Proceedings of the International MultiConference of Engineers and Computer Scientists 2015 Vol I,
IMECS 2015, March 18 - 20, 2015, Hong Kong

Establishing Relationships between Emotion Taxonomies Using the Vector Space Model

Lei Duan, Satoshi Oyama, Masahito Kurihara, Haruhiko Sato

***Abstract*—Due to different aspects that emotion-oriented research looks to capture, the emotion taxonomy used often differs among research efforts. Therefore, it is hard to coordinate the research efforts using different emotion taxonomies. On the other hand, due to the multiplicity of "emotion", emotion annotations more naturally fit the paradigm of multi-label classification since one instance (such as a sentence) may evoke a combination of multiple emotions. We thus propose bridging the gap between emotion taxonomies in the multi-label domain by leveraging the Vector Space Model and crowdsourcing. The relationships between source emotion taxonomy and target emotion taxonomy are formalized as a transformation mapping, which is established using the gold emotion annotations in the source taxonomy and the crowdsourced emotion annotations in the target taxonomy. Using the established mapping, associated emotions in the target taxonomy for an instance can be directly obtained according to its associated emotions in the source taxonomy. Experimental results on the real-world data demonstrate that the mapping established using the proposed models enables the gold emotions in the target taxonomy to be effectively estimated.**

***Index Terms*—emotion-oriented research, emotion taxonomy, vector space model, crowdsourcing, transformation mapping.**

I. Introduction

IN the area of emotion-oriented research, the first step towards solving an emotion-related problem is to adopt or create an appropriate emotion taxonomy. Simply put, an emotion taxonomy can be represented by the candidate emotion categories (also referred to as classes, labels, terms, or tags) applied to the collected instances (such as narrative sentences, movie clips, or music pieces). Due to different aspects that emotion-oriented research looks to capture or just inconsistency in terminology usage, the taxonomy used differs among research efforts. Even though the taxonomy of Ekman's six emotions [1] {*happiness*, *fear*, *anger*, *surprise*, *disgust*, *sadness*} has been used very broadly to cover a wide range of emotion-oriented research [2], [3], other emotion taxonomies are also adopted. For example, Trohidis *et al.* [4] used other six emotions {*amazed-surprised*, *happy-pleased*, *relaxing-calm*, *quiet-still*, *sad-lonely*, *angry-fearful*} based on the Tellegen-Watson-Clark taxonomy [5] to conduct the automated detection of emotion in music; the taxonomy used by the *manifold* emotion analyzer [6] consists of a collection of 32 emotions; the WordNet-Affect [7] even hierarchically organized a collection of 288 emotions.

Moreover, social and cultural background plays a significant role in emotion interpretation. A noteworthy example is that different from the English language-oriented research listed above, a lot of Japanese language-oriented research (e.g., [8], [9]) prefers to employ the taxonomy of Nakamura's ten emotions [10]: {*ki/yorokobi* (joy, delight, happiness; later referred to as **happiness**), *dō/ikari* (**anger**), *ai/aware* (sorrow, sadness, gloom; later referred to as **sadness**), *fu/kowagari* (**fear**), *chi/haji* (shame, shyness, bashfulness; later referred to as **shame**), *kō/suki* (liking, fondness; later referred to as **fondness**), *en/iya* (dislike, detestation, disgust; later referred to as **disgust**), *kō/takaburi* (**excitement**), *an/yasuragi* (**relief**), *kyō/odoroki* (surprise, amazement; later referred to as **surprise**)}. A complete discussion on the taxonomies used for emotion-oriented research is beyond the scope of this paper but can be found in Calvo *et al.* [11].

