Some Empirical Evidence for Annotation Noise in a Benchmarked Dataset

Beata Beigman Klebanov
Kellogg School of Management
Northwestern University
beata@northwestern.edu

Eyal Beigman
Washington University in St. Louis
beigman@wustl.edu

Abstract

A number of recent articles in computational linguistics venues called for a closer examination of the type of noise present in annotated datasets used for benchmarking (Reidsma and Carletta, 2008; Beigman Klebanov and Beigman, 2009). In particular, Beigman Klebanov and Beigman articulated a type of noise they call *annotation noise* and showed that in worst case such noise can severely degrade the generalization ability of a linear classifier (Beigman and Beigman Klebanov, 2009). In this paper, we provide quantitative empirical evidence for the existence of this type of noise in a recently benchmarked dataset. The proposed methodology can be used to zero in on unreliable instances, facilitating generation of cleaner gold standards for benchmarking.

1 Introduction

Traditionally, studies in computational linguistics use few trained annotators. Lately this might be changing, as inexpensive annotators are available in large numbers through projects like Amazon Mechanical Turk or through online games where annotations are produced as a by-product (Poesio et al., 2008; von Ahn, 2006), and, at least for certain tasks, the quality of multiple non-expert annotations is close to that of a small number of experts (Snow et al., 2008; Callison-Burch, 2009).

Apart from the reduced costs, mass annotation is a promising way to get detailed information about the dataset, such as the level of difficulty of the difference instances. Such information is important both from the linguistic and from the machine learning perspective, as the existence of a group of instances difficult enough to look like they have been labeled by random guesses can in the worst case induce the machine learner training on the dataset to misclassify a constant proportion of easy, non-controversial instances, as well as produce incorrect comparative results in a benchmarking setting (Beigman Klebanov and Beigman, 2009; Beigman and Beigman Klebanov, 2009) .

In this article, we employ annotation generation models to estimate the types of instances in a multiply annotated dataset for a binary classification task. We provide the first quantitative empirical demonstration, to our knowledge, of the existence of what Beigman Klebanov and Beigman (2009) call "annotation noise" in a benchmarked dataset, that is, for a case where instances cannot be plausibly assigned to just two classes, and where instances in the third class can be plausibly described as having been annotated by flips of a nearly fair coin. The ability to identify such instances helps improve the gold standard by eliminating them, and allows further empirical investigation of their impact on machine learning for the task in question.

2 Generative models of annotation

We present a graphical model for the generation of annotations. The basic idea is that there are different types of instances that induce different responses from annotators. Each instance may have a true label of "0" or "1", however, the researcher's access to it is mediated by annotators who are guessing the true label by flipping a coin, where the bias of the coin depends on the type of the instance. The bias of the coin essentially models the difficulty of label-

Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 438–446,
Los Angeles, California, June 2010.

ing the instance; coins biased close to 0 and 1 correspond to instances that are easy to classify; a fair coin represents instances that are very difficult if not impossible to classify correctly with the given pool of annotators. The model presented in Beigman Klebanov and Beigman (2009) is a special case with 3 types (A,B,C) where p_A=0, p_C=1 (easy cases), and $0<p_B<1$ represents the hard cases, the harder the closer p_B is to 0.5. Models used here are a type of latent class models (McCutcheon, 1987) widely used in the *Biometrics* community (Espeland and Handelman, 1989; Yang and Becker, 1997; Albert et al., 2001; Albert and Dodd, 2004).

The goal of modeling is to determine whether more than two types of instances need to be postulated, to estimate how difficult each type is, and to identify the troublemaking instances.

The graphical model is presented in figure 1. We assume the dataset of size N is a mixture of k different types of instances. The proportion of types is given by $\theta = (\theta_1, \ldots, \theta_k)$, and coin biases for each type are given by $p = (p_1, \ldots, p_k)$. Each instance is annotated by n i.i.d coinflips, and random variable $x \in \{0, \ldots, n\}$ counts the number of "1"s in the n annotations given to an instance. Each instance belongs to a type $t \in \{1, ..., k\}$, characterized by a coin with the probability p_t of annotating with the label "1". Conditioned on t, the number of "1"s in n annotations has a binomial distribution with parameter p_t: $\Pr(x = j|t) = \binom{n}{j} p_t^j (1 - p_t)^{n-j}$.

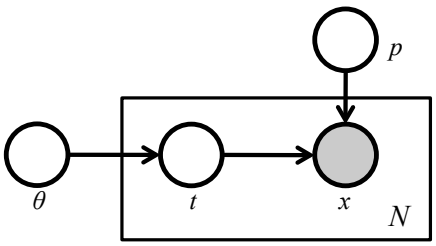

Figure 1: A graphical model of annotation generation.

