Estimation of Discourse Segmentation Labels from Crowd Data

Ziheng Huang
Department of Statistics
Columbia University
zh2220@columbia.edu

Jialu Zhong
Department of Computer Science
Columbia University
jialu.zhong@columbia.edu

Rebecca J. Passonneau
Center for Computational Learning Systems
Columbia University
becky@ccls.columbia.edu

Abstract

For annotation tasks involving independent judgments, probabilistic models have been used to infer ground truth labels from data where a crowd of many annotators labels the same items. Such models have been shown to produce results superior to taking the majority vote, but have not been applied to sequential data. We present two methods to infer ground truth labels from sequential annotations where we assume judgments are not independent, based on the observation that an annotator's segments all tend to be several utterances long. The data consists of crowd labels for annotation of discourse segment boundaries. The new methods extend Hidden Markov Models to relax the independence assumption. The two methods are distinct, so positive labels proposed by both are taken to be ground truth. In addition, results of the models are checked using metrics that test whether an annotator's accuracy relative to a given model remains consistent across different conversations.

1 Introduction

A single, spontaneous, spoken interaction can consist of multiple activities, such as to plan a future event, to complain about a past situation, or to carry out a transaction that might consist of subtasks. Speakers shift from one activity to the next with more or less awareness and explicit demarcation. To treat such conversational activities as a sequence of discrete units is a convenient oversimplification that is often resorted to (Bokaei et al., 2015; Galley et al., 2003; Passonneau and Litman, 1997). Systems that provide automated access to spoken language data often rely on segmentation of spoken discourse into sequential units for summarization (Wang and Cardie, 2012; Dielmann and Renals, 2005) or information retrieval (Ward et al., 2015). Research on the organization of spoken discourse also relies directly or indirectly on identification of such units to detect agreement among participants (Hillard et al., 2003; Somasundaran et al., 2007; Germesin and Wilson, 2009), multiparty meeting action items (Purver et al., 2007), decisions (Fernández et al., 2008), or answers to questions (Sun and Chai, 2007; Bosma, 2005). To support such research, there is a need for annotation methods to segment conversational interaction into sequential, multi-utterance units. We present and compare two methods to derive such data from crowdsourced annotations.

Crowdsourced annotation, where each item is labeled by a crowd of many independent annotators, is becoming more common in natural language processing. Examples include word sense (Bruce and Wiebe, 1999; Snow et al., 2008; Passonneau and Carpenter, 2014), named entities (Finin et al., 2010), and several other tasks in (Snow et al., 2008), including textual entailment. Three advantages to corpus annotation through application of a probabilistic model to crowdsourced labels, rather than reliance on interannotator agreement computed for a small number of trained annotators, are higher quality, lower cost, and a posterior probability for each ground truth label (Sheng et al., 2008; Snow et al., 2008; Passonneau and Carpenter, 2014). The latter serves as a confidence measure, which contrasts with interannotator agreement measures and with majority-voted labels, neither of which provides quality information for the ground truth labels on individual items. Previous work has demonstrated that model estimation of ground truth labels from crowd labels produces results superior to the crowd's majority vote, due to differences among annotators in the quality of their labels (Dawid and Skene, 1979; Snow et al., 2008; Passonneau and Carpenter, 2014). No previous work, however, provides model-based estimation of labels for sequential annotation from crowd labels.

For the discourse segmentation data presented here, annotators were presented with audio files of conversations and corresponding transcriptions into utterances. The annotation task was to identify each utterance that completes a discourse segment spanning one or more utterances, based on the speakers' conversational activities or intentions, as in (Passonneau and Litman, 1997). The annotations from y annotators for a conversation with x utterances can be represented as a $y \times x$ matrix, with cell values $n_{ij} \in \{0, 1\}$ to represent the binary segment boundary label assigned by annotator y_i at utterance x_j. Figure 1 illustrates part of such a matrix. The eight annotators for this conversation are on the y-axis and utterances 80 through 180 are on the

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2190–2200,
Lisbon, Portugal, 17-21 September 2015.

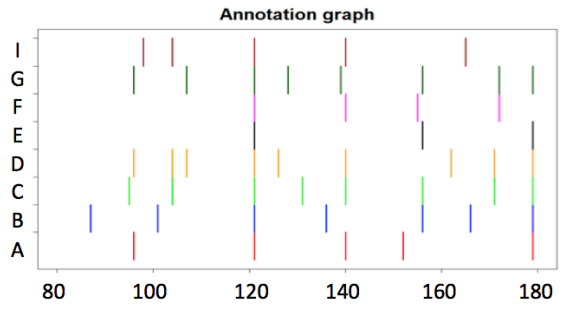


Figure 1: Annotation labels from eight annotators (A-G, I) on utterances 80 through 180 of a sample conversation. Vertical bars represent positive labels, with a different color for each annotator. Annotator H did not do this conversation.

x-axis. Colored bars represent positive labels, and each color represents a distinct annotator. The label distribution shown here is typical of our dataset: an annotator's positive labels are typically separated by several utterances, and annotators agree much more often on non-boundaries than on boundaries. Full consensus on a positive label is rare, but does occur. Here, all eight annotators assigned a positive label at utterance 120, six at utterance 178, and five at utterance 140.

Our work assumes that unobserved true labels condition the annotators' observed labels, and can be modeled as hidden states in a Markov-type process. Because an annotator rarely assigns positive labels for adjacent utterances, we assume that neither the true labels nor the observations are conditionally independent, and therefore are not generated by a simple Markov process. Our first model adapts the Double Chain Markov Model (Berchtold, 1999), designed to account for such cases. We then propose a second model that assumes that each annotator's labels are drawn from a Bernoulli distribution, that annotator performance is a parameter of the model, and that the state transitions are conditioned by an empirical distribution of discourse segment lengths. The two methods are quite distinct. Each thus serves as an evaluation of the other. The segment boundaries proposed by both models include all the majority vote cases, and in addition, cases voted on by a minority of relatively accurate annotators. We take segment boundaries proposed by both methods as ground truth. To further assess the results of the models, we assume that an annotator's accuracies should be consistent across the conversations she annotates.

