $[\![p]\!]$ equals 1 if the predicate p is true and 0 otherwise. If we assume that the number of categories is infinite and P_0 is the non-informative prior, then only categories y with the maximum number of votes will have non-zero associated probability $P(y|\mathbf{y})$. Consequently, the predictions of the aggregation model will be equivalent to the majority vote and the parameters θ will only affect the choice of a category in case of ties. The same effect will be observed in a more relaxed case when the number of classes is sufficiently large, the prior distribution is sufficiently uniform, and when there are no extremely accurate or extremely inaccurate experts present on the panel. Formally, the prediction of the aggregation model is guaranteed to agree with a majority vote as long as

$$\frac{P_{max}}{(1-P_{max})}\left(\frac{P_{max}(1-P_{max})}{P_{min}(1-P_{min})}\right)^{M+1} < \frac{\theta_{min}^M(1-\theta_{max})^{K-M}}{\theta_{max}^{M+1}(1-\theta_{min})^{K-M-1}}$$

where $P_{min} < P_0(y) < P_{max}$ for every category y, $\theta_{min} < \theta_k < \theta_{max}$ for every expert k, and M is the maximum number of votes associated with a single category for a given example. Clearly, this model is thereby insufficient for aggregation with a large number of potential categories.

2.2 MODELING CONFUSION SETS

It has been observed for many problems that not all mistakes from a large set $\mathcal{Y}\backslash\{y^*\}$ are equally likely. Instead, for every example x there is often a small subset of $\mathcal{Y}_c(x) \subset \mathcal{Y}$, referred to as confusion set, such that any reasonable expert will predict a label from $\mathcal{Y}_c(x)$. Instead of assuming that the votes of every expert are conditionally independent, we assume that the categories for the mistakes are drawn independently from this confusion set.

In order to define this model we need to estimate the confusion set $\mathcal{Y}_c(x)$, or the confusion distribution, defined as a distribution with a support in $\mathcal{Y}_c(x)$. If an expert k makes a mistake, the resulting category y_k is drawn from this confusion distribution. One potential approach would be to try to estimate the confusion distribution using both the input elements x and votes $\mathbf{y} = (y_1, \ldots, y_K)$. However, this is a hard learning problem, requiring a good feature representation of x and a large amount of training data. We propose a simpler technique, drawing confusion distributions from a Dirichlet process (Ferguson, 1973). The concentration parameter controlling the Dirichlet process, if chosen properly, will enforce a preference for smaller confusion sets.

The graphical model is shown in Figure 1 with the following formal definition. First, draw the true category y^* from the prior distribution P_0 and draw a measure G from a Dirichlet process defined by the concentration parameter α and the measure induced by the prior distribution P_0. Then, for each expert k:

- Decide if expert k is correct by drawing r_k from θ_k,
- If $r_k = 0$ set $y^{(k)}$ to y^*,
- Otherwise, draw $y^{(k)}$ from G.

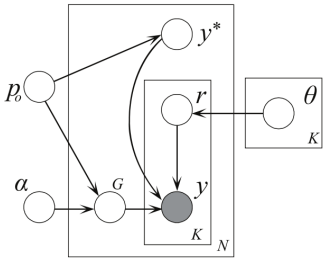

Figure 1: Type-agnostic model

This method, unlike the previously considered model which assumed conditional independence of experts, does not require that the set of potential categories $\mathcal{Y}$ is constant across the input examples. This property is important for aggregation of individual decisions for structured prediction problems, where $\mathcal{Y}$ is likely to change depending on the properties of the input sequence. As an example, consider the problem of predicting a parent of a vertex in a graph. In this case, the number of potential parents grows linearly with the size of the vertex set, whereas the size of the confusion set would not normally be strongly affected.

Note that the underlying model assumption is that categories for mistakes are exchangeable, unlike the previous model where they are assumed independent. This assumption, though not entirely realistic in all cases, results in a much better approximation of confusion distributions, allowing for smaller supports and explaining high agreement between incorrect experts.

Note also that in this section, we assumed that the number of categories is very large or infinite and the prior probabilities $P_0(y) \ll 1$ for every category y. Alternatively, we could use a modification of the model where instead of drawing a measure G from a Dirichlet process, a multinomial distribution is drawn from a Dirichlet prior.

2.3 INCORPORATING CATEGORY TYPES

In the preceding section, we argued that with a large or even unbounded set of possible categories we cannot reliably estimate parameters associated with each category y. This is a serious drawback as the expertise of a judge is in general dependent on the considered category and ignoring this property is likely to limit the potential accuracy of the aggregation method.

