Multi-Emotion Estimation in Narratives from Crowdsourced Annotations

Lei Duan
Graduate School of Information Science and Technology
Hokkaido University
Sapporo 060-0814, Japan
duan@eis.hokudai.ac.jp

Satoshi Oyama
Graduate School of Information Science and Technology
Hokkaido University
Sapporo 060-0814, Japan
oyama@ist.hokudai.ac.jp

Haruhiko Sato
Graduate School of Information Science and Technology
Hokkaido University
Sapporo 060-0814, Japan
haru@complex.ist.hokudai.ac.jp

Masahito Kurihara
Graduate School of Information Science and Technology
Hokkaido University
Sapporo 060-0814, Japan
kurihara@ist.hokudai.ac.jp

ABSTRACT

Emotion annotations are important metadata for narrative texts in digital libraries. Such annotations are necessary for automatic text-to-speech conversion of narratives and affective education support and can be used as training data for machine learning algorithms to train automatic emotion detectors. However, obtaining high-quality emotion annotations is a challenging problem because it is usually expensive and time-consuming due to the subjectivity of emotion. Moreover, due to the multiplicity of "emotion", emotion annotations more naturally fit the paradigm of multi-label classification than that of multi-class classification since one instance (such as a sentence) may evoke a combination of multiple emotion categories. We thus investigated ways to obtain a set of high-quality emotion annotations ({instance, multi-emotion} paired data) from variable-quality crowdsourced annotations. A common quality control strategy for crowdsourced labeling tasks is to aggregate the responses provided by multiple annotators to produce a reliable annotation. Given that the categories of "emotion" have characteristics different from those of other kinds of labels, we propose incorporating domain-specific information of *emotional consistencies* across instances and *contextual cues* among emotion categories into the aggregation process. Experimental results demonstrate that, from a limited number of crowdsourced annotations, the proposed models enable gold standards to be more effectively estimated than the majority vote and the original domain-independent model.

JCDL'15, June 21–25, 2015, Knoxville, Tennessee, USA.

DOI: http://dx.doi.org/10.1145/2756406.2756910.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*indexing methods, linguistic processing*; J.5 [**Arts and Humanities**]: *literature*

General Terms

Algorithm, Experimentation, Human Factors, Measurement, Reliability

Keywords

Multi-emotion annotation; Emotional Consistency; Contextual cue; Crowdsourcing; Human computation

1. INTRODUCTION

Humans, by nature, can be emotionally affected by literature, music, fine art, etc., so "emotion" (also referred to as affect, feeling, sentiment, mood, etc.) conveyed in digital contents is essential for enjoying such content. Such information is usually provided as emotion annotations provided by human annotators. The paradigms used for emotion annotation depend on the application. For simple applications, it is sufficient to annotate whether an instance (such as a narrative line, a movie clip, or a music piece) is emotive or the instance's emotional valence (positive or negative). Such annotations obviously simplify the complexity of human emotion and are thus not effective for more complicated applications such as expressive text-to-speech synthesis [25] and affective education support [8]. Some researchers in this area [2, 15] have considered human emotion as a single category, with only one particular emotion (e.g., *happiness* or *sadness*) appearing at a time. However, this assumption has been undermined by the results of psychology studies. It has been demonstrated that a single emotion category is unable to represent all possible emotional manifestations [24] and that some emotional manifestations are a combination of several emotion categories [33]. For example, Alm

[1] observed that the following sentence from H. C. Andersen's fairy tale "The Ugly Duckling[1]" refers to *happiness* and *sadness* simultaneously:

> He now felt glad at having suffered sorrow and trouble, because it enabled him to enjoy so much better all the pleasure and happiness around him; for the great swans swam round the new-comer, and stroked his neck with their beaks, as a welcome.

A single emotion category would fail to represent this multiplicity.

Consequently, much of the recent emotion-oriented research [18, 23, 29] has concentrated on exploiting the multifaceted nature of emotion so that an instance can be associated with a combination of multiple emotion categories. An enormous number of emotion annotations ({instance, multi-emotion} paired data) is generally necessary for such research to be used as input for expressive text-to-speech synthesis, reference material for affective education support, training data for machine learning algorithms supporting automatic emotion detection, etc. Moreover, the annotation quality directly affects the quality of the research.

One way to obtain an enormous number of emotion annotations is to use online crowdsourcing services, which are being used more frequently in the labeling community. The state-of-the-art is for each instance to be annotated by one crowdsourcing annotator. However, the emotion labeling is more subjective than most other labeling tasks. There are different tendencies and substantial variations among individuals when detecting emotions, so high-quality emotion annotations should be in accordance with the general consensus of large crowds. This means that the annotation quality greatly depends on the judgment of an annotator, and is actually problematic in most cases. It is thus necessary to develop quality control strategies.

A promising quality control strategy is to introduce redundancy by asking several annotators (a sub-set of the crowd) to work on an instance and then aggregating their responses (crowdsourced annotations) to produce a reliable annotation (crowd's opinion). This is also called "approximating the crowd" [12]. The simplest aggregation strategy, *majority vote*, is valid only if the number of annotators is large enough. It is based on the implicit assumption that all annotators have the same probability of making an error. If the number of annotators is less than a certain unknown number, the detrimental effect of the noisy responses is significant, so treating responses given by different annotators equally would produce poor results. However, collecting responses from a large number of annotators is impractical due to the high cost (time and expense). To alleviate this problem, a number of state-of-the-art statistical techniques [7, 10, 22, 31, 32] have been proposed for producing a reliable annotation from a limited number of annotator responses used for crowdsourced labeling[2].

A crowdsourced labeling task is a form of semantic interpretation in which the instances are signs, the labels are referents, and the annotators are interpreters, as illustrated by the triangle of reference [21] (Figure 1). The interpreter perceives the sign (e.g., a word, a sound, an image, a sentence) and through some cognitive process attempts to find

[1]http://www.surlalunefairytales.com/uglyduckling/index.html
[2]For a detailed discussion, see Section 6.2

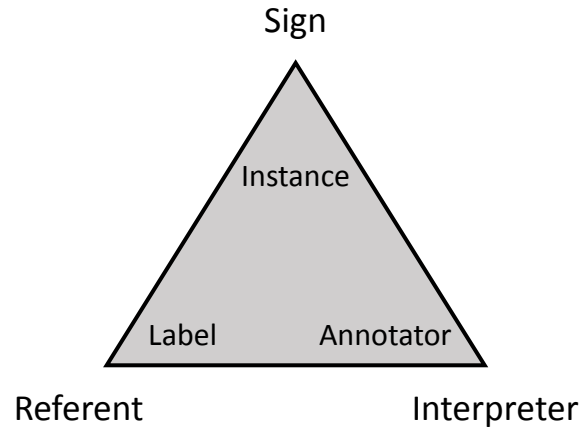

Figure 1: Triangle of Reference

the referent of that sign (e.g., an object, an idea, a class of things). As discussed further in section 6.2, most state-of-the-art quality control techniques include latent factors related to these three components, such as annotator bias [7], annotator expertise [32], and instance difficulty [31, 32]. However, they ignore the internal relationships among instances and among labels.[3] In other words, for each repeatedly annotated instance, the reliable annotation is produced separately.