Unfortunately, it is certain that, as of date, there does not exist an authoritative emotion taxonomy, meaning that no general emotion taxonomy has yet been agreed on [12]. Therefore, emotion-oriented research is faced with the problem that the emotion taxonomy used in one research may be different to the emotion taxonomy used in another research. Although different emotion taxonomies are founded on different psychological theories and fit specific purposes of particular emotion-oriented research in various fields, complications still arise when they are used:

1) It is hard to coordinate emotion-oriented systems using different emotion taxonomies to allow them work together.
2) An enormous number of emotion annotations is generally necessary for emotion-oriented research to be used as training data or reference material. Such annotations cannot be shared among studies using different emotion taxonomies, which results in waste of resources.
3) The lack of harmonization poses barriers among different emotion taxonomies and the barriers further complicate comparison experiments and benchmarking studies.

Therefore, how to establish relationships between emotion taxonomies is a necessary and important problem.

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 may evoke a combination of multiple emotion categories. It has been demonstrated that a single emotion category is unable to represent all possible emotional manifestations [13] and that some emotional manifestations are a combination of several emotion categories [14].

Given all that, it is both important and necessary to bridge the gap between emotion taxonomies in the multi-label domain. One typical example of the gap is that a text-oriented emotion detector classifies a sentence (e.g., "John has already killed three kittens on the bridge.") into associated emotions (e.g., {anger, disgust, sadness}) in the Ekman taxonomy. But a text-to-speech synthesis requires the sentence with its associated emotions (e.g., {sad-lonely, angry-fearful}) in the

Lei Duan, Satoshi Oyama, Masahito Kurihara and Haruhiko Sato are with Division of Synergetic Information Science in Graduate School of Information Science and Technology, Hokkaido University, Sappro, Japan, 060-0814. E-mail: duan@ec.hokudai.ac.jp, oyama@ist.hokudai.ac.jp, kurihara@ist.hokudai.ac.jp, haru@complex.ist.hokudai.ac.jp.

Tellegen-Watson-Clark taxonomy as the input for affective pronunciation. This means that the output of the emotion detector cannot be used as the input of the text-to-speech synthesis since the taxonomy used for the output does not match the taxonomy used for the input. To this end, our primary goal is to establish a mapping from emotion sets in one (source) taxonomy to emotion sets in another (target) taxonomy (e.g. {anger, disgust, sadness} $\rightarrow$ {sad-lonely, angry-fearful}) so that emotion-oriented systems using different emotion taxonomies can become interoperable.

We propose leveraging crowdsourcing to achieve this goal, since that on-line crowdsourcing services can provide an inexpensive means for outsourcing various kinds of tasks to hundreds of thousands of people, and it is being used more frequently in the annotation community. Suppose that there is a large collection of (sentence, associated emotions) paired data, where the associated emotions are selected from emotion taxonomy X. However, the information of emotion taxonomy Y is considered more important. We can first (randomly or deliberately) select a part of all sentences, and ask crowdsourcing annotators to assign the associated emotions in taxonomy Y to each of the selected sentences. Our goal is to establish the mapping from taxonomy X (the source taxonomy) to taxonomy Y (the target taxonomy) according to the obtained triplets: {(sentences, associated emotions in taxonomy X, assigned emotions in taxonomy Y)}. Using the established mapping, we can directly obtain the associated emotions in taxonomy Y for each sentence according to its associated emotions in taxonomy X.

Although data can be obtained from a crowdsourcing service at very low cost (time and expense), crowdsourcing annotators are rarely trained and generally do not have the abilities needed to accurately perform the offered task. Therefore, ensuring the quality of the responses is one of the biggest challenges in crowdsourcing.[1] A promising quality control strategy is to introduce redundancy by asking several annotators to perform each task. There are two main methods for processing crowdsourced annotations provided by multiple annotators. One is to aggregate the annotations to produce a reliable annotation, and establish models using a set of aggregated annotations. The other is to establish models directly from the obtained annotations. In this paper, we compared these two methods for establishing the mapping from the source emotion taxonomy to the target emotion taxonomy in Sections III-A and III-B.