However, we can assume that there exists a finite and relatively small set of category *types* such that experts' accuracy differs significantly across types but remains constant or similar for categories within each type. A method for associating types with categories may be derived either from the domain knowledge or from available information about experts' properties. Recall, for example, the previously mentioned task of predicting a relevant page from Wikipedia for a fragment of text. One expert may be better in predicting Wikipedia pages describing named entities, whereas another may be good at finding definitions for scientific terms. For learned experts, these differences may be prominent due to a number of reasons including the discrepancies in their underlying statistical models, differences in the views used to derive the representation, or differing distributional properties of the data used to estimate their parameters.

We associate two parameters with each pairing of expert k and type t. The first parameter is similar to the accuracy parameter θ_k in the type-agnostic model: $\theta_{k,t}$ is the probability of making a mistake given that the true category y^* has type $t \in 1, \ldots, T$. In other words, $1 - \theta_{k,t}$ is the recall of expert k on type t. It may seem that conditioning on the type of y^* is the only required modification to the type-agnostic model, but note that modeling only recall parameters is not sufficient as demonstrated by the following example.

Consider a difficult instance where an example generates little agreement among experts. Furthermore, assume that a weak expert j predicts a category y with type t even though all the remaining, more accurate, experts predict categories of type $t' \neq t$. Assume also that type t is 'harder' than type t' (i.e. all experts have much lower recall on type t) and consequently, $\theta_{k,t} \gg \theta_{k,t'}$ for all k. Clearly, an appropriate model should predict some category y' from type t' as it is reasonable to assume that experts have relatively balanced recall and precision. However, the recall-focused model will predict category y of type t because

$$(1 - \theta_{j,t}) \prod_{k \neq j} \theta_{k,t} \gg \prod_{k=1}^{K} (1 - \theta_{k,t'})^{[\![y^{(k)} = y']\!]} \theta_{k,t'}^{1 - [\![y^{(k)} = y']\!]}.$$

This problem occurs because the model does not penalize for false positives of type t.

In order to address this deficiency, we introduce another set of parameters $\varphi_k = (\varphi_{k,1}, \ldots, \varphi_{k,T})$ which defines a distribution of false positives over types for the expert k. If expert k makes a mistake on an example, we select the type for its vote from distribution φ_k. These two sets of parameters, θ and φ, allow us to control both precision and recall of experts for each type. If prior knowledge suggests that the classifiers are balanced, we can enforce this balance by constraining θ and φ, but we will not explore this direction in the paper.

Formally, the generative process proceeds as follows. First, we draw type t^* for the true category from the prior distribution ψ over types, then choose category y^* from the prior distribution P_{0,t^*} for the type t^*. Also, draw a measure G_t for every type t from a Dirichlet process defined by the concentration parameter α_t and the measure induced by the prior distribution $P_{0,t}$. Then for each expert k:

- Draw r_k from the recall parameters θ_{k,t^*}.
- If $r_k = 0$ set $y^{(k)}$ to y^*,
- Otherwise:
  - draw type $t^{(k)}$ for the vote from φ_k,
  - draw $y^{(k)}$ from $G_{t^{(k)}}$.

The corresponding graphical model is presented below in Figure 2.

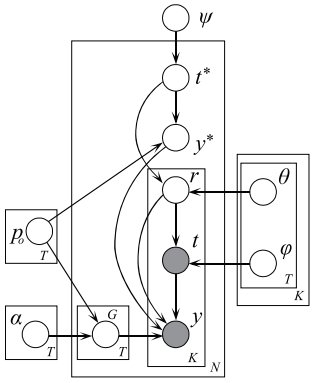

Figure 2: Aggregation with typed decisions.

Note that the number of categories associated with each type does not need to be unbounded, as a different distributional assumption can be made for different types. For example, if some type t contains a very small number of categories, it may be beneficial to draw $y^{(k)}$ for this type directly from $P_{0,t}$. Furthermore, if the number of categories is large but not sufficient to ignore probabilities $1 - P_{0,t}(y)$, we can use Dirichlet priors instead of Dirichlet processes as discussed in the preceding section.