In addition to the subjectivity characteristic as discussed above, "emotion" has its unique characteristics in comparison with other kinds of labels:

• Internal relationships among instances

Emotive expressions can not be divorced from their context [3]. The media used to convey emotion include (but are not limited to) narrative, music, cinema, facial expression, and body language. Take narrative, the main focus of this paper, as an example. As the genre of literature characterized by description, narrative usually subjects to certain emotional tendency, and characters in a narrative typically have distinct personalities. Both narrative emotional tendency and character personality tend to remain consistent across instances (sentences) in the same context. If they did not, the emotive expression of the utterances of expressive text-to-speech synthesis, a potential application, would result in unnatural pronunciation.

• Internal relationships among labels

Emotions (labels) expressed by a sentence likely relate to the emotions expressed by the subsequent sentences. For example, a boy scolded by his mother for some mistake would more likely feel sad and disgusted while the mother would more likely feel angry. It is thus beneficial to know the contextual cues among emotions.

In this work, we explore the domain-specific viability of crowdsourced emotion annotations in narratives. We asked crowdsourcing annotators to read narrative lines such as those shown in Figure 2 and spontaneously indicate the

[3]The relationship among annotators in a collaborative crowdsourcing task should normally be taken into account, but this is not relevant here since the annotators made their decisions independently in our experiments.

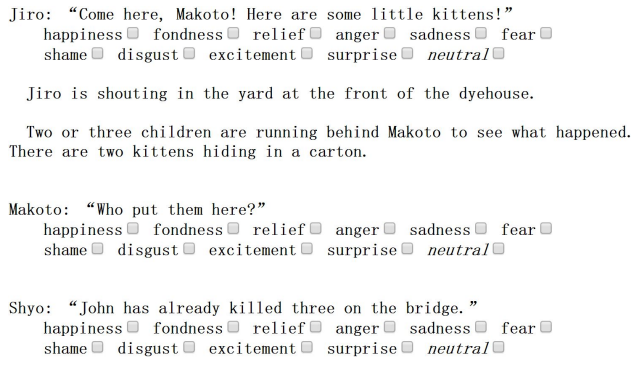

Figure 2: Example task input screen (translated from Japanese). Annotators were native Japanese language speakers. Both candidate emotions and lines were presented to annotators in their original Japanese form.

character's emotions expressed in each line. The true (gold standard) emotions for each line were then estimated by aggregating the obtained multi-emotion responses. As an extension to our previous work [10] on domain-independent multi-label estimation, we propose incorporating the internal relationships among lines (instances) and among emotions (labels) into the estimation process. The two relationships are respectively specified as the domain-specific information of *emotional consistencies* across lines and *contextual cues* among emotions. An expectation maximization (EM) based incremental algorithm is used to estimate the gold standard emotion annotations together with the parameters of the proposed models. The experimental results demonstrate that incorporating the domain-specific information into the estimation process makes the accuracy rates of the proposed models higher than those of the majority vote and the original domain-independent model for the same number of annotations.

The remainder of this paper is organized as follows. Section 2 briefly describes the sources of annotated narratives and candidate emotions. Section 3 reviews the original domain-independent multi-label estimation model, the foundation of our study, and introduces the two proposed domain-specific multi-emotion estimation models. Section 4 describes the use of the EM algorithm to estimate the gold standard emotion annotations together with the model parameters. Section 5 describes the experimental design and discusses the results obtained by applying the proposed models to real-world data. Section 6 provides background material by introducing related research. Section 7 summarizes the main points and suggests several future research directions.

2. DATASET DESCRIPTIONS

2.1 Aozora Library

The Aozora (Blue Sky) Library[4] is a Japanese online repository containing freely available books. It contains over 10,000 books of various genres (philosophy, history, art, etc.) published in Japanese for which copyrights have expired (50 years after the death of the copyright holder). The children's book genre in Aozora Library includes sub-categories such as *history books*, *beautiful arts and crafts books*, and *literature*. Children's narratives are included in the *literature* category, which contained 1217 books in December 2014. From this category we chose two narratives at random for our experiments, which are discussed in Section 5.

[4]http://www.aozora.gr.jp

Table 1: Distribution of separate expressions across emotions in Nakamura's dictionary, ordered by number of expressions per emotion.

Emotion	No. of expressions	Emotion	No. of expressions
Disgust	532	Fondness	197
Excitement	269	Fear	147
Sadness	232	Surprise	129
Happiness	224	Relief	106
Anger	199	Shame	65
		Total	2100

2.2 Emotive Expression Dictionary

Since the chosen narratives were in Japanese, we had to use a candidate emotion category set proven to be appropriate for the Japanese language. The Emotive Expression Dictionary [19] is a dictionary developed by Akira Nakamura over a period exceeding 20 years. It is a collection of over 2000 expressions describing emotional states that were collected manually from a wide range of literature. It is not a tool for emotion analysis *per se* but was converted into an emotive expression database by Ptaszynski *et al.* [23] in their research on emotion analysis of utterances in Japanese. This dictionary is a state-of-the-art example of a hand-crafted lexicon of emotive expressions. In particular, it uses ten emotion categories that appropriately reflect the Japanese language and culture. This classification is also applied in the lexicon itself. Each expression is classified as representing one specific emotion category, or more if applicable. The distribution of separate expressions across all emotion categories is represented in Table 1.

3. STATISTICAL MODELS

Multi-emotion estimation from crowdsourced annotations can be seen as an unsupervised multi-label classification problem. Two widely used methods for multi-label classification are the binary relevance (BR) method and the label combination or label power-set (LP) method [30].

The BR method decomposes the multi-label classification problem into several independent binary-label classification problems, one for each label in the set of candidate labels. The final labels for each instance are determined by aggregating the predictions from all binary estimators. In other words, the BR method does not consider dependency among candidate labels. This is reasonable only in the extreme case that labels are mutually independent. However, it is obvious that emotions are interrelated. Some emotions may reveal clues about others. For example, a line expressing *fear* may also express a certain degree of *anger* and/or *surprise*.