A. Probabilistic Model

Let $\mathcal{X}$ be a subset of emotions in taxonomy X, and $\mathcal{Y}$ be a subset of emotions in taxonomy Y. The simplest way to establish the mapping from taxonomy X to taxonomy Y is the maximum likelihood estimation (MLE), where the optimum target emotion set in Y for a source emotion set $\mathcal{X}$ is the one that achieves the maximum likelihood:

$$\arg\max_{\mathcal{Y}\subseteq Y} \Pr\left(\mathcal{Y} \mid \mathcal{X}\right).$$

Because the states of emotions in both source taxonomy X and target taxonomy Y are binary-valued, the MLE needs to estimate a target emotion set for each of the $2^{|X|}$ different source emotion sets in taxonomy X. This means that it is necessary to at least select $2^{|X|}$ sentences to cover all possible sub-sets in source taxonomy X. But at the practical level, it is too expensive and nearly impossible to select a sufficient number of sentences for every perspective and ask annotators to annotate them. Therefore, it is necessary to exploit more robust methods for solving this problem.

B. Vector Space Model

The Vector Space Model (VSM) is one of the most widely used models for information retrieval, mainly because of its conceptual simplicity and the appeal of the underlying metaphor of using spatial proximity for semantic proximity. In a typical VSM for information retrieval, each document is formalized as a vector, and each dimension corresponds to a separate term. If a term occurs in the document then its value in the vector is non-zero. In the mapping problem, emotion annotations (emotion category sets) and emotion categories can be seen as the counterparts of documents and terms. By leveraging the VSM, the problem of establishing a mapping $2^X \rightarrow 2^Y$ can be solved by establishing a linear transformation mapping f:

$$\mathbf{y} = f\left(\mathbf{x}\right) = \mathbf{A}\mathbf{x}\,, \tag{1}$$

where $\mathbf{x}$ and $\mathbf{y}$ are binary vectors corresponding to $\mathcal{X}$ and $\mathcal{Y}$ in Section I-A, and $\mathbf{A}$ is a real-valued $|Y| \times |X|$ transformation matrix of mapping f.[2] Using the VSM, the state of each emotion in target taxonomy Y can be seen as a linear combination of the emotions in source taxonomy X. Since that the VSM can overcome the shortage of the MLE with establishing the mapping from a limited number of annotations, in this paper, we focus on how to leverage the VSM to establish relationships (transformation mapping) between emotion taxonomies from crowdsourced annotations.

The remainder of this paper is organized as follows. Section II provides background material by introducing related research. Section III introduces the three proposed VSM-based models for establishing relationships between emotion taxonomies. Section IV describes the experimental design and discusses the results obtained by applying the proposed models to real-world data. Section V summarizes the main points and suggests several future research directions.

II. Background and Related Work

A. Emotion-oriented Research

Humans, by nature, can be emotionally affected by literature, music, fine art, etc. 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 [15] and therapeutic education of children with communication disorders [16]. Many researchers have thus concentrated on this area. Alm *et al.* [3] 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, using the sparse network of winnows (*SNoW*) learning architecture. Kim *et al.* [6] modeled emotion as

[1]For a detailed discussion, see Section II-B

[2]The detailed implementation is described in Section III.

ISBN: 978-988-19253-2-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

a continuous *manifold* and constructed a statistical model connecting it to documents and to a discrete set of emotions. Similar to our research, Danisman *et al.* [17] used the VSM for emotion classification in text. They showed that VSM-based classification on short sentences can be as good as other well-known classifiers including Naïve Bayes, SVM, and ConceptNet.

Several researchers concentrated on exploiting the multifaceted nature of emotion. Trohidis *et al.* [4] modeled emotion detection in music pieces as a multi-label classification task. Ptaszynski *et al.* [8] 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 Pelachaud *et al.* [18].

B. Crowdsourcing and Quality Control

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[3] (MTurk) and CrowdFlower[4], the concept of crowdsourcing has been successfully leveraged in various areas of computer science research, such as natural language processing [19]. There have also been several attempts in the emotion detection domain. Alm [2] 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.

There is no guarantee that all crowdsourcing annotators are sufficiently competent to complete the offered tasks. Therefore, ensuring the quality of the results is one of the biggest challenges in crowdsourcing. In addition to simple strategies such as offering incentive programs, 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.* [20] 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.* [21] 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. Snow *et al.* [19] demonstrated that by using an automatic bias correction algorithm, MTurk can be used effectively for a variety of natural language annotation tasks.