3 LEARNING AND INFERENCE

In this section, we discuss learning and inference procedures for the proposed models.[1]

We consider the more complex case of the typed model throughout this section. As commonly practiced, we analytically marginalize out measures G_t resulting in the Chinese Restaurant Process (CRP) (Pitman, 1995) generating categories for the mistakes. We search for the maximum likelihood estimates for the concentration parameters α_t, type priors ψ, recall parameters $\theta_{k,t}$ and false-positive rates $\varphi_{k,t}$ using the Expectation Maximization (EM) algorithm (Dempster et al., 1977). Though derivation of the EM algorithm is relatively straightforward, we include it for completeness. In order to simplify the notation we assume that the number of categories for every type t is infinite.

In the expectation step, the posterior probabilities of categories appearing in the list are computed as

$$\hat{P}(y^*|\mathbf{y}) \propto \psi_{t^*} \prod_{k=1}^{K} (1-\theta_{k,t^*})^{[\![y^{(k)}=y^*]\!]} \left(\theta_{k,t^*}\varphi_{k,t^{(k)}}\right)^{1-[\![y^{(k)}=y^*]\!]} \prod_{t=1}^{T} R_t(y^*,\mathbf{y}) \quad (2)$$

where t^* and $t^{(k)}$ are types of categories y^* and $y^{(k)}$, respectively, and $R_t(y^*,\mathbf{y})$ is the probability associated by CRP with the partition of the mistakes of type t into groups where each group correspond to a distinct category among those proposed in $\mathbf{y}$. Formally, if we denote by $B(y,\mathbf{y})$ the number of times category y appears in the list of votes $\mathbf{y}$ we can write

$$R_t(y^*,\mathbf{y}) = \frac{\Gamma(\alpha)\alpha^{\sum_{y\neq y^*, t(y)=t}[\![B(y,\mathbf{y})>0]\!]}}{\Gamma(\alpha+\sum_{\substack{y:y\neq y^*\\ t(y)=t}} B(y,\mathbf{y}))} \prod_{\substack{y:y\neq y^*\\ t(y)=t}} \Gamma(B(y,\mathbf{y})),$$

where Γ is the Gamma function.

Similar to expression (2), the total probability assigned to categories y^* of type t^* but not appearing in the list $\mathbf{y}$ is proportional to

$$\sum_{\substack{y^*:t(y^*)=t^*\\ B(y^*,\mathbf{y})=0}} \hat{P}(y^*|\mathbf{y}) \propto \psi_{t^*} \prod_{k=1}^{K} \theta_{k,t^*}\varphi_{k,t^{(k)}} \prod_{t=1}^{T} R_t(y^*,\mathbf{y}) \quad (3)$$

where y^* in r.h.s. is any category not from the list $\mathbf{y}$ (i.e. $B(y^*,\mathbf{y})=0$).

In the M-step the marginal counts computed in the E-step (2-3) are used to reestimate θ, φ and ψ:

$$\theta_{t,k} \propto \sum_{n=1}^{N}\sum_{y^*} \hat{P}(y^*|\mathbf{y}_n)[\![t^*=t]\!][\![y^{(k)}\neq y^*]\!],$$

$$\varphi_{k,t} \propto \sum_{n=1}^{N}\sum_{y^*} \hat{P}(y^*|\mathbf{y}_n)[\![y^{(k)}\neq y^*]\!][\![t^{(k)}=t]\!],$$

$$\psi_t \propto \sum_{n=1}^{N}\sum_{y^*} \hat{P}(y^*|\mathbf{y}_n)[\![t^*=t]\!].$$

We use the non-informative hyperprior for concentration parameters α_t, and reestimate them by performing line search on every iteration of EM.

The aggregate prediction with this model is derived by selecting the most probable category y^* from the posterior distribution (2). If the aggregation of individual decisions corresponding to some structured prediction problem is considered, additional structural constraints on the valid sequences of decisions can be enforced. In this case, local inference using the posterior distribution (2) can be replaced with global inference where marginal distributions for individual decisions are approximated by the distribution (2-3). However, these issues are largely orthogonal to the goal of the paper and have been studied previously in the context of majority voting (Sagae and Lavie, 2006).

4 EMPIRICAL EVALUATION

In this section, we experimentally evaluate both the type-agnostic model and its category type extension as introduced in Section 2. We first study the models with synthetic experiments, and then move on to the task of aggregating dependency parsers. We also demonstrate that modeling confusions sets with Dirichlet processes is crucial to the success of our model.

4.1 SYNTHETIC EXPERIMENTS

Let us begin by describing the procedure used to generate synthetic experts, whose output we then learn to aggregate. The aim of the following construction is to produce experts which rely on input data modalities (subsets of features) particularly informative for discriminating among a specific type of the output categories. In particular, note that we do not enforce exchangeability or conditional independence of their votes, thus modeling a realistic setting.