To take into account the dependency relationship among candidate emotions, we use the LP method. It treats each unique subset of labels in the set of candidate labels as an atomic "label" and defines a new single-label estimation problem, i.e., estimating each member of the power-set of

the candidate label set. We use the concept of "conjoint-emotion" to represent a subset of the candidate emotion set. For example, the two conjoint-emotions {*happiness*, *relief*} and {*happiness*, *excitement*} express two different kinds of "happiness": one is comparatively mild while the other is strong.

Problem Formulation:

Let I be the set of lines, J be the set of candidate emotions, and K be the set of annotators. The number of times that annotator k annotated line i with conjoint-emotion $\mathcal{L}$ is given by $n_{i\mathcal{L}}^{(k)} \in \mathbb{N}$ $(k \in K, i \in I, \mathcal{L} \subseteq J)$. The true conjoint-emotion, namely the multiple true emotions, for line i is denoted by $\mathcal{T}_i \subseteq J$ $(i \in I)$. The objective is to aggregate the set of annotations $\left\{n_{i\mathcal{L}}^{(k)} : k \in K, i \in I, \mathcal{L} \subseteq J\right\}$ to estimate the set of true conjoint-emotions $\{\mathcal{T}_i : i \in I\}$.

We first introduce the original model aimed at handling domain-independent multi-label estimation in Section 3.1, and then discuss how we extended it to handle domain-specific multi-emotion estimation by considering *emotional consistencies* and *contextual cues* in Sections 3.2 and 3.3.

3.1 Original Domain-independent Multi-label Estimation

The original domain-independent model was proposed in our previous work [10]. The graphical representation of the model is illustrated in Figure 3(a). The true conjoint-emotion for line i is determined by the maximum a posteriori (MAP) principal:

$$\mathcal{T}_i = \arg\max_{\mathcal{J} \subseteq J} \mathrm{E}\left[\mathcal{T}_i = \mathcal{J}\right] , \tag{1}$$

where the expectation of the true conjoint-emotion for line i is estimated as the conditional distribution given the annotations for line i:

$$\mathrm{E}\left[\mathcal{T}_i = \mathcal{J}\right] = \Pr\left[\mathcal{T}_i = \mathcal{J} \mid \left\{n_{i\mathcal{L}}^{(k)} : k \in K, \mathcal{L} \subseteq J\right\}\right] . \tag{2}$$

The bias for annotator k is defined as the probability that she annotated a line with conjoint-emotion $\mathcal{L}$ when the true conjoint-emotion of the line is $\mathcal{J}$:

$$\pi_{\mathcal{J}\mathcal{L}}^{(k)} := \frac{\sum_{i \in I} \gamma_{i\mathcal{J}} \cdot n_{i\mathcal{L}}^{(k)}}{\sum_{\mathcal{L} \in J} \sum_{i \in I} \gamma_{i\mathcal{J}} \cdot n_{i\mathcal{L}}^{(k)}} ,$$

where $\gamma_{i\mathcal{J}}$ is defined as the prior probability that $\mathcal{J}$ is the true conjoint-emotion for line i:

$$\gamma_{i\mathcal{J}} := \Pr[\mathcal{T}_i = \mathcal{J}] .$$

Since the annotators make their decisions independently, the numbers of times that an annotator annotated instance i with different conjoint-emotions when $\mathcal{J}$ is the true conjoint-emotion are distributed according to a multinomial distribution. Using Bayes' Theorem, we obtain the expectation in Equation (2):

$$\mathrm{E}\left[\mathcal{T}_i = \mathcal{J}\right] = \frac{\gamma_{i\mathcal{J}}}{z_i} \prod_{k \in K} \prod_{\mathcal{L} \subseteq J} \left(\pi_{\mathcal{J}\mathcal{L}}^{(k)}\right)^{n_{i\mathcal{L}}^{(k)}} , \tag{3}$$

where z_i is defined as the normalization constant:

$$z_i := \sum_{\mathcal{J} \subseteq J} \left(\gamma_{i\mathcal{J}} \cdot \prod_{k \in K} \prod_{\mathcal{L} \subseteq J} \left(\pi_{\mathcal{J}\mathcal{L}}^{(k)}\right)^{n_{i\mathcal{L}}^{(k)}}\right) ,$$

which ensures that the posterior distribution in Equation (2) is a valid probability density and can be integrated to one. The detailed derivation is given elsewhere [10].

3.2 Multi-emotion Estimation Considering Emotional Consistencies

Multi-emotion estimation is a domain-specific aspect of multi-label estimation. To improve the accuracy of the domain-independent estimation model described in Section 3.1, we first propose incorporating *emotional consistencies* into the estimation process. As discussed in Section 1, there are *emotional consistencies* across lines (as well as *contextual cues* among emotions) in narrative emotional tendency and in character personality. We thus define narrative emotional tendency as the distribution of separate expressions over conjoint-emotions:

$$\alpha_{\mathcal{J}} := \frac{\sum_{i \in I} \gamma_{i\mathcal{J}}}{|I|} .$$

It is obtained by maximum-likelihood estimation (MLE) of the proportion of lines expressing conjoint-emotion $\mathcal{J}$ over all lines.

Let $c\,(i)$ $(i \in I)$ denote the character speaking in line i and $I_{c(i)}$ $(i \in I)$ denote the lines spoken by character $c\,(i)$. Similar to the narrative emotional tendency, the character personality is also defined as the distribution over conjoint-emotions:

$$\beta_{c(i)\mathcal{J}} := \frac{\sum_{\imath \in I_{c(i)}} \gamma_{\imath\mathcal{J}}}{\left|I_{c(i)}\right|} .$$

It is obtained by MLE of the proportion of the lines expressing conjoint-emotion $\mathcal{J}$ over all the lines spoken by character $c\,(i)$.