The probability of observing j "1"s out of n annotations for an instance given θ and p is therefore $\Pr(x = j|\theta, p) = \sum_{t=1}^{k} \Pr(t|\theta) \cdot \Pr(x = j|t) = = \binom{n}{j} \sum_{t=1}^{k} \theta_t p_t^j (1 - p_t)^{n-j}$. The annotations are thus generated by a superposition of k binomials.

3 Data

3.1 Recognizing Textual Entailment -1

For the experiments reported here we use the 800 item test data of the first Recognizing Textual Entailment benchmark (RTE-1) from Dagan et al. (2006). This task drew a lot of attention in the community, with a series of benchmarks in 2005-2007.

The task is defined as follows: "*... textual entailment* is defined as a directional relationship between pairs of text expressions, denoted by *T* - the entailing "Text", and *H* - the entailed "Hypothesis". We say that *T* entails *H* if the meaning of *H* can be inferred from the meaning of T, as would typically be interpreted by people. This somewhat informal definition is based on (and assumes) common human understanding of language as well as common background knowledge" (Dagan et al., 2006). Further guidelines included an instruction to disregard tense differences, to accept cases where the inference is "very probable (but not completely certain)" and to avoid cases where the inference "has some positive probability that is not clearly very high." An example of a true entailment is the pair *T-H*: (T) Cavern Club sessions paid the Beatles £15 evenings and £5 lunchtime. (H) The Beatles perform at Cavern.

Although annotated by a small number of experts for the benchmark, the RTE-1 dataset has been later transferred to a mass annotation framework by Snow et al. (2008), who submitted simplified guidelines to the Amazon Mechanical Turk workplace (henceforth, AMT), collected 10 annotations per item from the total of 164 annotators, and showed that majority vote by Turkers agreed with expert annotation in 89.7% of the cases. We call the Snow et al. (2008) Turker annotations SRTE dataset, and use it in section 6. The instructions, followed by two examples, read: "Please state whether the second sentence (the Hypothesis) is implied by the information in the first sentence (the Text), i.e., please state whether the Hypothesis can be determined to be true given that the Text is true. Assume that you do not know anything about the situation except what the Text itself says. Also, note that every part of the Hypothesis must be implied by the Text in order for it to be true." The guidelines for Turkers are somewhat different from the original, not mentioning the issue of highly probable though not certain inference or a special treat-

ment of tense mismatch between *H* and *T*, as well as discouraging reliance on background knowledge.

Using Snow et al. (2008) instructions, we collected 20 annotations for each of the 800 items through AMT from the total of 441 annotators. Each annotator did the minimum of 2 items, and was paid $0.01 for 2 items, for the total annotator cost of $80. We used only annotators with prior AMT approval rate of at least 95%, that is, only people whose performance in previous tasks on AMT was almost always approved by the requester of the task. Our design is thus somewhat different from Snow et al. (2008), as we paid more and selected annotators with a stake in their AMT reputation.

3.2 Preparing the data for model fitting

We collected the annotations in two separate batches of 10 annotations per item, using the same set of instructions, incentives, and examples. We hypothesized that controlling for these elements, we would get two random samples from the same distribution of Turkers, and hence will have two samples to make sure a model fitted on one sample generalized to the other. It turned out, however, that a 3-Binomial model with a good fit on one of the samples was rejected with high probability for the other.[1] Thus, on the one hand, the variations between annotators in each sample were not as high as to preclude a model that captures only instance variability from fitting well; on the other hand, evidently, the two samples did not come from the same annotator distribution, but differed systematically due to factors we did not control for.[2] In order for our models not to inherit a systematic bias of any of the two samples, we mixed the two samples, and constructed two sets, BRTEa and BRTEb, each with 10 annotations per item, by randomly splitting the 20 answers per item into two groups, allowing the same annotator to contribute to different groups on different instances. Indeed, after the randomization, a model fitted for BRTEa produced excellent generalization on BRTEb, as we will see in section 4.2.

[1]For details of the model fitting procedure, see section 4.

[2]Such factors could be the hour and day of assignment, as the composition of AMT's global 24/7 workforce could differ systematically by day and hour.

4 Fitting a model to BRTE data

Using the model template presented in section 2, we successively attempt to fit a model with $k = 2, 3, \ldots$ until a model with a good fit is found or no degrees of freedom are left. For a given k, we fit the parameters θ and p using non-linear least squares trust-region method as implemented in the default version of MATLAB's *lsqnonlin* function. We then use χ^2 to measure goodness of fit; a model that cannot be rejected with 95% confidence (p>0.05) would be considered a good fit. In all cases N=800, n=10, as we use 10 annotations for each instance.