We first split the a of $|\mathcal{Y}| = 150$ output categories equally into $T = 3$ types. We then construct a Naïve Bayes model $\mathcal{B}$ selecting the conditional probability distributions to make some features more informative

[1] As before, for simplicity we assume that the prior probability of every category is negligible, $P_0(y) \ll 1$. A simple modification of the model can handle the general case.

for categories of each type. In particular, the conditional probability of a feature being active for a category of a given type was chosen from a uniform distribution over $(0.6, 1.0]$ with probability 0.1 and from $[0.0, 0.1)$, otherwise.

Each of the K experts was then generated by training a Naïve Bayes classifier with varying proportions of types of 1000 examples sampled from $\mathcal{B}$. The proportion coefficients were selected from a symmetric Dirichlet prior with parameters set to 0.5, which was also used to select proportions of the category types in the test data.

We trained both the type-agnostic and typed aggregation models on the predictions of K experts for 1000 test examples, letting K span $[3, 20]$.

Our baseline is the majority vote for each test example, with ties broken randomly. The accuracies of both models and the baseline were averaged over 5 runs and are shown in Figure 3. While the type-agnostic model performs similarly to the voted baseline, the typed model is able to learn and take advantage of the type specificity present in judges' expertise and significantly exceeds the performance of both the type-agnostic model and the baseline. The models and the baseline also outperform the single best expert for all the trials where the number of experts exceeds 3.

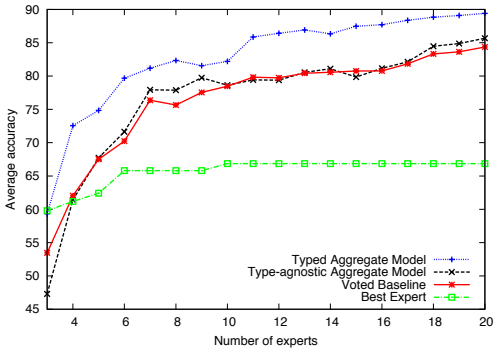

Figure 3: Experiments with Synthetic Data – The type-agnostic aggregation model performs on par with the voted baseline, while the typed model exploits type specificity to significantly outperform both the voted baseline and type-agnostic model.

Since the distributions of votes are observable, one may assume that careful modeling of confusion sets is not needed. To counter this argument, we replaced our generative model with a model which conditions variables on the set of observable categories. In this case, categories y are generated not from the entire set of potential categories, but only from a small subset of observable categories. As a result, $\prod_{t=1}^{T} R_t(y^*, \mathbf{y})$ in (2) is replaced by the product of these generation probabilities. The accuracy of the type-agnostic version of this model (omitted from Figure 3 to improve clarity) is significantly below the majority vote baseline when the number of experts is less than 16 (42.6% vs 67.5%, 66.1% vs 78.4%, 79.4% vs 79.9% with 5, 10 and 15 experts, respectively), although it does outperform the type-agnostic model insignificantly with 19 and 20 experts (85.0% vs 84.8% and 86.0% vs 85.7%). The typed version of this model demonstrates similar behavior, beating the majority-vote baseline only with the number of experts above 15 but staying at least 5% below its generative counterpart in all cases (with the single exception of $K = 19$ where this difference is 1%). This result suggests that a trivial conditional model is not appropriate for our task.

4.2 DEPENDENCY PARSING

The second set of experiments consider the problem of aggregating votes of syntactic dependency parsers. As experts for our model, we use parsers constructed by participants of the multilingual track of the CoNLL-2007 shared task (Nivre et al., 2007). Though we do not have access to the parsers, the organizers of the competition distributed dependency structures predicted by parsers on small test sets for all 10 considered languages: between 131 and 690 sentences and between 4513 and 5390 words, depending on the language. The numbers of experts slightly varies across the languages but always remains between 20 and 23. For further information on the datasets and the participating systems we refer the reader to the shared task report (Nivre et al., 2007).