As a domain-specific extension of Equation (2), the expectation of the true conjoint-emotion for line i is then estimated considering *emotional consistencies* in the narrative emotional tendency and in the character personality:

$$\mathrm{E}\left[\mathcal{T}_i = \mathcal{J}\right] = \Pr\left[\mathcal{T}_i = \mathcal{J} \mid \left\{n_{i\mathcal{L}}^{(k)} : k \in K, \mathcal{L} \subseteq J\right\}; \alpha_{\mathcal{J}}, \beta_{c(i)\mathcal{J}}\right] . \tag{4}$$

This means that the true conjoint-emotion of each line is determined not only by the annotations but also by the narrative emotional tendency and the character personality. The graphical representation of the model is illustrated in Figure 3(b). In a manner similar to that for Equation (3), using Bayes' Theorem, we obtain the expectation in Equation (4):

$$\mathrm{E}\left[\mathcal{T}_i = \mathcal{J}\right] = \frac{\alpha_{\mathcal{J}} \cdot \beta_{c(i)\mathcal{J}} \cdot \gamma_{i\mathcal{J}}}{z_i} \prod_{k \in K} \prod_{\mathcal{L} \subseteq J} \left(\pi_{\mathcal{J}\mathcal{L}}^{(k)}\right)^{n_{i\mathcal{L}}^{(k)}} , \tag{5}$$

where z_i is the normalization constant:

$$z_i = \sum_{\mathcal{J} \subseteq J} \left(\alpha_{\mathcal{J}} \cdot \beta_{c(i)\mathcal{J}} \cdot \gamma_{i\mathcal{J}} \cdot \prod_{k \in K} \prod_{\mathcal{L} \subseteq J} \left(\pi_{\mathcal{J}\mathcal{L}}^{(k)}\right)^{n_{i\mathcal{L}}^{(k)}}\right) .$$

Equation (5) demonstrates that the domain-specific model considering *emotional consistencies* automatically assigns higher weights to the conjoint-emotions that are more consistent with the narrative emotional tendency and with the character personality, and assigns lower weights to those that are less consistent.

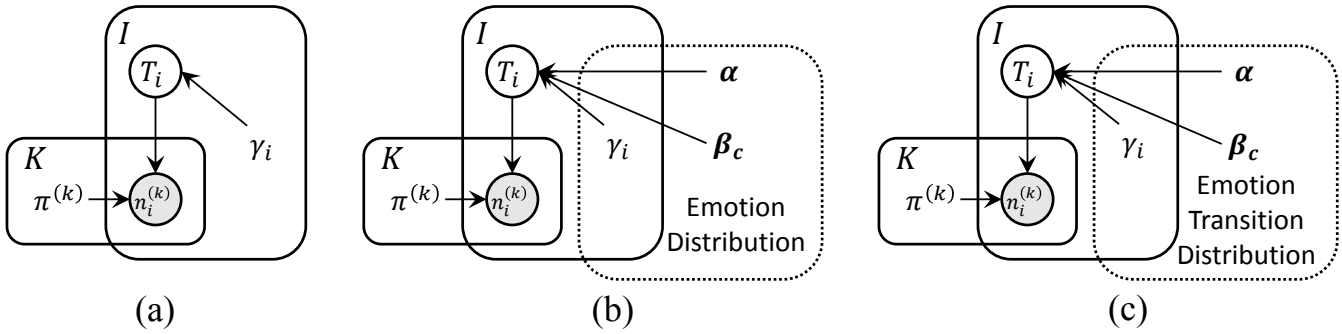

Figure 3: Graphical model representation for multi-emotion estimation: (a) original domain-independent model, (b) model considering *emotional consistencies*, and (c) model considering *emotional consistencies* with *contextual cues*.

3.3 Multi-emotion Estimation Considering Emotional Consistencies with Contextual Cues

As mentioned in Section 1, the emotions expressed by a sentence likely relate to the emotions expressed by the subsequent sentences, so it is beneficial to know the *contextual cues* among emotions. As a statistical measure of those "cues" we use the possibility of a conjoint-emotion following another conjoint-emotion. For example, the conjoint-emotion {*anger*, *disgust*} is more likely followed by {*sad*, *fear*} than {*happiness*, *fondness*}. This means that {*anger*, *disgust*} has a closer relationship with {*sad*, *fear*} than with {*happiness*, *fondness*}.

As an extension to the model proposed in Section 3.2, we propose estimating *emotional consistencies* across instances by using *contextual cues* among emotions. We use the idea of *bi-gram* to learn the transition distribution over conjoint-emotions. This means that the conjoint-emotion expressed by a line is conditional on the conjoint-emotion expressed by the previous line. Let $i-1$ be the line before line i and $\bar{\mathcal{J}}$ be the true conjoint-emotion of line $i-1$, which means

$$\bar{\mathcal{J}} = \arg\max_{\mathcal{J}\subseteq J} \mathrm{E}[\mathcal{T}_{i-1} = \mathcal{J}]\,.$$

The *contextual cues* are extracted using parameters $\{\alpha\}$, $\{\beta\}$, and $\{\gamma\}$, which are the counterparts of the parameters defined in Sections 3.1 and 3.2. They are estimated considering the true conjoint-emotions for two consecutive lines:

$$\begin{aligned}
\alpha_{\mathcal{J}} =& \frac{\sum_{i\in I} \Pr\left[\mathcal{T}_{i-1} = \bar{\mathcal{J}}, \mathcal{T}_i = \mathcal{J}\right]}{\sum_{i\in I} \Pr\left[\mathcal{T}_i = \bar{\mathcal{J}}\right]}\,,\\
\beta_{c(i)\mathcal{J}} =& \frac{\sum_{\imath\in I_{c(i)}} \Pr\left[\mathcal{T}_{\imath-1} = \bar{\mathcal{J}}, \mathcal{T}_\imath = \mathcal{J}\right]}{\sum_{\imath\in I_{c(i-1)}} \Pr\left[\mathcal{T}_\imath = \bar{\mathcal{J}}\right]}\,,\\
\gamma_{i\mathcal{J}} =& \frac{\Pr\left[\mathcal{T}_{i-1} = \bar{\mathcal{J}}, \mathcal{T}_i = \mathcal{J}\right]}{\Pr\left[\mathcal{T}_{i-1} = \bar{\mathcal{J}}\right]}\,.
\end{aligned}$$

The annotations for each line are provided by arbitrary annotators, but we need to obtain cases in which one conjoint-emotion followed by another as complete as possible. Therefore, we use a *cross-strategy* among annotators to compute the joint distribution for two consecutive lines:

$$\begin{aligned}
&\Pr[\mathcal{T}_{i-1} = \bar{\mathcal{J}}, \mathcal{T}_i = \mathcal{J}]\\
=&\frac{\sum_{k\in K} n^{(k)}_{(i-1)\bar{\mathcal{J}}} \cdot \sum_{k\in K} n^{(k)}_{i\mathcal{J}}}{\sum_{k\in K}\sum_{\jmath\subseteq J} n^{(k)}_{(i-1)\jmath} \cdot \sum_{k\in K}\sum_{\jmath\subseteq J} n^{(k)}_{i\jmath}}\,.
\end{aligned}$$

The differences among the original domain-independent model, the model considering *emotional consistencies*, and the model considering *emotional consistencies* with *contextual cues* are illustrated in Figure 3. The proposed domain-specific multi-emotion estimation models, (b) and (c), extend the original domain-independent model (a) by statistically analyzing *emotional consistencies* in narrative emotional tendency and character personality, which are estimated as probabilistic variables following different distributions.