4.1 Mixture of 2 Binomials

Suppose k=2, with types t_0 and t_1. The best fit yields p_0=0.237, p_1=0.867, $\theta_0=\frac{431}{800}$, θ_1=1-θ_0. The model (shown in figure 2) is a poor fit, with χ^2=73.66 well above the critical value of 14.07 for df=7, p=0.05.[3]

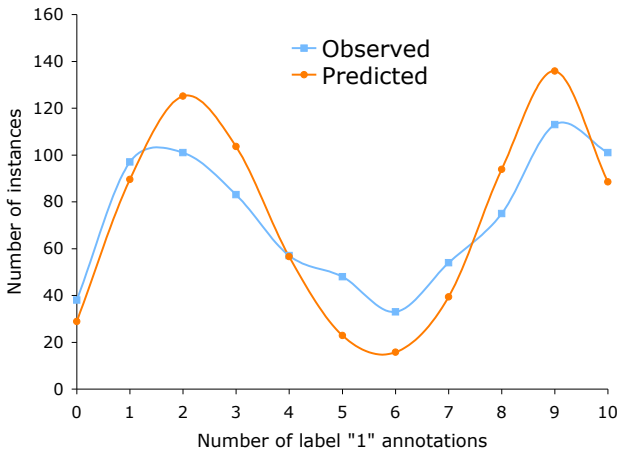

Figure 2: Fitting the model B1+B2 to BRTEa data. B1$\sim$ $\mathcal{B}$(10,0.237) on 431 instances, B2$\sim$ $\mathcal{B}$(10,0.867) on 369 instances. The point (x,y) means that there are y instances given label "1" in exactly x out of 10 annotations.

4.2 Model M: Mixture of 3 Binomials

Suppose now k=3. The best fitting model M=B1+B2+B3 is specified in figure 3; M fits the data very well. Assuming B1 and B3 reflect items

[3]For degrees of freedom, we take the number of datapoints being fitted (11), take one degree of freedom off for knowing in advance the total number of instances, and take off additional 3 degrees of freedom for estimating p_0, p_1, and θ_0 from the data. We are therefore left with 7 degrees of freedom in this case.

with uncontroversial labels "0" and "1", respectively, the model suggests that detecting "0" (no textual entailment) is somewhat more difficult for non-experts than detecting "1" (there is textual entailment) in this dataset, with the rate of incorrect predictions of about 20% and 10%, respectively.[4] The model also predicts that $\frac{159}{800} \approx 20\%$ of the data are difficult cases, with annotators flipping a close-to-a-fair coin (p=0.5487).

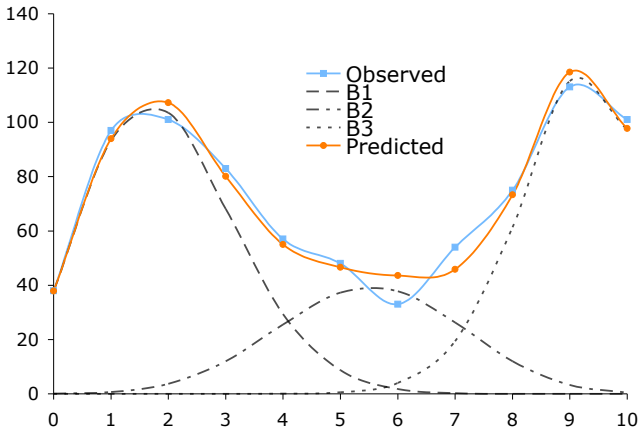

Figure 3: Fitting the model M=B1+B2+B3 to BRTEa data. B1$\sim$ $\mathcal{B}$(10,0.1978) on 343 instances, B2$\sim$ $\mathcal{B}$(10,0.5487) on 159 instances, B3$\sim$ $\mathcal{B}$(10,0.8942) on 298 instances. The binomials are shown in grey lines. The model M fits with χ^2=5.091; for df=5, this corresponds to p=0.4.

We use the dataset BRTEb to test the model developed on BRTEa. The model fits with χ^2=13.13, which, for df=10,[5] corresponds to p=0.2154.

We therefore conclude that, after eliminating systematic differences between annotators, we were unable to fit a model with two types of instances, whereas a model with three types of instances provides a good fit both for the dataset on which it is estimated and for a new dataset. This constitutes empirical evidence for the existence of a group of instances with near-random labels in this recently benchmarked dataset, at least for our pool of more than 400 non-expert annotators.