2 Related Work

Previous work on annotation of discourse into linear segments has used a variety of methods to derive ground truth segment boundaries. In (Passonneau and Litman, 1997), seven annotators annotated narrative monologues for segments based on speaker intention. Agreement levels for ground truth boundaries were based on statistical significance using Cochran's Q. In (Galley et al., 2003), three annotators segmented the ICSI meeting corpus into topical units, and majority agreement was taken as ground truth. A functional segmentation of meetings from the AMI multiparty meeting corpus based on involved participants was segmented by one annotator and finalized by a second annotator (Bokaei et al., 2015). Task-based segmentation of patron-librarian interactions (Passonneau et al., 2011) measured agreement among two annotators using Krippendorff's Alpha at an average of 0.77 (Krippendorff, 1980). The annotation task here mostly closely resembles (Passonneau and Litman, 1997), and uses a similar number of annotators. No prior work, however, applies a probabilistic model to crowd labels for discourse segmentation.

Estimation of ground truth from crowd labels has been applied to many tasks, but is especially useful where judgments are subjective, making ground truth difficult to arrive at. Application areas include disease prevalence estimation (Albert and Dodd, 2008), identification of craters in images of Venus (Smyth et al., 1995), curation of biological data (Rzhetsky et al., 2009), computer vision (Whitehill et al., 2009), patient history (Dawid and Skene, 1979), and clinical reports (2010). Smyth et al. (1995), Rogers et al., and (2010) and Raykar et al. (2010) discuss the advantages of probabilistically annotated corpora over majority vote. Much of this work is motivated by the observation that annotators have different accuracies, and the fact that when annotators have known accuracies it can be shown that a majority of inaccurate annotators can be wrong (Raykar et al., 2010; Passonneau and Carpenter, 2014). Equally important, information from inaccurate annotators informs the model inference. For example, an inaccurate annotator might be biased towards label m whenever the true label is z.

Dawid and Skene (1979) present a joint model of true labels, observed labels, and annotator performance. Perhaps its first application to NLP data was the Bruce and Wiebe (1999) investigation of word sense. It has also been applied to more fine-grained word sense with a direct comparison to trained annotator labels in (Passonneau and Carpenter, 2014). Snow et al. (2008) showed that application of the same model to noisy crowd annotations produced data of equal quality to five distinct published gold standards. Hovy et al. (2013) apply a simple and effective model to identify untrustworthy annotators and test it on the same datasets used in (Snow et al., 2008). As they point out, when ties occur among an even number of annotators, it's necessary to resort to a tie-breaking procedure, e.g., for utterance 155 in Figure 1 where four annotators assign a positive label and four do not.

In experiments on an existing dataset of word sense annotation, Dligach et al. (2010) compare singly annotated data with doubly annotated adjudicated data, using trained annotators. They find that with the same amount of data, machine learning performance improves with the doubly annotated adjudicated data by

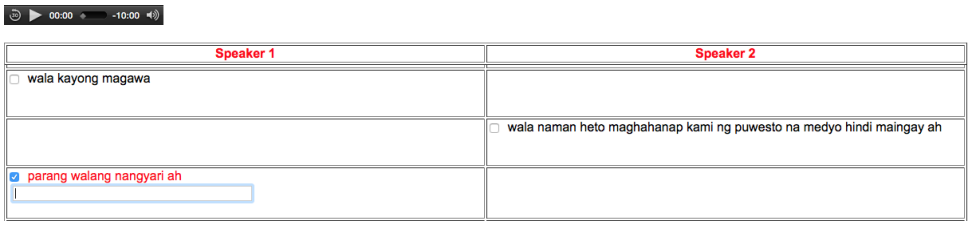


Figure 2: The annotation interface presented the audio control button on the upper left and the transcript below, with a scroll bar (not shown). Utterances from the two speakers are on the right and left sides, respectively. Each utterance had a checkbox; when selected, a textbox appeared to allow annotators to enter their segment descriptions.

a small amount, but that investing in more singly annotated labels leads to greater improvements. Their results on trained annotators, however, would not apply to our use case involving untrained annotators. In previous work, we found the cost per ground truth label of singly annotated data with trained annotators to be more than twice that for multiply annotated data with twenty untrained annotators (Passonneau and Carpenter, 2014). Half that many would have been sufficient for the Dawid & Skene model used there, which would reduce the cost by half again as much.[1]

3 Data and Annotation Task

The data consists of digital recordings and transcripts of fifty telephone calls between family members and friends who were native speakers of Tagalog. These were collected for the Babel program, sponsored by the Intelligence Advanced Research Projects Activity (IARPA). The calls ranged in length from about seven to ten minutes (μ = 9.67 minutes, σ=0.68 minutes). Transcripts provided by IARPA had an average of 364.66 utterances (min=239; max=475; σ=60.80).

The annotations were collected using Amazon Mechanical Turk. The task name and instructions were in English. The instructions were provided through a short video and text. Proficiency in Tagalog was assessed through a vocabulary test. Those who passed the vocabulary test were paid to do an initial annotation so we could ensure they understood the task. The initial task was based on a short Tagalog conversation that had been translated, annotated by a bilingual speaker of Tagalog and English, and verified by Passonneau. Annotators who understood the task and whose labels and descriptions seemed reasonable were admitted into the pool of annotators. A pool of nine annotators completed the qualifications. Each conversation was annotated by at least five annotators. Altogether, annotators assigned 5,567 labels to 164,097 utterances. Annotators' segments had a mean length of 21.85 utterances with a high standard deviation (σ = 19.32).

The interface designed for the annotation task is shown in Figure 2. Through the interface, annotators could read the transcript of a recorded conversation, and could play, pause or stop the audio. Each utterance had a checkbox for assigning a positive label if the annotator judged it to be the end of a segment. As shown, selection of a checkbox opened a text box for the annotator to enter a brief description of the segment. Table 1 in section 8 illustrates the descriptions assigned by six annotators to several segments.

4 Assumptions

Given the many labels from annotators, our goal is to estimate a ground truth label for each utterance position, where the label values represent a binary classification of segment boundaries. Our two models each assume there is a hidden *true* label that conditions an annotator's observed labels, and that can be estimated from the observed labels. How well the estimated ground truth fits the data thus depends on how well the model assumptions accord with the phenomenon of interest. The models do not account for annotator differences in the level of granularity they apply; cf. the contrast between lumpers and splitters in taxonomic classification of the natural world (Branch, 2014). Further, neither model takes linguistic features into account that annotators consider in deciding on segments, such as speaker attitude towards utterance content or speaker role in the conversational activity (Niekrasz and Moore, 2009). We find, however, much agreement between the two models on the proposed segment boundaries, and leave for future work the question of whether more complex models could accoount for differences in granularity or utterance features.