In the multi-label domain, Duan *et al.* [9] 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.* [22] studied inter-annotator agreement for multi-label 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. However, they did not answer the question of how many crowdsourcing annotators are needed to obtain quality comparable to that of expert annotators.

[3]http://www.mturk.com

[4]http://crowdflower.com

III. Statistical Models

Problem Formulation: Let $\mathbf{I}$ be the set of sentences, and X be the source emotion taxonomy. $\mathcal{X}_i \subseteq X$ $(i \in \mathbf{I})$ denotes the associated emotions of sentence i in taxonomy X. Let K be the set of crowdsourcing annotators, Y be the target emotion taxonomy. $I \subset \mathbf{I}$ denotes the set of sentences annotated with taxonomy Y. $\mathcal{K}_i \subseteq K$ $(i \in I)$ denotes the set of annotators who annotated sentence i with taxonomy Y. $\mathcal{Y}_i^k \subseteq Y$ $(k \in \mathcal{K}_i, i \in I)$ denotes the emotions assigned by annotator k for sentence i in taxonomy Y. Let $T = \{\mathcal{X}_i, \mathcal{Y}_i^k : k \in \mathcal{K}_i, i \in I\} \subseteq 2^X \times 2^Y$ be the set of obtained examples, where 2^X and 2^Y are the power sets of X and Y. The goal is to establish a mapping f: $2^X \to 2^Y$ from T, where f is chosen from a hypothesis class F, such that a loss function: $F \times 2^X \times 2^Y \to \mathbb{R}$ is minimized. Using the established mapping f, the associated emotions $\mathcal{Y}_i \subseteq Y$ for sentence $i \in \mathbf{I}$ can be directly obtained according to $\mathcal{X}_i$ without any extra effort.

A. VSM with Aggregated Annotations

We first defined two indicator vectors $\mathbf{x}$ and $\mathbf{y}$ corresponding to $\mathcal{X} \subseteq X$ and $\mathcal{Y} \subseteq Y$. The elements of the vectors are defined as follows (all vectors are assumed to be column vectors in this paper):

$$\mathbf{x}_{(\imath)} = \begin{cases} 1, & \imath \in \mathcal{X} \\ -1, & \imath \notin \mathcal{X} \end{cases} \quad \mathbf{y}_{(\imath)} = \begin{cases} 1, & \imath \in \mathcal{Y} \\ -1, & \imath \notin \mathcal{Y} \end{cases} .$$

This means that $\mathbf{x}$ (or $\mathbf{y}$) is a binary vector: if an element's corresponding emotion exists in $\mathcal{X}$ (or $\mathcal{Y}$), its value is 1, and -1 otherwise.

We have described in Section I-B that using the VSM, establishing the mapping f can be solved by constructing the real-valued $|Y| \times |X|$ linear transformation matrix $\mathbf{A}$ in Equation (1). We propose using the distance between two vectors to formalize the loss function, such that $\mathbf{A}$ can be estimated by minimizing the sum of the distances between the mapping vector and the aggregated vector of all annotated sentences:

$$\mathbf{A}^\star = \arg\min_{\mathbf{A}} \left\{ \sum_{i \in I} \mathrm{dis}\left(\mathbf{A}\mathbf{x}_i, \bar{\mathbf{y}}_i\right) \right\}, \tag{2}$$

where $\mathrm{dis}(\cdot,\cdot)$ denotes the distance between two vectors, $\bar{\mathbf{y}}_i$ is the vector corresponding to the aggregated annotation of the annotation set $\{\mathcal{Y}_i^k : k \in \mathcal{K}_i\}$. The c-th $(1 \le c \le |Y|)$ row in matrix $\mathbf{A}$ is the transformation vector from the states (exist or not) of the emotions in the source emotion set $\mathcal{X}_i$ to the state of the c-th emotion in the aggregated emotion set $\bar{\mathcal{Y}}_i$.