[4]We note that any conclusions from the model hold for the particular 800 item dataset in question, and not for the task of recognizing textual entailment in general, as the dataset is not necessarily a representative sample. In fact, we know from Dagan et al. (2006) that these 800 items are not a random sample, but rather what remained after some 400 instances were removed due to disagreements between expert annotators or due to the judgment of one of organizers of the RTE-1 challenge.

[5]No parameters are fitted using the BRTEb data.

5 Could annotator heterogeneity provide an alternative explanation?

In the previous section, we established that instance heterogeneity can explain the observations. We might however ask whether a different model could provide a similarly fitting explanation. Specifically, heterogeneity among annotators has been seen as a major source of noise in the aggregate data and there are several works attempting to separate high quality annotators from low quality ones (Raykar et al., 2009; Donmez et al., 2009; Sheng et al., 2008; Carpenter, 2008). Could we explain the observed behavior with a model with only two types of instances that allows for annotator heterogeneity?

In this section we construct such a model. We show that this model entails an instance distribution that is a superposition of two normal distributions. We subsequently show that the best fitting two-Gaussian model does not provide a good fit.

We use a generation model similar to those in (Raykar et al., 2009; Carpenter, 2008) but with weaker parametric assumptions. The graphical model is given in figure 4.

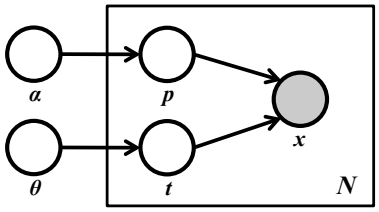

Figure 4: Annotation generation model with annotator heterogeneity.

We assume there are two types of instances $t \in \{0,1\}$ with the proportions $\theta = (\theta_0, \theta_1)$. The $2n$ probabilities $p = (p_{t1}, \ldots, p_{tn})$ for $t = 0,1$ correspond to coins drawn independently from some distribution with parameter $\alpha = (\alpha_1, \ldots, \alpha_n)$. We make no assumption on the functional form apart from a positive probability to draw a value between 0 and 1, this in particular is true for the beta distribution used in (Raykar et al., 2009; Carpenter, 2008). As before, the number of "1"s attributed to an instance of type t is a random variable x, determined

by independent flips of the n coins that correspond to the value of t. The marginal distribution of x is:

$$Pr(x = j|\theta, \alpha) =$$

$$= \sum_{t=0,1} \Pr(t|\theta) \int_{[0,1]^n} \Pr(p_t|\alpha) \cdot \Pr(x = j|p_t, t, \alpha) dp_t$$

$$= \sum_{t=0,1} \theta_t \int_{[0,1]^n} \Pr(p_t|\alpha) \left(\sum_{|S|=j} \prod_{i \in S} p_{ti} \prod_{i \notin S} (1 - p_{ti}) \right) dp_t$$

Let $x_1, \ldots, x_N$ be the random variables corresponding to the number of "1"s attributed to instances $1, \ldots, N$. W.l.g we assume instances $1, \ldots, N'$ are all of type t_0 ($N' = \theta_0 \cdot N$) and the rest of type t_1. Since $0 \leq x_j \leq n$ it follows that $\mathbb{E}(x_j), Var(x_j) < \infty$ for $j = 1, \ldots, N$. If for each instance the coin-flips are independent, we can think of this as a two step process where we first draw the coins and then flip them. Thus, $x_1, \ldots, x_{N'}$ are i.i.d and the central limit theorem implies that the average number of "1"s on t_0 instances, namely the random variable $y_0 = \frac{1}{N'} \sum_{j=1}^{N'} x_j$ has an approximately normal distribution.[6] Making the same argument for the distribution of y_1 for instances of type t_1, it follows that the number of "1"s attributed to an instance of any type $y = y_0 + y_1$ would have a distribution that is a superposition of two Gaussians.

The best least-squares fit of all two-Gaussian models to BRTEa data is produced by G=N1+N2, N1$\sim$ $\mathcal{N}$(2.22, 1.73) on 418 instances, N2$\sim$ $\mathcal{N}$(9.07,1.41) on 382 instances; G is shown in figure 5. G fits with χ^2=36.77, much above the critical value χ^2=11.07 for df=5, p=0.05. We can thus rule out annotator heterogeneity as the only explanation of the observed pattern of responses.

6 Testing M on SRTE data

We further test M on the annotations collected by Snow et al. (2008) for the same 800 item dataset. While the instructions and the task were identical in BRTEa, BRTEb, and BRTE datasets, and in all cases each item was given 10 annotations, the incentive design was different (see section 3).