As discussed in section 2, we assume that annotators are not equally accurate, and that a probabilistic model based on the distribution of observed labels can do better than majority vote. Inspired by the type of probabilistic model proposed in (Dawid and Skene, 1979) and extended in (Bruce and Wiebe, 1999; Passonneau and Carpenter, 2014), annotator accuracy is a parameter of our second model. As described in detail in subsequent sections, the two models proposed here rely on distinct assumptions and inference methods. They nevertheless propose many of the same labels. We take each model to provide independent evidence for the ground truth labels, thus the final labels are those voted on by both models.

[1] Twenty labels per item were collected in order to provide tight estimates for item difficulty. This, however, requires a model with a parameter for item difficulty, which had not yet been implemented for this data.

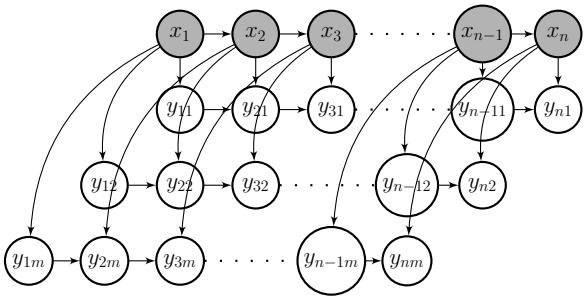


Figure 3: Graphical model of Double Chain Dynamic Hidden Markov Model for a conversation with m annotators and n utterances. The x_t are the hidden states, and the y_{jl} are the observed labels from annotator l at utterance j.

In addition, we assume that annotators' accuracies should be relatively consistent across conversations, and we measure how well each model's results support this assumption. We base the assumption on the observation that the annotation task is the same for all conversations, and an annotator's relative ability to do the task should not change significantly. The annotators all had the same initial training, and did about the same number of conversations. The conversations all had similar conditions of collection, similar participants, and similarly mundane topics and conversational activities that most annotators would be familiar with. The subjects that were discussed included parties, watching tv, siblings, money, jobs, spouses, medical issues, birthdays, and so on.

5 Double Chain Dynamic Hidden Markov Model

The first model we propose combines the Double Chain Markov Model (Berchtold, 1999) and dynamic Bayesian networks (Martinez and Sucar, 2008). The double chaining involves the dependence of observations on immediately prior observations. Figure 3 shows that for all $y_{tl}, t \geq 2$, observation y_{tl} depends on observation $y_{(t-1)l}$. The emission matrix at the first utterance x_1 is thus a 2×2 matrix, while all subsequent emission matrices are $2 \times 2 \times 2$. As in (Martinez and Sucar, 2008), the observed states can be regarded as a composition of m independent chains, where m is the number of annotators for the conversation. Also, the l^{th} annotator's observation at the t^{th} utterance depends not only on the same hidden state x_t, but also on the last observation $y_{(t-1)l}$.

Assume in a conversation, there are m annotators and n utterances. The model $\Theta = \{\pi, \gamma, A, B\}$ can be described as follows:

- a set of hidden states, i.e the true labels: $x_t \in \{0, 1\}$, $t \in \{1, 2, \ldots, n\}$. $x_t = 1$ represents the t^{th} utterance is a true boundary and 0 otherwise;
- a set of observed variables: $y_{tl} \in \{0, 1\}$, $l \in \{1, 2, \ldots, m\}$ annotators, $t \in \{1, 2, \ldots, n\}$ utterances. $y_{tl} = 1$ represents that the l^{th} annotator annotates t^{th} utterance to a true boundary and 0 otherwise;
- Θ is a vector of parameters. To be more specific, the elements are:
  - the probability of the initial hidden state: $\pi_{x_1}, x_1 \in \{0, 1\}$. Note $\pi_0 + \pi_1 = 1$.
  - the probabilities of the initial emission matrix. Note that the initial emission matrix is a 2×2 matrix: $\gamma_l \in \{c_{x_1, y_{1l}}\}$, $x_1, y_{1l} \in \{0, 1\}, l \in \{1, 2, \ldots, m\}$. For annotator l, $c_{x_1, y_{1l}}$ is the probability of emitting from x_1 to y_{1l}.
  - the transition matrix between hidden states, $A \in \{a_{x_{t-1}, x_t}\}$, $x_{t-1}, x_t \in \{0, 1\}$, $t \in \{2, 3, \ldots, n\}$. a_{x_{t-1}, x_t} is the probability of transitioning from x_{t-1} to x_t.
  - the emission matrices, $B_l \in \{b_{x_t, y_{(t-1)l}, y_{tl}}\}$, $x_t, y_{(t-1)l}, y_{tl} \in \{0, 1\}, l \in \{1, 2, \ldots, m\}$, $t \in \{2, 3, \ldots, n\}$. Note that the emission matrix is a $2 \times 2 \times 2$ matrix as each observed state depends on current hidden state as well as the previous observation, i.e., $b_{x_t, y_{(t-1)l}, y_{tl}}$ is the probability of emitting from x_t to y_{tl} and transitioning from $y_{(t-1)l}$ to y_{tl}.

A graphical sketch of the DCD HMM model is shown in Figure 3. The target function $F = P(x, y|\Theta)$ is:

$$F = \pi_{x_1} \prod_{l=1}^{m} c_{x_1, y_{1l}} \prod_{t=2}^{n} a_{x_{t-1}, x_t} \prod_{l=1}^{m} \prod_{t=2}^{n} b_{x_t, y_{(t-1)l}, y_{tl}}$$

We can derive a marginal distribution over y and have the *likelihood* as:

$$L(\Theta) = P(y|\Theta) = \sum_{x} P(x, y|\Theta)$$

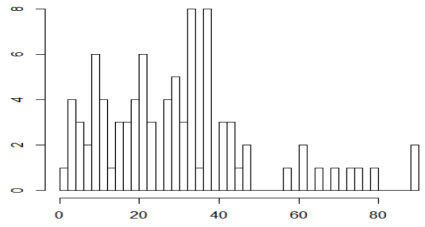
Our goal is to find the parameters (Θ) that maximize the above function. Bayes Net Toolbox for Matlab (Murphy, 2001) is used for the inference. Expectation-Maximization (EM) with Junction Tree inference for the E-step is used for learning the parameters. The Junction Tree Algorithm is a method to calculate marginals by propagation on the graph. It runs as follows: 1) Initialize: Pick a proper root and initialize all variables; 2) Collect: Pass message from each child of a node through separators to the parent node and update the node with collected evidence; 3) Distribute: Send back message to each child of the node through separators and update the child with distributed evidence; 4) Normalize: Normalize cliques connected by a separator so they agree with each other: e.g., for {AB} and {BC}, if we have $\sum_A \{AB\} = \sum_C \{AB\}$, propagation is complete.