4. INFERENCE ALGORITHM

Inference of the gold standard emotion annotations and the parameters can be greatly simplified if we use the EM algorithm. The EM algorithm is an efficient iterative procedure for computing the maximum-likelihood solution in presence of hidden/missing data. It is widely used in crowdsourcing related research [7, 32, 22, 10]. We treat the MLEs of the parameters

$$\begin{aligned}
&\{\alpha_{\mathcal{J}} : \mathcal{J}\subseteq J\}\,\text{(narrative emotional tendency)}\,,\\
&\left\{\beta_{c(i)\mathcal{J}} : i\in I, \mathcal{J}\subseteq J\right\}\,\text{(character personality)}\,,\\
&\{\gamma_{i\mathcal{J}} : i\in I, \mathcal{J}\subseteq J\}\,\text{(prior distribution)}\,,\\
&\left\{\pi^{(k)}_{\mathcal{J}\mathcal{L}} : k\in K, \mathcal{J}\subseteq J, \mathcal{L}\subseteq J\right\}\,\text{(annotator bias)}\,,\\
&\{z_i : i\in I\}\,\text{(normalization constant)}\,,
\end{aligned}$$

as the hidden data, with the expectations of the true conjoint-emotions for each line

$$\{\mathrm{E}[\mathcal{T}_i = \mathcal{J}] : i\in I, \mathcal{J}\subseteq J\}$$

as unobserved variables (missing data). The observed variables

$$\left\{n^{(k)}_{i\mathcal{L}} : i\in I, k\in K, \mathcal{L}\subseteq J\right\}$$

can be directly calculated from the obtained crowdsourced annotations.

We therefore proceed as follows.

(1) Initialization

Obtain the initial estimates of unobserved variables {E}:

$$\mathrm{E}\left[\mathcal{T}_i = \mathcal{J}\right] = \frac{\sum_{k \in K} n_{i\mathcal{J}}^{(k)}}{\sum_{k \in K} \sum_{\jmath \subseteq J} n_{i\jmath}^{(k)}} ,$$

which is an intuitive way to assign the MLEs. This approach can be computationally demanding because it is equivalent to estimating a $|J|$-dimensional joint distribution for each instance over the candidate emotions. Because the emotion states are binary-valued, the joint distribution requires the probabilities of $2^{|J|}$ different assignments of values. For all but the smallest $|J|$, the explicit representation of the joint distribution is unmanageable from every perspective. Therefore, at the practical level, it is too expensive and nearly impossible to acquire a sufficient number of samples from annotators to robustly estimate the high-dimensional joint distribution. An effective strategy to overcome the adverse effect of data sparsity is to represent the underlying joint distribution more compactly. By using the conditional independence properties, we can approximate the joint distribution from a finite number of annotations. Several methods are available for such approximation [6]. We used a Bayesian network as it was previously shown to be effective for multi-label estimation [10].

(2) Maximization

Estimate the MLEs of hidden parameters $\{\alpha\}$, $\{\beta\}$, $\{\gamma\}$, $\{\pi\}$, and $\{z\}$ using the equations in Section 3 with the current estimates of {E}, which means

$$\Pr\left[\mathcal{T}_i = \mathcal{J}\right] = \mathrm{E}\left[\mathcal{T}_i = \mathcal{J}\right] .$$

(3) Expectation

Estimate the expected values of the unobserved variables, {E}, using Equation (5) with the current estimates of the parameters calculated in step (2).

(4) Alternation

Alternately perform steps (2) and (3) until the likelihood for all annotations $\Pr\left[\left\{n_{i\mathcal{L}}^{(k)} : k \in K, i \in I, \mathcal{L} \subseteq J\right\}\right]$ converges. Since all lines are independently annotated, we have

$$\Pr\left[\left\{n_{i\mathcal{L}}^{(k)} : k \in K, i \in I, \mathcal{L} \subseteq J\right\}\right] = \prod_{i \in I} z_i .$$

At this point, the $\mathcal{J}$ with the maximum $\mathrm{E}[\mathcal{T}_i = \mathcal{J}]$ is the true conjoint-emotion for line i, as specified in Equation (1).

One of the characteristics of the EM algorithm is that, after alternately performing steps (2) and (3), only one of the unobserved variables for a line has a probability converging towards 1 while the other unobserved variables for the line have a probability converging towards 0. In other words, it is unlikely that any of the expectations {E} is between 0.1 and 0.9.

5. EMPIRICAL STUDY

To evaluate the effectiveness of the proposed models, we needed narratives in which the lines express clear emotions. Since children typically have an elementary level of psychological development, narratives written for them usually have vibrant affection tints and distinct character personalities as the aim is to better attract the attention of children.

Table 2: Annotation frequencies of emotions and *neutral*, ordered by total frequency

Emotion	"Love"	"Apple"	Total
Relief	516	362	878
Anger	242	623	865
Sadness	522	298	820
Happiness	458	306	764
Fondness	467	226	693
Excitement	379	270	649
Disgust	279	265	544
Neutral	*120*	*352*	*472*
Surprise	190	243	433
Fear	164	107	271
Shame	84	68	152
Total (except *Neutral*)	3301	2768	6069

Table 3: Statistics for the experiments

	"Love"	"Apple"	Total
No. of lines	63	78	141
No. of characters	12	9	21
No. of annotators	30	57	84
No. of annotations	1890	2340	4230
Avg. no. of checked labels per annotation	1.75	1.18	1.43
Avg. no. of annotations per line	30	30	30

The proportion of speaking lines in children's narratives is also higher than that in other narrative genres. Therefore, children's narratives are commonly used in emotion-oriented research [2, 8]. These characteristics of children's narratives are also the focal points of our research. We thus chose two Japanese children's narratives, "Although we are in love"[5] ("Love" for short) and "Little Masa and a red apple"[6] ("Apple" for short), from the Aozora Library[7] as the annotated texts. We conducted the experiments using the Lancers crowdsourcing service[8].

Due to different aspects that emotion-oriented research looks to capture, the candidate emotion category set used differs among research efforts. While the six basic emotions (*happiness*, *fear*, *anger*, *surprise*, *disgust*, and *sadness*) [11] are typically used [1, 2, 29], we used ten candidate emotions in order to provide more choices to the annotators and thereby enable us to perform an in-depth study on multi-emotion estimation. They were taken from the "Emotive Expression Dictionary"[9], which contains emotions proven to be appropriate for the Japanese language and culture [23]. An example task input screen is shown in Figure 2. If none of the candidate emotions was felt, the annotator would check *neutral*. The candidate emotions are shown in Table 2 with their annotation frequencies. Other statistics for the experiments are shown in Table 3.