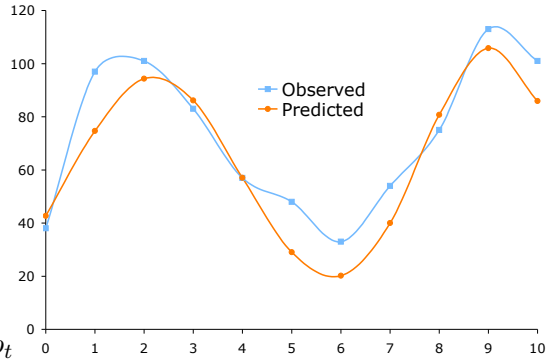

Figure 5: Model G's fit to BRTEa data, G= N1+N2, a mixture of two Gaussians.

Figure 6 shows that model M=B1+B2+B3 does not fit well, as SRTE dataset exhibits a rather different distribution from both BRTE datasets. In particular, it is clear that had a model been fitted on SRTE data, the coin flipping probabilities for the clear types, B1 and B3, would have to be moved towards 0.5; that is, an average annotator in SRTE dataset had worse ability to detect clear 0s and clear 1s than an average BRTE annotator. We note that BRTEa and BRTEb agreed with expert annotation in 92.5% and 90.8% of the instances, respectively, both better than 89.7% in SRTE.[7] Since we offered somewhat better incentives in BRTE, it is tempting to attribute the observed better quality of BRTE annotations to the improved incentives, although it is possible that some other uncontrolled AMT-related factor is responsible for the difference between the datasets, just as we found for our original two collected samples (see section 3.2).

Supposing the main source of misfit is difference in incentives, we conjecture that the difference between the 441 BRTE annotators and the 164 SRTE ones is due to the existence in SRTE of unmotivated, or "lazy" annotators, that is, people who flipped the same coin on every instance, no matter what type. Our hypothesis is that once an annotator is diligent (and motivated) enough to pay attention to the data, her annotations can be described by model M, but some annotators are not sufficiently diligent.

[6] It can be shown that $y_0 \sim \mathcal{N}(\mu, \sigma)$ for $\mu = n \cdot \mathbb{E}_{Dist(\alpha)}(p)$ and $\sigma = \sqrt{Var_{Dist(\alpha)}(p) \cdot n}$, using the expectation and variance of the coin parameter for type t_0 instances. For example, for a beta distribution with parameters α and β these would be $\mu = \frac{\alpha}{\alpha+\beta} n$ and $\sigma = \sqrt{\frac{\alpha\beta}{\alpha+\beta} n}$.

[7] Turker annotations were aggregated using majority vote, as in Snow et al. (2008) section 4.3.

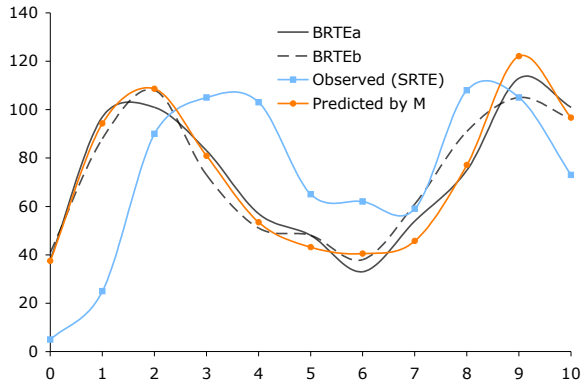

Figure 6: Model M's fit to SRTE data. BRTEa and BRTEb are shown in grey lines.

In this model we assume there are three types of instances as before, and two types of annotators $a \in \{D, L\}$, for Diligent and Lazy, with their proportions in the population $\xi = (\xi_D, \xi_L)$. The corresponding graphical model is shown in figure 7.

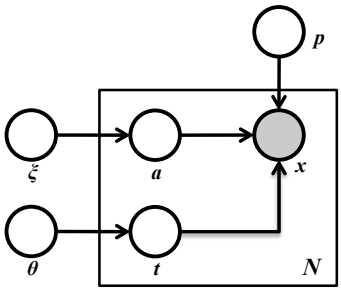

Figure 7: Annotation generation with diligent and lazy annotators.

We assume that diligent annotators flip coins corresponding to the types of instances, whereas lazy annotators always flip the same coin p_L.