After convergence from EM, junction tree propagation is again used for inference, and the model produces

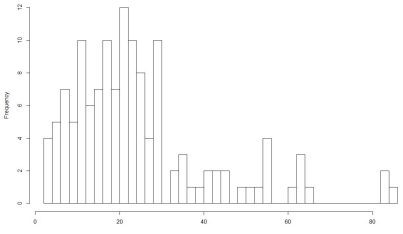
a probability for each ground truth label. We take the label to be positive if the posterior probability is greater than 0.5; as shown in section 8, probabilities tend to be very high or very low.

6 Interval-dependent HMM

The second model, Interval-dependent HMM, imposes a constraint on the state transitions between two positive labels based on the empirical distribution of intervals between observed labels. Initially, we examined known distributions. The Poisson, for example, represents the probability of events in an interval as an average rate. The model based on the Poisson did not perform particularly well. Histograms of interval sizes from different conversations have similar shapes, however, as illustrated in Figure 4. Although more of the probability is towards 20 to 40 utterances in Figure 4a, and between 15 and 35 utterances in Figure 4b, we assume these small differences in the two distributions are mainly due to sampling variation. As discussed in preceding sections, the model we present here assumes that the probability of a true label at time t_i is a function of the interval length $t_i - t_j$, where t_j is the most recent time of a true label. The observed data for all annotators on all conversations provides a set of time intervals to construct the empirical distribution.



(a) First sample conversation



(b) Second sample conversation

Figure 4: Histograms of interval lengths between all observed labels for two conversations.

To assess whether we have sufficient data to reliably construct the empirical distribution, we performed fifty iterations of random divisions of the data into two samples. For each pair of samples, we measured the maximum distance between pairs of cumulative distribution function (CDF) curves, and used the two-sample Kolmogorov-Smirnov test to measure the goodness of fit of the two curves. Figure 5 shows an example comparison of two CDF curves which have a maximum gap

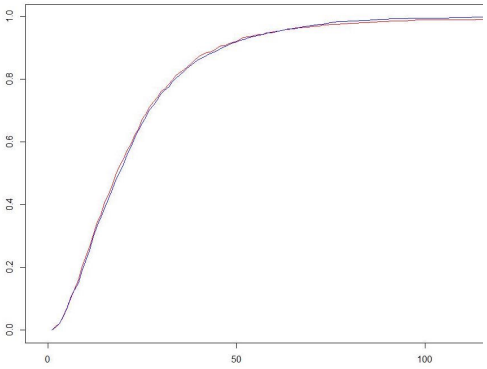


Figure 5: A plot of two CDF curves for a random split of the data. The curves are almost identical; the maximum gap is 0.0175. A two sample K-S test has a p-value of 0.79.

of 0.0175 and a K-S p-value of 0.7866. The mean maximum distance between pairs of curves was 0.014, with a standard deviation of 0.009, both of which are quite small. The p-values for the K-S test ranged from 0.4 to 0.96, which fail to reject the hypothesis that the pairs of samples are from the same distribution. While the two measures are not conclusive evidence that we have sufficient data to construct the empirical distribution, they are supportive. Further, reliance on estimates of the empirical distribution are preferable to a known distribution that does not fit the data, such as the Poisson.

The model can be described as follows:

- the observations $Y_{ij} \in 0, 1, i \in 1, 2, \cdots, N, j \in 1, 2, \cdots, J$;
- the true labels $Z_i \in 0, 1, i \in 1, 2, \cdots, N$;
- the 2×2 annotator performance matrices B_j;
- the initial state probability π

Given N utterances, J annotators, the initial state probability π and four cells in each annotator's performance matrix B_j, where B_{j11} represents the true positives (the probability that given a ground truth positive label, annotator a_j assigns a positive label), B_{j10} represents false negatives, B_{j01} represents false positives, and B_{j00} represents true negatives. $\pi = 1$ is the probability that the first hidden state is a boundary and $\pi = 2$ means it is not. Our objective is to find the parameter vector $\theta = (\pi, B)$ that maximize the likelihood $P(Y|\theta)$, and to use this θ to estimate the true labels Z.

To solve:

$$\underset{\theta}{\text{Argmax}} \log\left[P(Y|\theta)\right] = \underset{\theta}{\text{Argmax}} \log\left[\sum_{Z} P(Y, Z|\theta)\right]$$

we use expectation-maximization (EM).

E step First, we should find the lower bound of our optimization object: $\underset{\theta}{\text{Argmax}} \log\left[\sum_{Z} P(Y, Z|\theta)\right]$; by

Jensen's inequality, we have:

$$\log\left[\sum_Z P(Y,Z|\theta)\right] = \log\left[\sum_Z \frac{P(Y,Z|\theta)}{Q_\theta(Z)} Q_\theta(Z)\right] \geq \sum_Z Q_\theta(Z)\log\left[\frac{P(Y,Z|\theta)}{Q_\theta(Z)}\right]$$

$Q_\theta(Z)$ is a function of θ which satisfies that $\sum_Z Q_\theta(Z) = 1$. The equality holds if and only if

$$\frac{P(Y,Z|\theta)}{Q_\theta(Z)} = c \qquad \text{for all Z}$$

Note that c is a constant. In the E step we need to calculate the Q function to maintain the equality. By straightforward algebra, we get $Q_\theta = P(Z|Y,\theta)$.

