Crowdsourced Semantic Matching of Multi-Label Annotations

Lei Duan, Satoshi Oyama, Masahito Kurihara and Haruhiko Sato
Graduate School of Information Science and Technology, Hokkaido University
Sapporo, Japan
duan@ec.hokudai.ac.jp, {oyama, kurihara}@ist.hokudai.ac.jp, haru@complex.ist.hokudai.ac.jp

Abstract

Most multi-label domains lack an authoritative taxonomy. Therefore, different taxonomies are commonly used in the same domain, which results in complications. Although this situation occurs frequently, there has been little study of it using a principled statistical approach. Given that (1) different taxonomies used in the same domain are generally founded on the same latent semantic space, where each possible label set in a taxonomy denotes a single semantic concept, and that (2) crowdsourcing is beneficial in identifying relationships between semantic concepts and instances at low cost, we proposed a novel probabilistic cascaded method for establishing a semantic matching function in a crowdsourcing setting that maps label sets in one (source) taxonomy to label sets in another (target) taxonomy in terms of the semantic distances between them. The established function can be used to detect the associated label set in the target taxonomy for an instance directly from its associated label set in the source taxonomy without any extra effort. Experimental results on real-world data (emotion annotations for narrative sentences) demonstrated that the proposed method can robustly establish semantic matching functions exhibiting satisfactory performance from a limited number of crowdsourced annotations.

1 Introduction

In a multi-label domain, each instance is associated with the subset of candidate labels (also referred to as classes, categories, terms, or tags) that most appropriately denote the relationship between a semantic concept and the instance. The first step towards solving a problem in a multi-label domain is to adopt or construct an appropriate taxonomy, i.e., the candidate label set applied to the collected instances. In most multi-label domains, there is no formal agreement on what kinds of labels exist and how to define them, and of course, not everyone will agree on what a "standard" taxonomy should be in that domain. Therefore, the taxonomy used may differ among projects in the same domain for various reasons such as differences in the aspects to be capture or simply researchers' personal preferences.

A typical example can be found in the domain of emotion-oriented research. It has been demonstrated that a single emotion category is unable to represent all possible emotional manifestations [Russell and Fernández-Dols, 1997] and that some emotional manifestations are a combination of several emotion categories [Widen *et al.*, 2004]. Therefore, due to the multifaceted nature of emotion, an instance (such as a sentence, a movie clip, or a music piece) is more naturally to be associated with a combination of multiple emotion categories. Even though the taxonomy of Ekman's six basic emotions [Ekman, 1992] (*happiness*, *fear*, *anger*, *surprise*, *disgust*, and *sadness*) has been used very broadly to cover a wide range of emotion-oriented research, other emotion taxonomies are also widely used. For example, Trohidis *et al.* [2008] used six other emotions {*amazed-surprised*, *happy-pleased*, *relaxing-calm*, *quiet-still*, *sad-lonely*, and *angry-fearful*} based on the Tellegen-Watson-Clark taxonomy [Tellegen *et al.*, 1999] for the automated detection of emotion in music. The taxonomy used by the *manifold* emotion analyzer [Kim *et al.*, 2013] consists of 32 emotions while the WordNet-Affect thesaurus [Strapparava and Valitutti, 2004] has a hierarchically organized collection of 288 emotions. Social and cultural factors also play a significant role in emotion interpretation. A noteworthy example of this is the difference between the English-language-oriented research mentioned above and Japanese-language-oriented research (e.g., [Ptaszynski *et al.*, 2013; Duan *et al.*, 2014]) which tends to use the taxonomy of Nakamura's ten emotions [Nakamura, 1993]. A complete discussion of the taxonomies used for emotion-oriented research is beyond the scope of this paper but can be found in Calvo and D'Mello [2010].

The lack of an authoritative emotion taxonomy means that emotion-oriented research is faced with the problem of inconsistency in taxonomy usage. For example, a text-oriented emotion detector classifies a sentence (e.g., Shyo: "John has already killed three kittens on the bridge.") into associated emotions (e.g., {sad-lonely, angry-fearful}) in the Tellegen-Watson-Clark taxonomy while a text-to-speech synthesis system requires the sentence and the associated emotions (e.g., {anger, disgust, sadness}) in Ekman's taxonomy as the input for affective pronunciation. This means the output of the emotion detector cannot be used as the input to the text-to-

speech synthesis system since the taxonomy used for the output does not match the taxonomy used for the input.

The same problem is found in other multi-label domains, such as film genre classification (using the list of genres from IMDB[1] or Netflix[2]) and text categorization (using Reuters Topics[3] or the Mozilla Directory[4]). The barriers among different taxonomies often result in complications, such as

1. making it hard to coordinate systems that use different taxonomies so that they work together (as described above);
2. disallowing multi-label annotations (used as training data or reference material) to be shared among systems using different taxonomies, resulting in a waste of resources;
3. complicating comparison experiments and benchmarking studies among systems using different taxonomies.

Given all this, it is both important and necessary to bridge the gap between taxonomies in the same multi-label domain.