[5]http://www.aozora.gr.jp/cards/001475/files/52111_47798.html
[6]http://www.aozora.gr.jp/cards/001475/files/52113_46622.html
[7]For a detailed description, see Section 2.1
[8]http://www.lancers.jp
[9]For a detailed description, see Section 2.2

For the emotion annotations to be reliable, they should be in accordance with the general consensus of large crowds. The majority vote strategy most objectively reflects the general consensus if the number of annotators is large enough. Therefore, we obtained gold standards by having each line annotated 30 times and then taking the majority vote. That is, the most often annotated conjoint-emotion for a line was used as the gold standard for that line. For the "Love" narrative, we asked each of the 30 annotators to annotate each line one time, which ensured that each annotator annotated the complete set of the lines. For the "Apple" narrative, the annotation task was divided into small parts and distributed to the annotators in a nonspecific manner, so the 30 annotations for every line were provided by arbitrary annotators, and few, if any, of them annotated the complete set of the lines. This is a more realistic situation since it is not a good idea to submit a very large task to a crowdsourcing service because a big task tends to diminish annotator enthusiasm or even cause annotators to avoid the task. We conducted the "Apple" task in this way simply to examine the effects of "arbitrary annotator interference" on the model results.

Although our proposed models can handle a line being annotated more than once by an annotator, to collect opinions as widely as possible at a fixed cost, it is still best to avoid this situation even though an annotator may interpret a line differently at different times. Therefore, in our experiments, all the annotations for a line were obtained from different annotators. This means that the values of the observed variable $n_{i\mathcal{L}}^{(k)}$ $(k \in K, i \in I, \mathcal{L} \subseteq J)$ was either 0 or 1.

To determine the effect of the number of annotators per line on accuracy, we randomly split the 30 annotators who annotated a particular line into various numbers of groups of equal size. We used five different group sizes: 3 (ten groups), 5 (six groups), 10 (three groups), 15 (two groups), and 30 (one group). The true conjoint-emotion for each line was estimated given the annotations within each group using the following four models:

- *MV*: **M**ajority **V**ote;
- *OL*: **O**riginal domain-independent multi-**L**abel estimation model;
- *EC*: domain-specific multi-**E**motion estimation model considering *emotional **C**onsistency*;
- *EC+*: domain-specific multi-**E**motion estimation model considering *emotional **C**onsistency* with *contextual cues*.

MV and *OL* are the baselines to which we compare the results for *EC* and *EC+*.

Both the estimation result and the gold standard for a line can be regarded as a binary vector. It is unreasonable to check whether the two binary vectors match exactly. For example, {*happiness*, *fondness*} is closer to {*happiness*, *relief*} than {*anger*, *disgust*}. Therefore, the average *Simple Matching Coefficient* is used to evaluate the performance of the proposed models, i.e., the average proportion of correct emotions between the estimation results and the gold standards for all lines within a group.

The MV results for 30 annotators in Figure 4 represent the accuracy (1.00) of the gold standard group. For both narratives, when the group size was 3, 5, 10, or 15, all three statistical models achieved better average accuracies than the MV model. Although the accuracies of the statistical

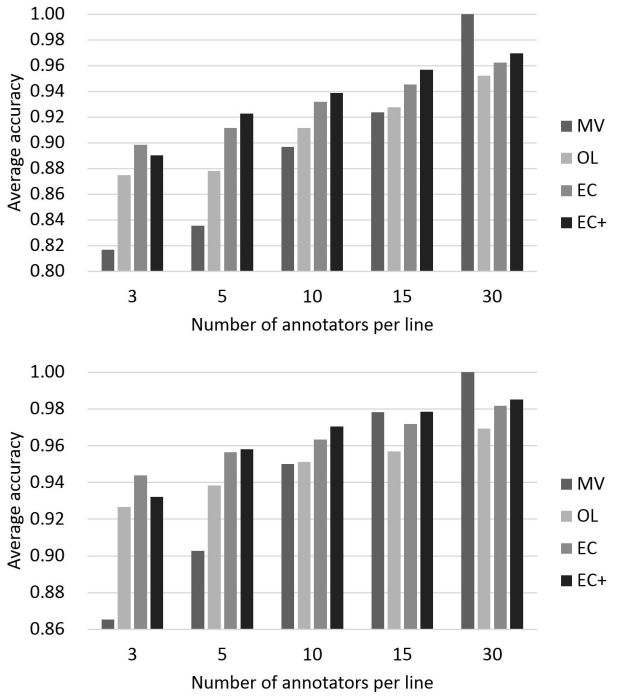

Figure 4: Average accuracy for "Although we are in love" narrative (upper) and "Little Masa and a red apple" narrative (lower).

models increased with the group size, the *EC* and *EC+* models consistently outperformed the *OL* model and had accuracies greater than 90 % for five or more annotators per line. This means that considering *emotional consistencies* in narrative emotional tendency and in character personality is effective for multi-emotion estimation, and five would be a reasonable number of annotators for each line to achieve satisfactory performance. Moreover, the average accuracy of the *EC+* model increased fastest and exceeded that of the *EC* model when the group size was five or more. The reason for this phenomenon is that, when the group size was 3, the adverse effect of data sparsity was dominant, and the quantity of annotations was insufficient to well learn the transition distribution over conjoint-emotions. However, once the quantity was sufficient, the superiority of the *EC+* model, which considers *emotional consistencies* with *contextual cues*, became evident. Finally, none of the models was particularly sensitive to the effect of "arbitrary annotator interference" in the "apple" narrative.

All the models were run on a workstation with an Intel Core i7-3770 3.40-GHz 4-core processor, 8-GB RAM, and the Windows 7 64-bit operating system. We found that even the most complicated model, the *EC+* model, converged in less than 10 seconds when using the annotations of 30 annotators as the input. There are two reasons for this performance. One is that the computational complexity of the proposed models is linear in the number of lines, the number of annotators, and the number of candidate emotions and the dataset used was not so large. The other is that we did not estimate the expectations {E} over all possible conjoint-emotions ($2^{|J|}$) for all the sentences (I). In fact, we

only estimated those of conjoint-emotions that have been annotated in a sentence by at least one annotator.