M step In this part, we should maximize our lower bound:

$$\underset{\theta}{\text{Argmax}} \sum_Z Q_{\theta^{(n)}}(Z)\log\left[\frac{P(Y,Z|\theta)}{Q_{\theta^{(n)}}(Z)}\right]$$

Since $\log[Q_\theta(Z)]$ is a term not related to θ, $P(Z|Y,\theta) \propto P(Z,Y|\theta)$. Our problem becomes:

$$\underset{\theta}{\text{Argmax}} \sum_Z P(Y,Z|\theta^{(n)})\log[P(Y,Z|\theta)]$$

$\theta^{(n)}$ is the parameter we get from the last iteration, and the Q function is fixed in this M step. We cannot use the forward-backward algorithm to optimize, because the first order Markov property does not hold: $P(Z_i = 1)$ is a function of the last positive label $Z_j = 1$ at time j such that $j < i$, and for all k such that $j < k < i$, $Z_k = 0$. To make use of the Markov property, we rely on a hidden variable U_i to save the interval length between i and j. The hidden parameter space is then expanded to $X_t = (Z_t, U_t)$, where U_t denotes the size of the interval between the current position t_i and the most recent t_j with a positive label. If the true label $Z_{t_i} = 0$, then $U_{t_i} = t_i - t_j$, and if $Z_{t_i} = 1$, then $U_{t_i} = 0$. This gives $t+1$ possible states for each t: the t states for $Z_t = 0$, and one state for $Z_t = 1$.

In this problem, given a length N conversation, there are $N+1$ hidden states at each moment. $X_t = 1$ means $(Z_t = 1, U_t = 0)$, $X_t = 2$ means $(Z_t = 0, U_t = 1)$, $X_t = 3$ means $(Z_t = 0, U_t = 2)$, and so on.

The transition matrix at each t for the cases represented by $P(X_t = k|X_{t-1} = l)$, which is with size $(t+1)\times(t+2)$, will necessarily be very sparse. For example, given an empirical function $f(n) = P(x = n|x \geq n)$, the transition matrix from $t = 4$ to $t = 5$ can be written:

$$\begin{pmatrix} f(1) & 1-f(1) & 0 & 0 & 0 & 0 \\ f(2) & 0 & 1-f(2) & 0 & 0 & 0 \\ f(3) & 0 & 0 & 1-f(3) & 0 & 0 \\ f(4) & 0 & 0 & 0 & 1-f(4) & 0 \\ f(5) & 0 & 0 & 0 & 0 & 1-f(5) \end{pmatrix}$$

After this transformation, X_{t+1} is independent to all X_k for any $k < t$ provided that X_t is given. With X as the new hidden state, we can estimate the HMM parameter by adding some constraints. Replacing the Z in the object function with X, we can rewrite the object function as:

$$\begin{aligned}
&\sum_X P(Y,X|\theta^{(n)})\log[P(Y,X|\theta)] \\
=&\sum_X P(Y,X|\theta^{(n)})\left[\text{Log}P(X_1) + \sum_{t=1}^{N-1}\log P(X_{t+1}|X_t) + \sum_{t=1}^{N}\log P(Y_t|X_t)\right] \\
=&\sum_X P(Y,X|\theta^{(n)})\left[\log[\pi_{X_1}] + \sum_{t=1}^{N-1}\log\left[A_{X_t,X_{t+1}}\right] + \sum_{t=1}^{N}\log[B_{X_t,Y_t}]\right]
\end{aligned}$$

The object is split into three independent parts: the first part is for the initial state distribution π, the second for the transition probability matrix A, and the third is the emission matrix B. For the first term, because in the moment $t = 1$, X_t can just be 1 or 2, we have the optimization problem:

$$\begin{aligned}
&\underset{\pi}{\text{Argmax}} \sum_{i=1}^{2} P(Y, X_1 = i|\theta^{(n)})\log[\pi_i] \\
&\text{s.t} \qquad \pi_1 + \pi_2 = 1 \\
&\qquad \pi_3 = \pi_4 = ... = \pi_{N+1} = 0
\end{aligned}$$

We can easily solve this optimization problem by the Lagrange multiplier: we have the update formula:

$$\pi_1^{(n+1)} = P(X_1 = 1|Y,\theta^{(n)})$$

$$\pi_2^{(n+1)} = P(X_1 = 2|Y,\theta^{(n)})$$

$$\pi_i^{(n+1)} = 0 \qquad \text{for} \quad i > 2$$

Both can be solved by the traditional forward-backward algorithm after this transformation. $\theta^{(n)}$ is the parameter set we get from the last iteration.

The second term can be ignored, since we use the known empirical distribution as the transition matrix; it is therefore a constant term.

The third term can be rewritten as:

$$\begin{aligned}
&\sum_{t=1}^{N} P(Y,X|\theta^{(n)})\log[B_{X_t,Y_t}] \\
=&\sum_{t=1}^{N}\sum_{i=1}^{N+1}\sum_{j=1}^{J}\sum_{k=0}^{1} I(Y_{t,j} = k)P(X_t = i, Y|\theta^{(n)}) \\
&\log[B_{j,i,k}]
\end{aligned}$$

So our problem is:

$$\underset{B}{\text{Argmax}} \sum_{t=1}^{N} \sum_{j=1}^{J} \sum_{i=1}^{N+1} \sum_{k=0}^{1} I(Y_{t,j} = k)$$
$$P(X_t = i, Y|\theta^{(n)}) \log B_{j,i,k}$$
$$\text{s.t} \quad \sum_{k=0}^{1} B_{j,i,k} = 1 \qquad \text{For all i, j}$$
$$B_{j,i_1,k} = B_{j,i_2,k} \qquad \text{For all j, k and } i_1, i_2 \geq 2$$

The second constraint here means that, if this is not a true boundary, a given annotator j will have the same emission matrix no matter what U is. This optimization can also be solved by Lagrange multiplier, where the update formula is as follows. For $i = 1$:

$$B_{j,i=1,k}^{(n+1)} = \frac{\sum_{t=1}^{N} P(Y, Z_t = 1|\theta^{(n)}) I(Y_{t,j} = k)}{\sum_{t=1}^{N} P(Y, Z_t = 1|\theta^{(n)})}$$

For any $i \neq 1$, the matrix B is the same given j:

$$B_{j,i\neq 1,k}^{(n+1)} = \frac{\sum_{t=1}^{N} P(Y, Z_t \neq 1|\theta^{(n)}) I(Y_{t,j} = k)}{\sum_{t=1}^{N} P(Y, Z_t \neq 1|\theta^{(n)})}$$

Now we have the update function for θ. After convergence, we will have π and B. It is straightforward to transfer these parameters for the new space to our original HMM problem. This completes the M step.