Let n_D and n_L=$n - n_D$ be the number of diligent and lazy annotations given to a certain instance, thus $\Pr(n_D{=}r|\xi){=}\binom{n}{r}\xi_D^r\xi_L^{n-r}$, and the probability of observing j label "1" annotations for an instance of type t is given by:

$$\Pr(x = j|t, \xi, p) = \sum_{r=1}^{n} \left[\binom{n}{r} \xi_D^r \xi_L^{n-r} \times \right.$$

$$\times \left[\sum_{(j_1,j_2)\in S} \binom{r}{j_1} p_t^{j_1} (1-p_t)^{r-j_1} \times \right.$$

$$\left. \left. \times \binom{n-r}{j_2} p_L^{j_2} (1-p_L)^{n-r-j_2} \right] \right]$$

where S={(j_1, j_2):j_1+j_2=j; $j_1 \leq$r;$j_2 \leq$n-r}. Finally, $\Pr(x{=}j|\theta, \xi, p){=}\sum_{t=1}^{k} \theta_t \Pr(x{=}j|t, \xi, p)$.

We assume that model M provides the values for θ and p for all diligent annotators, and estimate ξ and p_L, the proportion of the lazy annotators and the coin they flip. The best fitting model yields ξ=(0.79,0.21), and p_L=0.74, predicting that about one-fifth of SRTE annotators are lazy.[8] This model fits with χ^2=14.63, which is below the critical level of χ^2=15.51 for df=8,p=0.05, hence a hypothesis that model M behavior for the diligent annotators and flipping a coin with bias 0.74 for the lazy ones generated the SRTE data cannot be rejected with high confidence. We note that Carpenter (2008) arrived at a similar conclusion – that there are quite a few annotators making random guesses in SRTE dataset – by means of jointly estimating annotator accuracies.

7 Discussion

To summarize our findings: With systematic differences between annotators smoothed out, there is evidence that non-expert annotators performing RTE task on RTE-1 test data tend to flip a close-to-fair coin on about 20% of instances, according to the best fitting model.[9] This constitutes, to our knowledge, the first empirical evidence for the existence of the kind of noise termed annotation noise in Beigman Klebanov and Beigman (2009). Given Beigman Klebanov and Beigman (2009) warning against annotation noise in test data and their finding in Beigman and Beigman Klebanov (2009) that annotation noise in training data can potentially devastate a linear classifier learning from the data, the immediate usefulness of our result is that instances of this difficult type can be identified, removed from the dataset before further benchmarking, and pos-

[8] A more precise statement is that there are about one-fifth lazy potential annotators in the SRTE pool for any given item. It is possible that the length of stay of an annotator in the pool is not independent of her diligence; for example, Callison-Burch (2009) found in his AMT experiments with tasks related to machine translation that lazy annotators tended to stay longer and do more annotations.

[9] Beigman Klebanov and Beigman (2009) discuss the connection between noise models and inter-annotator agreement.

sibly used in a controlled fashion for subsequent studies of the impact of annotation noise on specific learning algorithms and feature spaces for this task.

The current literature on generating benchmarking data from AMT annotations overwhelmingly considers annotator heterogeneity as the source of observed discrepancies, with instances falling into two classes only. Our results suggest that, at least in RTE data, instance heterogeneity cannot be ignored.

It also transpired that small variations in incentives (as between SRTE and BRTE), and even unknown factors possibly related to differences in the composition of AMT's workforce can lead to systematic differences in the resulting annotator pools, which results in annotations that are described by models with somewhat different parameter values. This can potentially limit the usefulness of our main finding, because it is not clear how reliable the identification of hard cases is using any particular group of Turkers. While this is a valid concern in general, we show in section 7.1 that many items consistently found to be hard by different groups of Turkers warrant at least an additional examination, as they often represent borderline cases of highly or not-so-highly probable inferences, corruption of meaning by ungrammaticality, or difficulties related to the treatment of time references and background knowledge.

Finally, our findings seem to be at odds with the fact that the 800 items analyzed here were left after all items on which two experts disagreed and all items that looked controversial to the arbiter were removed (see section 3). One potential explanation is that things that are hard for Turkers are not necessarily hard for experts. Yet it is possible that two or three annotators, graduate students or faculty in computational linguistics, are an especially homogenous and small pool of people to base gold standard annotations of the way things are "typically interpreted by people" upon. Furthermore, there is some evidence from additional expert re-annotations of this dataset that some controversies remain; we discuss relation to expert annotations in section 7.2.

7.1 Hard cases

We examine some of the instances that in all likelihood belong to the difficult type, according to Turkers. We focus on items that received between 4 and 7 class "1" annotations in SRTE and in each of our two datasets (before randomization).

(1) **T**: Saudi Arabia, the biggest oil producer in the world, was once a supporter of Osama bin Laden and his associates who led attacks against the United States. **H**: Saudi Arabia is the world's biggest oil exporter.