6. BACKGROUND AND RELATED WORK

6.1 Emotion-oriented Research

Emotion is an important access point in digital libraries and online repositories. Analyzing how we are affected is a vital research direction in digital media processing as it is potentially applicable to many further emotion-related applications, including expressive text-to-speech synthesis [25] and therapeutic education of children with communication disorders [8]. Many researchers have thus concentrated on this area. Alm *et al.* [2] investigated the importance of various features for emotion analysis and classified the emotional affinity of sentences in the narrative domain of children's fairy tales. Kim *et al.* [15] modeled emotion as a continuous *manifold* and constructed a statistical model connecting it to documents and to a discrete set of emotions. A number of machine learning algorithms have been proposed for classifying music by mood in the music digital library domain [4, 14].

Several researchers concentrated on exploiting the multifaceted nature of emotion. Trohidis *et al.* [29] modeled emotion detection in music pieces as a multi-label classification task. Liu *et al.* [18] proposed an implicit video multi-emotion tagging method. Ptaszynski *et al.* [23] did an experiment on multi-emotion analysis of certain characters in narratives. A complete discussion of emotion-oriented research is beyond the scope of this paper but can be found in Calvo *et al.* [5].

Traditional emotion classification research is aimed at detecting single or multiple emotion(s) from an instance using a trained detector. Our work is aimed at estimating multiple emotions directly from crowdsourced annotations. The goal is to prepare high-quality emotion annotations at low cost for use with emotion-related applications and training data for emotion detectors. To the best of our knowledge, there is no work comparable to our proposed models, except our previous work [10] (the *OL* model in Section 5, introduced in Section 3.1).

6.2 Crowdsourcing and Quality Control

For the term *crowdsourcing*, Howe [13] offers the following definition:

> Crowdsourcing represents the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call. This can take the form of peer-production (when the job is performed collaboratively), but is also often undertaken by sole individuals. The crucial prerequisite is the use of the open call format and the large network of potential laborers.

Simply put, crowdsourcing is an economical and efficient approach to performing tasks that are difficult for computers but relatively easy for humans. With the recent expansion of crowdsourcing platforms such as Amazon Mechanical Turk[10] (MTurk) and CrowdFlower[11], the concept of crowdsourcing has been successfully leveraged in various areas of computer science research, including natural language processing [27] and computer vision [28]. There have also been several attempts in the emotion detection domain. Alm [1] analyzed the characteristics of sentences with high-agreement crowdsourced emotion annotations. He tentatively hypothesized that some characteristics of high-agreement annotations may show particular affinity with certain emotions. Lee *et al.* [16] compared the music emotion annotations collected from music experts with annotations collected using MTurk. They showed that the overall distribution of emotions and agreement rates from music experts and MTurk were comparable.

Although annotations can be obtained from a crowdsourcing service at very low cost (time and expense), there is no guarantee that all annotators are sufficiently competent to complete the offered tasks. In fact, crowdsourcing annotators are rarely trained and generally do not have the abilities needed to accurately perform the offered task. Some annotators may even simply submit random responses as a means to earn easy money. Therefore, ensuring the annotation quality from noisy responses is one of the biggest challenges in crowdsourcing.

A simple strategy would be to offer incentive programs for the annotators, such as giving monetary bonuses to high-performance ones and denying payments to low-performance ones. In addition, several approaches geared toward efficient quality control have been applied. For example, MTurk provides a pre-qualification system to assess the skill level of a prospective annotator, and CrowdFlower enables requesters to inject a collection of tasks with known correct answers into their tasks to automatically measure an annotator's performance.

Meanwhile, various statistical schemes have been proposed to aggregate multiple variable-quality annotations from non-expert annotators to yield results that rival gold standards. Dawid *et al.* [7] presented a method for inferring the unknown health state of a patient given diagnostic tests by several clinicians, where the biases of the annotators (clinicians) were modeled by a confusion matrix. Whitehill *et al.* [32] presented a model for simultaneously estimating the true label of each repeatedly labeled instance, the expertise of each annotator, and the difficulty of each question. Welinder *et al.* [31] incorporated into their bird image classification model all the above factors, along with a normalized weight vector for each worker, where each weight indicates relevance to the worker. Snow *et al.* [27] demonstrated that by using an automatic bias correction algorithm, MTurk can be used effectively for a variety of natural language annotation tasks. Lin *et al.* [17] took a decision-theoretic approach to estimating the correct answer for a task that can have a countably infinite number of possible answers. Ertekin *et al.* [12] presented an algorithm that works in an online fashion to produce a weighted combination of a subset's votes that approximates the crowd's opinion. Oyama *et al.* [22] investigated the use of not only crowdsourced annotations, but also annotators' self-reported confidence scores. These works focused on single-label estimation.

In the multi-label domain, Duan *et al.* [10] proposed a method for estimating multiple true labels for each repeatedly multi-labeled instance, with flexible incorporation of label dependency into the label-generation process. Nowak *et al.* [20] studied inter-annotator agreement for multi-label

[10]http://www.mturk.com
[11]http://crowdflower.com

image annotation. They found that using the majority vote strategy to generate one annotation set from several responses can filter out noisy responses of non-experts to some extent.

Some research in the machine learning community addressed the problem of supervised learning directly from crowdsourced annotations. Sheng *et al.* [26] explored several methods for choosing which instances should get more labels and how to include label uncertainty information when training classifiers. Donmez *et al.* [9] proposed simultaneously estimating annotator accuracies and training a classifier using annotator responses to actively select the next instance for annotating.

7. CONCLUSION

We investigated multi-emotion estimation from crowdsourced annotations, where an instance (such as a narrative line, a movie clip, or a music piece) can be associated with a combination of multiple emotion categories. The original domain-independent multi-label estimation model ignored the internal relationships among instances and among labels. To improve the accuracy, we extended it to specify those two relationships as the domain-specific information by making use of the characteristics of emotion, the *emotional consistencies* across instances, and the *contextual cues* among emotions. An EM-based incremental algorithm was devised for estimating the gold standard emotion annotation for each narrative line together with the model parameters. The objective was to determine how many crowdsourcing annotators have to provide annotations in the emotion labeling task in order for the aggregated annotation to be accurate. The experimental results demonstrate that incorporating *emotional consistencies* across instances enables the gold standard, i.e., the general consensus of large crowds, to be effectively estimated from the responses of a limited number (about five) of annotators. They also demonstrate that incorporating *contextual cues* among emotions improves the accuracy.

Emotion annotations are important metadata for narrative texts in digital libraries and emotion-oriented research. The annotation quality directly affects the quality of the research. Collecting high-quality annotations from both experts and large crowds can be expensive and time-consuming. The proposed models enable the cost of preparing high-quality annotations for emotion-oriented research to be reduced, with minimal degradation in the quality of the results.