7 Model Checking

No ground truth labels are available to evaluate our models. We check the model results, however, in three ways. One, we consider labels proposed by both models to be stronger evidence than labels proposed only by one. Two, we measure the consistency of annotators on the assumption that the same annotator should have relatively consistent performance across conversations, relative to the same model. The third way we can check the models is to examine the descriptive labels that annotators assign to segments to determine whether descriptions for the same segment from different annotators are consistent. In this section, we describe the two consistency metrics.

We measure how consistently the label quality from annotator a_i surpasses that for $a_j, i \neq j$, for all pairs of annotators using a metric to measure inconsistency and strength of inconsistency (I&SI) (de Vries, 1998). We also apply a variant we refer to as Directional Consistency (DC), which takes into account how often annotator a_i surpasses annotator a_j. To measure annotators' performance relative to the inferred true labels, we use F-score, the harmonic mean of recall and precision. Recall is the ratio of true positives to the sum of true positives and false negatives; precision is the ratio of true positives to the sum of true positives and false positives. A square matrix of annotator *dominance* is first constructed to give a count of how many conversations there are where a_i has a higher F measure than $a_j, i \neq j$. A linear dominance ordering $>$ of all annotators has an inconsistency score I that is incremented by 1 for each pair of annotators where $a_i > a_j$ in the linear ordering and $(a_i, a_j) \neq (a_j, a_i)$ in the matrix. I is minimal if no other ordering has fewer inconsistencies. The strength of the inconsistency IS for a linear ordering is incremented by the difference in rank between a_i and a_j for every inconsistent pair in the linear ordering. The I&SI method finds an ordering that minimizes I and SI. To check the results of our models, we compare the I&SI value of the dominance matrix associated with the model results against a simulated random matrix. If the model results are significantly more consistent than the simulation, the model produces a consistent ranking of annotators.

We propose a Directional Consistency index ($DC \in [0, 1]$) which considers the number of times a_i has a higher F measure than a_j (Leiva et al., 2008). Where X is the dominance matrix:

$$DC = \frac{\sum_{i=1}^{n} \sum_{j=i+1}^{n} |x_{ij} - x_{ji}|}{N}$$
$$N = \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} x_{ij}$$

DC values closer to zero indicate less consistency in differences among annotators, and the converse for values closer to 1. High DC values for the results of our models thus indicates better performance of the model in predicting consistent annotator behavior.

8 Results and Model Checking

The results consist of the true labels assigned by each model to each conversation, and estimates of the annotators' performance relative to the model's ground truth labels. Note that as the conversation is not a parameter of either model, after estimation of the empirical distribution of segment lengths, the data for each conversation is treated separately.

To provide a concrete illustration, we first review the data for a typical conversation. Table 1 presents the segments derived from both models for an extract from conversation 945, which had six annotators, and the annotator's segment descriptions. We selected a conversation with an even number of annotators to illustrate that an arbitrary choice must be made, given a 50/50 vote split. We take ties as true positives to provide a more conservative baseline. We first discuss this example conversation in detail to explain the kinds of cases where the models differ from majority voting. Then we present summary results on the fifty conversations for majority voting compared with the two models.

In Table 1 a description at n gives the annotator's interpretation of the kind of conversational activity that ends with n. When annotators agree on a positive label that ends a segment, they might not agree on the utterance that starts the segment, so their descriptions will not necessarily be about the same segments. From

Utt	Description
191	C: S1 and S2 are talking about the status of their children's studies I: S1 and S2 are talking about their children's education
216	A: S1 and S2 spoke about their children's studies E: S2 then shared that he's going to Laguna-Muntinlupa tomorrow. S1 said that S2 has many orders. S2 shared that he's striving hard in order for her kids to graduate college. . . . The two laughed at each other about S1's children not getting traits from S1 I: S1 and S2 are talking about who their children took after
217	C: S1 and S2 are joking about the traits their children got from them D: They are talking about S1s daughter that she is good at academics and that she got her being smart from her mom and nothing from S1. S1 said even if she got nothing from him as long as she will just study hard its okay
241	A: S1 and S2 spoke about their time of sleeeping C: S1 and S2 tell each other what time they usually go to sleep D: They are talking about the time that they go to sleep. S2 said sometimes by ten, eleven or twelve midnight. S1 said sometimes he goes out one in the morning. Sometimes he goes to sleep at ten or eleven in the evening too

Table 1: Annotator descriptions for conversation 945 for a sequence of four segment boundaries hypothesized by both models. A description from annotator a_i at utterance n indicates a_i assigned a positive label, and gives the annotator's interpretation of the kind of interaction that ended at utterance n. Underlined utterance numbers indicate cases where at least six annotators assigned a positive label.

the table, however, we see a a pattern that is consistent for most of the data: abstracting over the descriptions gives a good indication of what's going on in the segments that are defined by the positive labels assigned by both models. The descriptions from C and I at 191, for example, describe the first segment as the speakers talking about their children's education. A's similar description at 216 indicates that A ended the segment later than C and I. E and I describe the second segment as being about the children, including who they take after. C's description about who the children take after occurs at a later utterance. The third segment goes into detail about the children's traits, and the fourth is about what time the speakers go to sleep.

Across all fifty conversations, ID HMM assigns more positive labels than the majority, and DCD HMM assigns more than ID HMM. Totals for each labeling

Method	Total
Majority	683
ID HMM	991
DCD HMM	1324

Table 2: Total positive labels assigned by each method.

Utt	Annotators	DCD HMM	ID HMM
11	2 (A,I)	1.00	0.99
42	3 (B,E,I)	1.00	1.00
43	3 (A,C,D)	1.00	1.00
67	6 (A,B,C,D,E,I)	1.00	1.00
114	2 (A,C)	1.00	0.98
126	2 (D,E)	1.00	1.00
127	1 (C)	**0.63**	*0.02*
144	2 (A,D)	1.00	0.98
147	2 (C,I)	0.90	0.65
191	2 (C,I)	0.90	0.73
216	3 (A,E,I)	1.00	1.00
217	2 (C,D)	1.00	0.66
241	3 (A,C,D)	1.00	1.00
276	3 (B,C,D)	1.00	1.00
282	1 (A)	**1.00**	*0.29*
300	*5* (A,B,D,E,I)	1.00	1.00
356	2 (C,I)	**0.98**	*0.27*
357	4 (A,B,D,E)	1.00	1.00

Table 3: Comparison of positive predictions from majority voting (N=8, underlined; ties are taken as positive), DDC HMM (N=18), and the ID HMM (N=15) for conversation 945. Probabilities in bold are for boundaries proposed by only one model; italics are for probabilities below the 0.5 threshold to be considered true boundaries.

method are in Table 2. Wherever the majority vote predicts a true label, both models always do. If ID HMM posits a boundary at an utterance, DCD HMM also does, but DCD HMM predicts additional ones. Because all the ID HMM labels are also identified by DCD HMM, these are the final labels we propose.