(2) **T**: Seiler was reported missing March 27 and was found four days later in a marsh near her campus apartment. **H**: Abducted Audrey Seiler found four days after missing.

(3) **T**: The spokesman for the rescue authorities, Linart Ohlin, said that the accident took place between 01:00 and dawn today, Friday (00:00 GMT) in a disco behind the theatre, where "hundreds" of young people were present. **H**: The fire happened in the early hours of Friday morning, and hundreds of young people were present.

(4) **T**: William Leonard Jennings sobbed loudly as was charged with killing his 3-year-old son, Stephen, who was last seen alive on Dec.12, 1962. **H**: William Leonard Jennings killed his 3-year-old son, Stephen.

Labeling of examples 1-4 seems to hinge on the assessment of the likelihood of an alternative explanation. Thus, it is possible that the biggest producer of oil is not the biggest exporter, because, for example, its internal consumption is much higher than in the second-biggest producer. In 2, abduction is a possible cause for being missing, but how relatively probable is it? Similarly, fire is a kind of accident, but can we infer that there was fire from a report about an accident? In 4, could the man have sobbed because on top of loosing his son he was also being falsely accused of having killed him? Experts marked all five as true entailments, while many Turkers had reservations.

(5) **T**: Bush returned to the White House late Saturday while his running mate was off campaigning in the West. **H**: Bush left the White House.

(6) **T**: De la Cruz's family said he had gone to Saudi Arabia a year ago to work as a driver after a long period of unemployment. **H**: De la Cruz was unemployed.

(7) **T**: Measurements by ground-based instruments around the world have shown a decrease of up to 10 percent in sunlight from the late 1950s to the early 1990s. **H**: The world is about 10 per cent darker than half a century ago.

In examples 5-7 time seems to be an issue. If Bush returned to White House, he must have left it beforehand, but does this count as entailment, or is the hypothesis referencing a time concurrent with the text, in which case T and H are in contradiction? In 6, can H be seen as referring to some time more than a year ago? In 7, if the hypothesis is taken to be stated in mid- or late-2000s, the time of annotation, half a century ago would reach to late 1950s, but it is possible that further substantial reduction occurred between early 1990s mentioned in the text and mid 2000s, amounting to much more than 10%. Experts labeled example 5 as false, 6 and 7 as true.

(8) **T**: On 2 February 1990, at the opening of Parliament, he declared that apartheid had failed and that the bans on political parties, including the ANC, were to be lifted. **H**: Apartheid in South Africa was abolished in 1990.

(9) **T**: Kennedy had just won California's Democratic presidential primary when Sirhan shot him in Los Angeles on June 5, 1968. **H**: Sirhan killed Kennedy.

Labeling examples 8 and 9 (both true according to the experts) requires knowledge about South African and American politics, respectively. Was the ban on ANC the only or the most important manifestation of apartheid? Was abolishing apartheid merely an issue of declaring that it failed? In 9, killing is a potential but not necessary outcome of shooting, so details of Robert Kennedy's case need to be known to the annotator to render the case-specific judgment.

(10) **T**: The version for the PC has essentially the same packaging as those for the big game consoles, but players have been complaining that it offers significantly less versatility when it comes to swinging through New York. **H**: Players have been complaining that it sells significantly less versatility when it comes to swinging through New York.

(11) **T**: During his trip to the Middle East that took three days, Clinton made the first visit by an American president to the Palestinian Territories and participated in a three-way meeting with Israeli Prime Minister Benjamin Netanyahu and Palestinian President Yasser Arafat. **H**: During his trip to the east of the Middle which lasted three days, the Clinton to first visit to American President to the occupied Palestinian territories and participated in meeting tripartite cooperation with Israeli Prime Minister Benjamin Netanyahu and Palestinian President, Yasser Arafat.

(12) **T**: The ISM non-manufacturing index rose to 64.8 in July from 59.9 in June. **H**: The non-manufacturing index of the ISM raised 64.8 in July from 59.9 in June.

(13) **T**: Henryk Wieniawski, a Polish-born musician, was known for his special preference for resurrecting neglected or lost works for the violin. **H**: Henryk Wieniawski was born in Polish.

Examples 10-13 were labeled as false by experts, possibly betraying over-sensitivity to the failings of language technology. *Sells* is not an ideal substitution for *offers*, but in a certain sense versatility is sold as part of a product. In 11-13, some Turkers felt the hypothesis is not too bad a rendition of the text or of its part, while experts seemed to hold MT to a higher standard.