The dependency structures represent syntactic relations between words in a sentence (Tesniére, 1959). Every word has at most one other word as its *syntactic head* and each such relation has an associated type. The number of potential relation types depends on a language and linguistic formalism; for datasets in the CoNLL-2007 shared task, the number of relation types is between 20 and 69. We aggregate the experts' votes on the level of words, for each word each expert k provides a pair $y^{(k)} = (h^{(k)}, r^{(k)})$, where $h^{(k)}$ is an index of another word in the sentence and $r^{(k)}$ is a predicted relation type. Given an average number of types of about 40 and an average sentence length of around 20, one could estimate that for every word in a sentence there are approximately 1,000 potential relations and this number varies across sentences. On the contrary, the size of the confusion sets is relatively small, the experts predict only 3.6 distinct pairs $(h^{(k)}, r^{(k)})$ averaged over examples and languages. These observations suggest that our aggregation model should be appropriate for the problem.

We estimate a separate model for each language. For the experiments, we vary the number of experts by selecting K parsers from the list. They are chosen at random while ensuring that the distribution of their accuracy is similar to the distribution of accuracy of the entire set of experts for the considered language. To achieve this, we first split experts in K buckets of approximately equal size depending on their accuracy and then select a random expert from each bucket.[2]

We found that the EM algorithm converges quickly with essentially no decrease in the likelihood function after 10 iterations, as was also the case with artificial data experiments. Furthermore, this method appears insensitive to the initialization parameters.

In Figure 4, we present accuracy of the aggregate systems averaged over 10 languages and compare it to the accuracy of the majority vote baseline. Note that in practice it is unlikely that we can afford to have 20 different parsers in an ensemble, so of a higher practical interest is the part of the curve with a smaller number of experts. These improvements from using the aggregation model are statistically significant with p-value < 0.05 for all the experiments.[3] The improvement was consistent across all languages: the aggregation model outperformed the baseline on all languages in all experiments except for 2 and 1 languages out of 10 with the number of experts equal to 3 and 8, respectively.

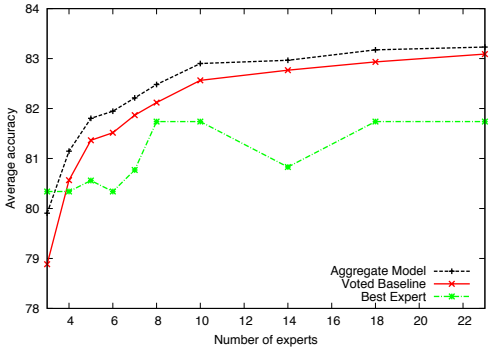

Figure 4: Dependency Parsing Experiments – Our aggregation model outperforms both the best expert and voted baseline over an average of 10 languages. In particular, note that this improvement is more pronounced in the more practical case of fewer experts.

We do not use types in these experiments, primarily because accurate estimation of parameters associated with types is not feasible for such a small number of examples. Note that among the available examples for every language approximately 25% are not informative as all experts are in perfect agreement. In a practical aggregation scenario, where the experts are available, it is easy to obtain the votes on an arbitrary amount of unlabeled data and this would allow for use of more powerful aggregation models. In this case, existing work on error analysis for dependency parsing may suggest appropriate ways to introduce types of relations (e.g., (Nivre and McDonald, 2008)).

[2]Note that this resulted in non-monotonicity of best-expert accuracy in Figure 4.

[3]We used the permutation test (Diaconis and Efron, 1983) to measure significance.

5 RELATED WORK

The problem of combining expert predictions has been extensively studied in statistics, see (Genest and Zidek, 1986; Kahn, 2004) for a thorough review. However, most of the research was focused on aggregating probability distributions and on categorical events with a small number of categories. A significant body of work exists regarding aggregating human opinions (Cooke, 1991), and though relevant, primarily focuses on specific aspects of human reasoning about uncertainty.

In machine learning, ensemble methods have played a prominent role (Dietterich, 2000). However, most existing research, with rare exception (Kahn, 2004; Klementiev et al., 2008), considers supervised methods which use labeled data both to learn the constituent experts and to learn how to combine them within an ensemble. This may not be optimal when we consider applying these systems to the data from new domains; it has been observed that relative performance of systems is likely to change when ported to a new domain (Globerson and Roweis, 2006). We are not aware of any previous work focused on the problem of unsupervised aggregation of experts' votes for the problem with large, infinite or variable number of output categories. Simple heuristics for combining classifiers, such as the majority vote, are commonly practiced for these problems (e.g. (Sagae and Lavie, 2006)), emphasizing the need for better aggregation techniques.

Another related research direction has focused on methods for joint semi-supervised learning of constituent judges (Rosenberg and Bartlett, 2007; Liang et al., 2008), deriving a learning method which forces the constituent judges to agree on the unlabeled data. Despite being potentially more powerful, this approach may not always be feasible, as often it is not possible to train all the judges jointly, or some of them may not even be statistical classifiers.