Table 3 shows the positive labels predicted for conversation 945 by majority vote, and by our two models. Column one is the utterance number, and again, underlining indicates cases where the voted baseline would assign a positive label. Column two lists the annotators who assigned a positive label, and columns three and four show the posteriors assigned by the two models; for all utterances not listed in the table, the posteriors are below 0.5. Low posteriors for ID HMM where DCD HMM proposed a boundary are in italics.

Ann	Maj	DCD HMM	ID HMM
A	0.68	0.71	0.68
B	0.57	0.36	0.38
C	0.40	0.63	0.46
D	0.59	0.56	0.55
E	0.73	0.50	0.52
I	0.43	0.55	0.49

Table 4: F-measure for annotators in conversation 945 for majority vote labels and both models; recall that the true labels for each model are different, and that DCD HMM hypothesizes more true labels than ID HMM.

For each model, the annotator can be ranked by the F-scores relative to the model predictions. When one of the models agrees with a minority of annotators,

Model	I&SI	DC
Majority	I=1, SI=3, p=0.008	p=0.0600
DCD HMM	I=2, SI=5, p=0.02	p=0.0014
ID HMM	I=0, SI=0, p=0	p=0.0001

Table 5: Consistency of annotators

the minority consists of the annotators considered by the model to have higher performance, as given by F-measure. The three sets of F-scores for the six annotators in 945 are shown in Table 4. Annotator performance given the two models is very similar; the Pearson correlation is 0.80. F-scores based on the majority baseline, however, do not correlate well with DCD ($\rho = -0.5$) or ID ($\rho = 0.49$). In eight cases where DCD posits a true label for conversation 945, and only 2 annotators voted positive, the pair never includes B, the least accurate annotator by DCD (see Table 4), and always includes one of the top three annotators (A,C,D). In the two cases where only one annotator voted positive, it was A or C, one of the two top DCD HMM annotators. Both models consider A to be the best annotator. C is relatively good in the DCD HMM model and relatively poor in the ID HMM model.

I&SI tests whether there exists a linear ordering of the annotators such that their relative performance across conversations is consistent. DC tests whether an ordering $a_i > a_j$ is based on relatively more frequent dominance of a_i over a_j. Table 5 shows that majority vote and the two models produce results that lead to high I&SI consistency, based on the statistically significant p-values. The majority vote p-value for DC, however, is not statistically significant. By the more stringent DC measure, the labels from the two HMM variants are superior to the majority vote labels.

The list of descriptions from annotators at utterance n represents the semantics of the hypothesized segment ending at n. Semantic consistency for a given segment serves as another check on the output of the model, because the human descriptions of the activity within the segment do not conflict. In general, this is the case for both models, but less so for DCD HMM. For conversation 945 illustrated in Table 3, there are three positive labels proposed by DCD HMM that are missing from the ID HMM predictions. These are at 127 where only annotator C had a positive label, 282 where only annotator A had a positive label, and 356 where annotators C and I had a positive label. Annotators B, C and D, for example, describe a segment ending at utterance 276 as the speakers discussing Facebook, whereas annotator A locates the end of the Facebook segment at utterance 282. The DCD HMM model posits a true label at 276 but not at 282, in contast to ID HMM.

9 Discussion

The two models for estimating ground truth labels from crowd labels advance previous work on probabilistic models for annotation by handling sequential data. We have argued that for our data, the Markov assumption must be relaxed. The two models handle this in distinct ways. The first model assumes that each state can be decomposed into multiple aspects, and that states and observations are conditionally dependent on the previous point in time. The second model builds in a parameter for annotator performance, as in previous work that adopts the Dawid and Skene (1979) model. Both assign more ground truth labels than majority voting, and avoid the problem with the majority vote method of ties where there are an even number of annotators. The results of the two models are very similar, but DCD HMM hypothesizes more boundaries, and therefore ranks some annotators differently.

Here we check the models by comparing them to each other, through analysis of each annotator's consistency across multiple conversations, and through inspection of the semantics of annotators' descriptions. Our future work will use the models generatively to predict a subset of the data for a given annotator, based on a model fit to all but the held out data. To do so, we would extend the models with an additional parameter for the conversation, to account for the observation that while all conversations seem to fit the same empirical distribution, there are differences across conversations.

10 Conclusion

Annotation and machine learning of discourse segmentation covers several types of units, including topical segments (Galley et al., 2003), meeting units in which action items are identified or decisions made (Purver et al., 2007; Fernández et al., 2008), transaction subtasks for ordering library books (Passonneau et al., 2014), or speaker involvement (Bokaei et al., 2015). This work relies on manual transcription, and draws on many sources of knowledge for machine learned models, including turn-taking, prosody, and linguistic features. The segmentation annotation can be linear (Galley et al., 2003; Bokaei et al., 2015; Passonneau and Litman, 1997; Passonneau et al., 2014) or hierarchical (Purver et al., 2007; Fernández et al., 2008; Passonneau et al., 2011). The differences in methods and results across this body of work, points to a need for more datasets for research on the organization of discourse into activity units. The results presented here support this research agenda by providing a reliable and cost-effective method to estimate ground truth discourse segment labels from crowd labels.

Acknowledgments

The authors thank Bob Carpenter for discussions during the early stages of the data analysis, and for helpful feedback on the paper. We thank the IARPA Babel program manager, Mary Harper, for giving us permission to annotate the Babel data.

References

Paul S. Albert and Lori E. Dodd. 2008. On estimating diagnostic accuracy from studies with multiple raters and partial gold standard evaluation. *Journal of the American Statistical Association*, 103(481):61–73.

André Berchtold. 1999. The Double Chain Markov Model. *Communications in Statistics: Theory and Methods*, 28(348):2569–2589.