There are many aggregation strategies proposed by crowdsourcing researchers. To simplify the computation, we adopted the majority vote, the most commonly used strategy. This means that $\mathbf{y}_i$ is the corresponding vector of the most frequently annotated emotion set among annotations $\{\mathcal{Y}_i^k : k \in \mathcal{K}_i\}$.

After constructing the optimal transformation matrix $\mathbf{A}^\star$, for any sentence with its associated emotions $\mathcal{X}$ in source taxonomy X, we can obtain its associated emotions $\mathcal{Y}$ in target taxonomy Y as

$$y \begin{cases} \in \mathcal{Y}, & \mathbf{A}^{\star}\mathbf{x}_{(y)} > 0 \\ \notin \mathcal{Y}, & \mathbf{A}^{\star}\mathbf{x}_{(y)} \leq 0 \,, \end{cases}$$

where $y \in Y$.

B. VSM with Combination of Crowdsourced Annotations

1) Ordinary Combination: Though we can use the aggregated crowdsourced annotations to establish the mapping, the information on the distribution of annotators' responses is missing in the establishing process, e.g., the assigned emotion annotations in taxonomy Y for each sentence are truncated to one binary-valued vector. We introduce a combination of multiple annotations to establish the mapping between two emotion taxonomies in a crowdsourced setting directly from multiple annotators.

The ordinary combination approach treats responses given by different annotators equally. The process of finding the optimal transformation matrix is implemented as

$$\mathbf{A}^{\star} = \arg\min_{\mathbf{A}} \left\{ \sum_{i \in I} \sum_{k \in \mathcal{K}_i} \mathrm{dis}\left(\mathbf{A}\mathbf{x}_i, \mathbf{y}_i^k\right) \right\} , \qquad (3)$$

where the c-th ($1 \leq c \leq |Y|$) row in matrix $\mathbf{A}$ is the transformation vector from the states of the emotions in the source emotion set $\mathcal{X}_i$ to the state of the c-th emotion in the annotation $\mathcal{Y}_i^k$.

2) Weighted Combination: Both Equations (2) and (3) treat responses given by different annotators equally. In crowdsourcing, it is natural to assume that some annotators perform better than others. To compensate for the variability in the accuracy of crowdsourcing annotators and thereby establish a more accurate mapping, we impose an additional weighting measure on the establishing process: giving more weight to the annotations provided by high-performance annotators and less weight to those provided by low-performance annotators.

As a result, the process of constructing the optimal transformation matrix (Equation (3)) is rewritten as

$$\mathbf{A}^{\star} = \arg\min_{\mathbf{A}} \left\{ \sum_{i \in I} \sum_{k \in \mathcal{K}_i} w_i^k \cdot \mathrm{dis}\left(\mathbf{A}\mathbf{x}_i, \mathbf{y}_i^k\right) \right\} . \qquad (4)$$

Note that the weights are specific to both a sentence and an annotator. This means that the mapping target vector ($\mathbf{A}^{\star}\mathbf{x}$) should be nearer to the annotations given by more accurate annotators for more easily comprehended sentences.

If we assume that most annotators give reliable annotations, an annotator is more qualified if the annotations provided by him/her are more similar to the annotations provided by other annotators. The similarity between the annotation provided by annotator k and the annotations provided by other annotators for sentence i is defined as

$$s_i^k = \frac{\sum_{\hat{k} \in \mathcal{K}_i, \hat{k} \neq k} \mathrm{sim}\left(\mathbf{y}_i^k, \mathbf{y}_i^{\hat{k}}\right)}{|\mathcal{K}_i| - 1} ,$$

where the similarity between two indicator vectors is defined as

$$\mathrm{sim}\left(\mathbf{y}_1, \mathbf{y}_2\right) = \frac{\sum_{y \in Y} \left(2 - |\mathbf{y}_{1(y)} - \mathbf{y}_{2(y)}|\right)}{2 \cdot |Y|} , \qquad (5)$$

i.e., the proportion of emotions with the same state between $\mathbf{y}_1$ and $\mathbf{y}_2$. Finally, the weight of annotator k for sentence i in Equation (4) is defined as

$$w_i^k = \frac{s_i^k}{\sum_{\hat{k} \in \mathcal{K}_i} s_i^{\hat{k}}} .$$

This is to ensure that the sum of the annotator weights for each sentence equals to 1.