The exchangeability assumption on expert votes was considered in (Mendel and Sheridan, 1986). However, they assume that all the experts votes are exchangeable when conditioned on the true category, whereas our assumption is that only categories predicted by incorrect experts are exchangeable. Therefore, unlike

our case, they are not able to associate expertise parameters with individual experts.

Distance-based models (Mallows, 1957; Lebanon and Lafferty, 2003) were used in the context of unsupervised aggregation of rankers (Klementiev et al., 2008) and more recently extended to include input data domain specificity (Klementiev et al., 2009). All these models make conditional independence assumptions which is not realistic for the problem class considered in this paper. There have been previous attempts to incorporate types in aggregation models (Klementiev et al., 2009) in this context. However, in their model types (domains) are associated with input example and the model attempts to distinguish between types purely on the basis of agreement patterns. This approach is orthogonal to the direction studied in this paper and may be less suitable when the number of available experts is small.

6 CONCLUSIONS

In this work, we have proposed an unsupervised learning framework for aggregating experts which generate votes from a large (possibly, infinite) set of categories. It is then extended to incorporate the notion of category types, since judges' expertise are often dependent on the types of decisions they are asked to make. The efficient learning procedure we propose exploits the agreement between judges to estimate the model parameters and does not require supervision, which is typically very expensive to obtain for complex prediction tasks. We first study the predictive performance of type-agnostic and typed models on simulated data and demonstrate that incorporating the notion of category types can significantly boost model performance. We further evaluate the model on aggregating dependency parsers. We train the model on the parses produced by the CoNLL-2007 shared task participants for 10 languages, and again show an improvement over our voted baseline.

Acknowledgements

This work is supported by NSF grant ITR IIS-0428472, DARPA funding under the Bootstrap Learning Program, MMCI Excellence Cluster at Univ. Saarland, and MIAS – a DHS-IDS Center for Multimodal Information Access and Synthesis at UIUC.

References

Cooke, R. M. (1991). *Experts in Uncertainty: Opinion and Subjective Probability in Science*. Oxford University Press.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithms. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Diaconis, P. and Efron, B. (1983). Computer-intensive methods in statistics. *Scientific American*.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Proc. of the First International Workshop on Multiple Classifier Systems*.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2).

Genest, C. and Zidek, J. V. (1986). Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1(1):114–135.

Globerson, A. and Roweis, S. T. (2006). Nightmare at test time: robust learning by feature deletion. In *ICML*.

Hinton, G. (1999). Products of experts. In *ICANN*.

Kahn, J. M. (2004). *Bayesian aggregation of probability forecasts on categorical events*. PhD thesis, Stanford University.

Klementiev, A., Roth, D., and Small, K. (2008). Unsupervised rank aggregation with distance-based models. In *ICML*.

Klementiev, A., Roth, D., Small, K., and Titov, I. (2009). Unsupervised rank aggregation with domain-specific expertise. In *IJCAI*.

Lebanon, G. and Lafferty, J. (2003). Conditional models on the ranking poset. In *NIPS*.

Liang, P., Klein, D., and Jordan, M. (2008). Agreement-based learning. In *NIPS*.

Liu, Y.-T., Liu, T.-Y., Qin, T., Ma, Z.-M., and Li, H. (2007). Supervised rank aggregation. In *WWW*.

Mallows, C. L. (1957). Non-null ranking models. *Biometrika*, 44:114–130.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19.

Mendel, M. and Sheridan, T. (1986). Optimal combination of information from multiple sources. Technical report, Massachusetts Institute of Technology.

Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., and Yuret, D. (2007). The CoNLL 2007 shared task on dependency parsing. In *Proc. of the CoNLL Shared Task, EMNLP-CoNLL*.

Nivre, J. and McDonald, R. (2008). Integrating graph-based and transition-based dependency parsers. In *Proc. of the Assoc. for Computational Linguistics*.

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, 102(2):145–158.

Rosenberg, D. and Bartlett, P. L. (2007). The rademacher complexity of co-regularized kernelclasses. In *AISTATS*.

Sagae, K. and Lavie, A. (2006). Parser combination by reparsing. In *NAACL*.

Tesniére, L. (1959). *Eléments de syntaxe structurale*. Editions Klincksieck, Paris.

Wallsten, T. S., Budescu, D. V., Erev, I., and Diederich, A. (1999). Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making*, 10(3):243 – 268.