7.2 Turkers vs experts

Model M puts 159 items in the difficult type B2. While M is the best fitting model, it is possible to find a model that still fits with $p>0.05$ but places a smaller number of items in B2, in order to obtain a conservative estimate on the number of difficult cases. The model with B1$\sim$ $\mathcal{B}(10, 0.21)$ on 373 items, B2$\sim$ $\mathcal{B}(10,0.563)$ on 110 items, B3$\sim$ $\mathcal{B}(10,0.89)$ on 327 items still produces a fit with $p>0.05$, but going down to 100 instances in B2 makes it impossible to find a good fit with a 3 type model. There are therefore about 110 difficult cases by a conservative estimate. Assuming there remain 110 hard cases in the 800 item dataset for which even experts flip a fair coin, we expect about 55 disagreements between the 800 item gold standard from RTE-1 and a replication by a new expert, or an agreement of $\frac{745}{800}$=93% on average. This estimate is consistent with reports of 91% to 96% replication accuracy for the expert annotations on various subsets of the data by different groups of experts (see section 2.3 in Dagan et al. (2006)).

Acknowledgments

We would like to thank the anonymous reviewers of this and the previous draft for helping us improve the paper significantly. We also thank Amar Cheema for his advice on AMT.

References

Paul Albert and Lori Dodd. 2004. A Cautionary Note on the Robustness of Latent Class Models for Estimating Diagnostic Error without a Gold Standard. *Biometrics*, 60(2):427–435.

Paul Albert, Lisa McShane, Joanna Shih, and The U.S. National Cancer Institute Bladder Tumor Marker Network. 2001. Latent Class Modeling Approaches for Assessing Diagnostic Error without a Gold Standard: With Applications to p53 Immunohistochemical Assays in Bladder Tumors. *Biometrics*, 57(2):610–619.

Eyal Beigman and Beata Beigman Klebanov. 2009. Learning with Annotation Noise. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics*, pages 280–287, Singapore.

Beata Beigman Klebanov and Eyal Beigman. 2009. From Annotator Agreement to Noise Models. *Accepted to Computational Linguistics*.

Chris Callison-Burch. 2009. Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk. In *Proceedings of the Empirical Methods in Natural Language Processing Conference*, pages 286–295, Singapore.

Bob Carpenter. 2008. Multilevel Bayesian Models of Categorical Data Annotation. *Unpublished manuscript*, last accessed 28 July 2009 at lingpipe.files.wordpress.com/2009/01/anno-bayes-entities-09.pdf.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL Recognising Textual Entailment Challenge. In J. Quiñonero Candela, I. Dagan, B. Magnini, and F. d'Alché-Buc, editors, *Machine Learning Challenges*, pages 177–190. Springer.

Pinar Donmez, Jaime Carbonell, and Jeff Schneider. 2009. Efficiently Learning and Accuracy of Labeling Sources for Selective Sampling. In *Proceedings of the 15th International Conference on Knowledge Discovery and Data Mining*, pages 259–268, Paris, France.

Mark Espeland and Stanley Handelman. 1989. Using Class Models to Characterize and Assess Relative Error in Discrete Measurements. *Biometrics*, 45(2):587–599.

Allan McCutcheon. 1987. *Latent Class Analysis*. Newbury Park, CA, USA: Sage.

Massimo Poesio, Udo Kruschwitz, and Jon Chamberlain. 2008. ANAWIKI: Creating Anaphorically Annotated Resources through Web Cooperation. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, Marrakech, Morocco.

Vikas Raykar, Shipeng Yu, Linda Zhao, Anna Jerebko, Charles Florin, Gerardo Hermosillo Valadez, Luca Bogoni, and Linda Moy. 2009. Supervised Learning from Multiple Experts: Whom to Trust when Everyone Lies a Bit. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 889–896, Montreal, Canada.

Dennis Reidsma and Jean Carletta. 2008. Reliability Measurement without Limits. *Computational Linguistics*, 34(3):319–326.

Victor Sheng, Foster Provost, and Panagiotis Ipeirotis. 2008. Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers. In *Proceedings of the 14th International Conference on Knowledge Discovery and Data Mining*, pages 614–622, Las Vegas, Nevada, USA.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and Fast - But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In *Proceedings of the Empirical Methods in Natural Language Processing Conference*, pages 254–263, Honolulu, Hawaii.

Luis von Ahn. 2006. Games with a Purpose. *Computer*, 39(6):92–94.

Ilsoon Yang and Mark Becker. 1997. Latent Variable Modeling of Diagnostic Accuracy. *Biometrics*, 53(3):948–958.