Our exploration of this human computation issue produced promising results that encourage us to overcome the limitations of our present work and continue our study in this area. We plan to enhance our research efforts in several ways. First, our experiments were conducted on a small dataset, two children's narratives. We plan to explore whether the proposed models are also accurate for larger datasets. Second, *bi-gram*s were used to learn the transition distribution over conjoint-emotions. It would be worth investigating whether *tri-gram*s or higher *n-gram*s yield better results, especially when dealing with larger contexts. Third, every instance was annotated by an equal number of annotators. However, for simple instances, few (one or two) annotators may be sufficient. This means that taking into account instance difficulty could further reduce preparation cost. We thus plan to design an effective mechanism for automatically identifying the difficulties of instances, such as using the annotators' annotation histories and the time needed for annotating an instance.

Our proposed models are for estimating the gold standard for crowdsourcing tasks that implicitly contain information for "consistency" and "context". This general idea may also be applicable to tasks such as art style annotation (music composers and movie directors generally have distinctive styles, which generally remains consistent in their works), *parts of speech* tagging (a word's tag may depend on the tags of neighboring words), and social network analysis (a text message and the replies often have consistent emotional tendency, and the feelings of a poster and repliers may have contextual cues). We plan to perform more experiments to test the feasibility and validity of the proposed models across different domains.

Acknowledgements

This work was supported in part by JSPS KAKENHI 24650061.

8. REFERENCES

[1] C. O. Alm. Characteristics of high agreement affect annotation in text. In *Proceedings of the Fourth Linguistic Annotation Workshop (LAW)*, pages 118–122. Association for Computational Linguistics, 2010.

[2] C. O. Alm, D. Roth, and R. Sproat. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pages 579–586. Association for Computational Linguistics, 2005.

[3] L. F. Barrett, B. Mesquita, K. N. Ochsner, and J. J. Gross. The experience of emotion. *Annual Review of Psychology*, 58:373, 2007.

[4] K. Bischoff, C. S. Firan, W. Nejdl, and R. Paiu. How do you feel about dancing queen?: deriving mood & theme annotations from user tags. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital libraries*, pages 285–294. ACM, 2009.

[5] R. A. Calvo and S. D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *Affective Computing, IEEE Transactions on*, 1(1):18–37, 2010.

[6] C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *Information Theory, IEEE Transactions on*, 14(3):462–467, 1968.

[7] A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied Statistics*, pages 20–28, 1979.

[8] M. d. R. D. Dias, S. d. S. B. L. d. Faria, S. C. M. Ibrahim, et al. I'm like a river: a health education instrument for stuttering. *Revista de Psicologia da IMED*, 5(2), 2013.

[9] P. Donmez, J. G. Carbonell, and J. Schneider. Efficiently learning the accuracy of labeling sources for selective sampling. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM, 2009.

[10] L. Duan, S. Oyama, H. Sato, and M. Kurihara. Separate or joint? estimation of multiple labels from crowdsourced annotations. *Expert Systems with Applications*, 41(13):5723–5732, 2014.

[11] P. Ekman. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200, 1992.

[12] S. Ertekin, C. Rudin, and H. Hirsh. Approximating the crowd.

[13] J. Howe. Crowdsourcing: A definition. *Crowdsourcing: Tracking the Rise of the Mmateur*, 2006.

[14] X. Hu and J. S. Downie. Improving mood classification in music digital libraries by combining lyrics and audio. In *Proceedings of the 10th annual Joint Conference on Digital libraries*, pages 159–168. ACM, 2010.

[15] S. Kim, F. Li, G. Lebanon, and I. Essa. Beyond sentiment: The manifold of human emotions. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 360–369, 2013.

[16] J. H. Lee and X. Hu. Generating ground truth for music mood classification using mechanical turk. In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 129–138. ACM, 2012.

[17] C. H. Lin, Mausam, and D. S. Weld. Crowdsourcing control: Moving beyond multiple choice. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI). Catalina Island, CA, USA. August 2012*, pages 491–500, 2012.

[18] Z. Liu, S. Wang, Z. Wang, and Q. Ji. Implicit video multi-emotion tagging by exploiting multi-expression relations. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–6. IEEE, 2013.

[19] A. Nakamura. Kanjo hyogen jiten [dictionary of emotive expressions]. *Tokyodo*, 1993.

[20] S. Nowak and S. Rüger. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the International Conference on Multimedia Information Retrieval*, pages 557–566. ACM, 2010.

[21] C. K. Ogden and I. A. Richards. The meaning of meaning. *New York*, 1923.

[22] S. Oyama, Y. Baba, Y. Sakurai, and H. Kashima. Accurate integration of crowdsourced labels using workers' self-reported confidence scores. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2554–2560. AAAI Press, 2013.

[23] M. Ptaszynski, H. Dokoshi, S. Oyama, R. Rzepka, M. Kurihara, K. Araki, and Y. Momouchi. Affect analysis in context of characters in narratives. *Expert Systems with Applications*, 40(1):168–176, 2013.

[24] J. A. Russell and J. M. Fernández-Dols. *The Psychology of Facial Expression*. Cambridge university press, 1997.

[25] M. A. M. Shaikh, A. R. F. Rebordao, and K. Hirose. Improving tts synthesis for emotional expressivity by a prosodic parameterization of affect based on linguistic analysis. In *Proceedings of the 5th International Conference on Speech Prosody (SP5)*, 2010.

[26] V. S. Sheng, F. Provost, and P. G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 614–622. ACM, 2008.

[27] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast—but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263. Association for Computational Linguistics, 2008.

[28] A. Sorokin and D. Forsyth. Utility data annotation with amazon mechanical turk. *Urbana*, 51(61):820, 2008.

[29] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. P. Vlahavas. Multi-label classification of music into emotions. In *International Society for Music Information Retrieval (ISMIR)*, volume 8, pages 325–330, 2008.

[30] G. Tsoumakas, I. Katakis, and I. Vlahavas. Mining multi-label data. In *Data Mining and Knowledge Discovery Handbook*, pages 667–685. Springer, 2010.

[31] P. Welinder, S. Branson, P. Perona, and S. J. Belongie. The multidimensional wisdom of crowds. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2424–2432, 2010.

[32] J. Whitehill, T.-f. Wu, J. Bergsma, J. R. Movellan, and P. L. Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2035–2043, 2009.

[33] S. C. Widen, J. A. Russell, and A. Brooks. Anger and disgust: Discrete or overlapping categories? In *2004 APS Annual Convention, Boston College, Chicago, IL*, 2004.