Note that our proposed models can also be used in the single-label domain as well, where the associated emotion in target taxonomy Y is the one with the maximum value in the mapping vector:

$$y^{\star} = \arg\max_{y \in Y} \mathbf{A}^{\star}\mathbf{x}_{(y)} .$$

IV. Empirical Study

To evaluate the effectiveness of the proposed models, we needed narratives in which the sentences 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. Therefore, children's narratives are commonly used in emotion-oriented research [3], [16]. 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].

In the experiment, we chose two typical emotion taxonomies as the source taxonomy and the target taxonomy. One is Ekman's emotion taxonomy (including six emotion categories), which is the most commonly used emotion taxonomy in emotion-oriented research. The other is Nakamura's emotion taxonomy (including ten emotion categories), which were taken from the "Emotive Expression Dictionary"[10] and were proven to be appropriate for the Japanese language and culture [8]. To perform *mutual validation* between two taxonomies, both of the two emotion taxonomies are annotated to the sentences in these two narratives. All annotators were native Japanese language speakers. Both sentences and Nakamura's taxonomy were presented in their original Japanese form. Ekman's taxonomy was presented in its original English form with Japanese explanations. Crowdsourcing annotators were asked to read the narrative sentences, and spontaneously indicate the character's emotions expressed in each sentence. If none of the candidate emotions was felt, the annotator would check *neutral*. The two taxonomies are presented separately to arbitrary annotators, and few, if any, of them annotated sentences with both the taxonomies.

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.

[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] http://www.aozora.gr.jp
[8] http://www.lancers.jp

ISBN: 978-988-19253-2-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Table I
TRANSFORMATION ACCURACIES

Training narrative	Ekman→Nakamura		Nakamura→Ekman	
"Love"	AA	0.779	AC	0.814
	OC	0.816	OC	0.828
	WC	**0.841**	WC	**0.843**
"Apple"	AA	0.856	AC	0.865
	OC	**0.877**	OC	0.884
	WC	0.875	WC	**0.902**

Therefore, we obtained gold emotions for each sentence by having the sentence annotated 30 times and then taking the majority vote. That is, an emotion category is in the gold emotion set of a sentence when it is assigned by more than half of the annotators for that sentence.

To implement the *mutual validation*, we used the sentences in one of the narratives to establish two mappings (Ekman→Nakamura, Nakamura→Ekman), and then using an established mapping and the gold emotions in the source taxonomy of each sentence in the two narratives to estimate the associated emotions in the target taxonomy for each of the sentences. The mappings were estimated using the following three models:

- *AA*: VSM with **A**ggregated **A**nnotations;
- *OC*: VSM with **O**rdinary **C**ombination of crowdsourced annotations;
- *WC*: VSM with **W**eighted **C**ombination of crowdsourced annotations.

Since both the estimated emotions and the gold emotions for a sentence can be regarded as a binary vector, the average *Simple Matching Coefficient* (Equation (5)) is used to evaluate the performance of the proposed models, i.e., the average proportion of correct emotions between the estimated emotions and the gold emotions for all sentences.

The experiment results are shown in Table I. For both narratives to be used to establish the transformation mapping, both the OC and WC models achieved better accuracies than the AA model. This means that the *combination* strategy is more effective than the *aggregation* strategy. Moreover, in most cases, the WC model performed better than the OC model. This demonstrates that crowdsourcing annotators have variable levels of expertise, so it is better to treat the variable-quality crowdsourced annotations considering annotator's expertise. Finally, transformation accuracies from Nakamura's taxonomy to Ekman's taxonomy (the third column) are higher than the accuracies of the reversed transformation (the second column). The reason for this phenomenon is that there are ten emotions in Nakamura's taxonomy and six emotions in Ekman's taxonomy. The transformation from taxonomy with more emotions to taxonomy with less emotions tends to be more accurate than the transformation from taxonomy with less emotions to taxonomy with more emotions.