Mohammad Hadi Bokaei, Hossein Sameti, and Yang Liu. 2015. Linear discourse segmentation of multiparty meetings based on local and global information. *IEEE Transactions on Audio, Speech and Language Processing*, 23(11):1879–1891.

Wauter Bosma. 2005. Extending answers using discourse structure. In *RANLP Workshop on Crossing Barriers in Text Summarization Research*.

Glenn Branch. 2014. Whence lumpers and splitters? `http://ncse.com/blog/2014/11/whence-lumpers-splitters-0016004`, December.

Rebecca F. Bruce and Janyce M. Wiebe. 1999. Recognizing subjectivity: a case study of manual tagging. *Natural Language Engineering*, 1(1):1–16.

A. P. Dawid and A. M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28.

Han de Vries. 1998. Finding a dominance order most consistent with a linear hierarchy: a new procedure and review. *Animal Behavior*, 55:827–843.

Alfred Dielmann and Steve Renals. 2005. Multistream dynamic Bayesian network for meeting segmentation. In Samy Bengio and Herv Bourlard, editors, *Machine Learning for Multimodal Interaction*, volume 3361 of *Lecture Notes in Computer Science*, pages 76–86. Springer Berlin Heidelberg.

Dmitriy Dligach, Rodney D. Nielsen, and Martha Palmer. 2010. To annotate more accurately or to annotate more. In *Proceedings of the Fourth Linguistic Annotation Workshop (LAW IV)*, pages 64–72.

Raquel Fernández, Matthew Frampton, Patrick Ehlen, Matthew Purver, and Stanley Peters. 2008. Modelling and detecting decisions in multi-party dialogue. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*.

Tim Finin, Will Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. 2010. Annotating named entities in twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 80–88, Stroudsburg, PA, USA. Association for Computational Linguistics.

Michel Galley, Kathleen R. McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 562–569, Sapporo, Japan, July. Association for Computational Linguistics.

Sebastian Germesin and Theresa Wilson. 2009. Agreement detection in multiparty conversation. In *Proceedings of the 2009 International Conference on Multimodal Interfaces*, ICMI-MLMI '09, pages 7–14, New York, NY, USA. ACM.

Dustin Hillard, Mari Ostendorf, and Elizabeth Shriberg. 2003. Detection of agreement vs. disagreement in meetings: Training with unlabeled data. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Companion Volume of the Proceedings of HLT-NAACL 2003–short Papers - Volume 2*, NAACL-Short '03, pages 34–36, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1120–1130, Atlanta, Georgia, June. Association for Computational Linguistics.

Klaus Krippendorff. 1980. *Content analysis: An introduction to its methodology*. Sage Publications, Beverly Hills, CA.

David Leiva, Antonio Solanas, and Lluis Salafranca. 2008. Testing reciprocity in social interactions: A comparison between the directional consistency and skew-symmetry statistics. *Behavior Research Methods*, 40(2):626–634.

Miriam Martinez and L. Enrique Sucar. 2008. Learning dynamic Naive Bayesian classifiers. In *Proceedings of the Twenty-First International FLAIRS Conference*, pages 655–659.

Kevin Murphy. 2001. Bayes Net toolbox for Matlab. *Computing Science and Statistics*, 33(2):1024–1034.

John Niekrasz and Johanna Moore. 2009. Participant subjectivity and involvement as a basis for discourse segmentation. In *Proceedings of the SIGDIAL 2009 Conference*, pages 54–61, London, UK, September. Association for Computational Linguistics.

Rebecca J. Passonneau and Bob Carpenter. 2014. The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, 2:311326.

Rebecca J. Passonneau and Diane Litman. 1997. Discourse segmentation by human and automated means. *Computational Linguistics*, 23.1:103–139. Special Issue on Empirical Studies in Discourse Interpretation and Generation.

Rebecca J. Passonneau, Irene Alvarado, Phil Crone, and Simon Jerome. 2011. PARADISE-style evaluation of a human-human library corpus. In *Proceedings of the SIGDIAL 2011 Conference*, pages 325–331, Portland, Oregon, June. Association for Computational Linguistics.

Rebecca J. Passonneau, Boxuan Guan, Cho Ho Yeung, Yuan Du, and Emma Conner. 2014. Aspectual properties of conversational activities. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 228–237, Philadelphia, PA, U.S.A., June. Association for Computational Linguistics.

Matthew Purver, John Dowding, John Niekrasz, Patrick Ehlen, Sharareh Noorbaloochi, and Stanley Peters. 2007. Detecting and summarizing action items in multi-party dialogue. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*.

Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. 2010. Learning from crowds. *Journal of Machine Learning Research*, 11:1297–1322.

Simon Rogers, Mark Girolami, and Tamara Polajnar. 2010. Semi-parametric analysis of multi-rater data. *Statistical Computing*, 20:317–334.

Andrey Rzhetsky, Hagit Shatkay, and W. John Wilbur. 2009. How to get the most out of your curation effort. *PLoS Computational Biology.*, 5(5):1–13.

Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. 2008. Get another label? Improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the Fourteenth ACM International Conference on Knowledge Discovery and Data Mining (KDD)*.

Padhraic Smyth, Usama Fayyad, Michael Burl, Pietro Perona, and Pierre Baldi. 1995. Inferring ground truth from subjectively-labeled images of Venus. In *Advances in Neural Information Processing Systems 7*, pages 1085–1092. MIT Press.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast - but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 254–263, Honolulu.

Swapna Somasundaran, Josef Ruppenhofer, and Janyce Wiebe. 2007. Detecting arguing and sentiment in meetings. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue (SIGdial 07)*, page 2634.

Mingyu Sun and Joyce Y. Chai. 2007. Discourse processing for context question answering based on linguistic knowledge. *Know.-Based Syst.*, 20(6):511–526, August.

Lu Wang and Claire Cardie. 2012. Focused meeting summarization via unsupervised relation extraction. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '12, pages 304–313, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nigel G. Ward, Steven D. Werner, Fernando Garcia, and Emilio Sanchis. 2015. A prosody-based vector-space model of dialog activity for information retrieval. *Speech Communication*, 28:85–96.

Jacob Whitehill, Paul Ruvolo, Tingfan Wu, Jacob Bergsma, and Javier Movellan. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Proceedings of the 24th Annual Conference on Advances in Neural Information Processing Systems*.