V. CONCLUSION

In emotion-oriented research, different emotion taxonomies are employed on the basis of what information is considered important for the goals being pursued, cultural or social differentiating factors, researchers' personal preferences, or just inconsistency in terminology usage. There does not exist an authoritative emotion taxonomy, meaning that there is no formal agreement on what kinds of emotions exist and how to define them. This further complicates emotion-oriented research. From this viewpoint, we focused on leveraging crowdsourcing to establish relationships between different emotion taxonomies. This can benefit in a few ways. First, emotion-related applications can have the advantage of using emotion classifiers that have already been vetted, and one that may also come with annotated emotion corpus, which can be used to train other classifiers or just supplement the original dataset if the usage restrictions on the corpus allow for that. Finally, it makes different emotion classifiers and applications comparable. We proposed three models, *AA*, *OC*, and *WC*. In order to test the efficiency of these models, we conducted the experiment on real-world crowdsourcing data with the sentences with two narratives and two typical emotion taxonomies. Experimental results demonstrate that the transformation mapping established using the proposed models enables the gold emotions in the target taxonomy for a narrative sentence to be effectively estimated, directly from its associated emotions in the source emotion taxonomy.

Our experiment were conducted on a small dataset, two children's narratives. We plan to explore whether the proposed models are also accurate for larger datasets. Our proposed models are for establishing relationships between emotion taxonomies. This general idea may also be applicable in other domain. For most labeling problems, a number of candidate class terms are used, and of course, not everyone will agree on what a "standard" list of terms should be – such as the emotion-oriented research illustrated above and film genre classification (taking the list of genres from IMDB[9] or etflix[10]). Therefore, we also plan to extend our research across different domains in future work.

ACKNOWLEDGMENT

This work was supported in part by JSPS KAKENHI 24650061.

REFERENCES

[1] P. Ekman, "An argument for basic emotions," *Cognition & Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

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

[3] 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)*. Association for Computational Linguistics, 2005, pp. 579–586.

[4] 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)*, vol. 8, 2008, pp. 325–330.

[5] A. Tellegen, D. Watson, and L. A. Clark, "On the dimensional and hierarchical structure of affect," *Psychological Science*, vol. 10, no. 4, pp. 297–303, 1999.

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

[7] C. Strapparava and A. Valitutti, "Wordnet affect: an affective extension of wordnet." in *LREC*, vol. 4, 2004, pp. 1083–1086.

[8] 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*, vol. 40, no. 1, pp. 168–176, 2013.

[9] http://www.imdb.com/genre
[10] http://www2.netflix.com/allgenreslist

ISBN: 978-988-19253-2-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

[9] L. Duan, S. Oyama, H. Sato, and M. Kurihara, "Separate or joint? estimation of multiple labels from crowdsourced annotations," *Expert Systems with Applications*, vol. 41, no. 13, pp. 5723–5732, 2014.

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

[11] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *Affective Computing, IEEE Transactions on*, vol. 1, no. 1, pp. 18–37, 2010.

[12] M. Schröder, *Speech and Emotion Research: An overview of research frameworks and a dimensional approach to emotional speech synthesis*. Institut für Photetik, Universität des Saarlandes, 2004.

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

[14] 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.

[15] 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.

[16] 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*, vol. 5, no. 2, 2013.

[17] T. Danisman and A. Alpkocak, "Feeler: Emotion classification of text using vector space model," in *AISB 2008 Convention Communication, Interaction and Social Intelligence*, vol. 1, 2008, p. 53.

[18] C. Pelachaud, *Emotion-oriented systems*. John Wiley & Sons, 2013.

[19] 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*. Association for Computational Linguistics, 2008, pp. 254–263.

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

[21] 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)*, 2009, pp. 2035–2043.

[22] 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*. ACM, 2010, pp. 557–566.

ISBN: 978-988-19253-2-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)