A taxonomy is constructed for a certain latent semantic space where each possible label set in the taxonomy denotes a unique semantic concept. Although different taxonomies in the same domain are proposed on the basis of different theories and fit the specific purposes of particular systems in various fields, they are generally constructed for the same latent semantic space. This insight led us to our key idea: in the latent semantic space, the label sets in different taxonomies should be semantically similar even if they do not share any common label – as long as they are associated to the same instance. Therefore, our primary goal is to represent semantic relationships between label sets in one taxonomy and those in another taxonomy in terms of their proximity in the latent semantic space. More specifically, our goal is to establish a semantic matching function that maps label sets in one (source) taxonomy to label sets in another (target) taxonomy (e.g., $\{$anger, disgust, sadness$\} \rightarrow \{$sad-lonely, angry-fearful$\}$) in terms of the semantic distance between them.

Suppose that there is a large collection of <instance, associated label set> paired data, where the associated label sets are selected from taxonomy S, even though the information in taxonomy T is more important. Clearly, annotating all the instances using taxonomy T would be tedious. In other words, it is better to make use of the semantic concepts denoted by their associated label sets in taxonomy S to reduce cost. We can first (randomly or systematically) select a portion of all instances and assign the associated label sets in taxonomy T to each of the selected instances. Then, the semantic mapping from taxonomy S (the source taxonomy) to taxonomy T (the target taxonomy) can be established using the obtained triplets: $\{$<instance, associated label set in S, assigned label set in T>$\}$. Using the established mapping, we can detect the associated label set in taxonomy T for each (both annotated and unannotated) instance directly from its associated label set in taxonomy S without any extra effort. One of our main contributions here is to show how we exploit the transformation of semantic concepts from the source taxonomy to the target taxonomy.

We propose leveraging crowdsourcing to achieve this goal since crowdsourcing is beneficial in identifying relationships between semantic concepts and instances at low cost (time and expense). However, crowdsourcing annotators are rarely trained and generally do not have the abilities needed to accurately perform the annotation task. Therefore, ensuring the quality of the responses is one of the biggest challenges in crowdsourcing.[5] A promising quality control strategy is to introduce redundancy by asking several annotators to perform each task and then aggregating their annotations to produce a reliable annotation. Our study thus focused on how to exploit effective aggregation so that semantic matching functions can be accurately established in a crowdsourcing setting with a minimum of human effort.

2 Background and Related Work

2.1 Semantic Matching

Interoperability among people of different cultures and languages, having different viewpoints and using different terminology, has always been a huge problem and is an important problem to be solved in semantic matching. More specifically, semantic matching is a type of ontology matching that relies on semantic information encoded in ontologies, like classifications, XML schemas, and label sets in taxonomies (the scenario in this research), to identify those nodes in two structures that semantically correspond to one another. It has been applied in areas such as resource discovery, data integration, data migration, and query translation.

2.2 Multi-label Learning

In multi-label learning, each instance can be associated with multiple labels simultaneously. Learning from multi-label data is an emerging and promising research topic and has attracted significant attention, mainly motivated by applications such as topic categorization of news articles and web pages, semantic annotation of images and videos, and emotion analysis in music and narratives. A good survey on multi-label learning was presented by Tsoumakas *et al.* [2010].

2.3 Crowdsourcing and Quality Control

Crowdsourcing is an economical and efficient approach to performing tasks that are difficult for computers but relatively easy for humans. However, there is no guarantee that all crowdsourcing annotators are sufficiently competent to perform the offered tasks accurately. Therefore, ensuring the quality of the results is one of the biggest challenges in crowdsourcing. Dawid and Skene [1979] presented a model for inferring the unknown health state of a patient given diagnostic tests by several clinicians. This model is used as a learning tool in our proposed method, as described in Algorithm 3. Duan *et al.* [2014] investigated multi-label estimation from crowdsourced annotations in the multi-label learning domain, with flexible incorporation of label dependency into the label-generation process.

[1] http://www.imdb.com/genre

[2] http://dvd.netflix.com/AllGenresList

[3] http://vocab.org/reuters_topics/1.0/

[4] http://www.dmoz.org/

[5] For a detailed discussion, see Section 2.3

3 Statistical Models

Problem Formulation

Let $\mathbf{I}$ be the set of instances, S be the source taxonomy, and T be the target taxonomy. $s = \left(s^{(1)}, s^{(2)}, ..., s^{(|S|)}\right)$ and $t = \left(t^{(1)}, t^{(2)}, ..., t^{(|T|)}\right)$ are binary vectors: if a label is present in a label set, the value of the corresponding element in the vector is 1, and 0 otherwise. s_i $(i \in \mathbf{I})$ denotes the associated label vector of instance i in taxonomy S. Let $I \subset \mathbf{I}$ be the set of instances annotated using taxonomy T, K be the set of crowdsourcing annotators, and $\mathcal{K}_i \subseteq K$ $(i \in I)$ be the set of annotators who annotated instance i using taxonomy T. t_{ik} $(k \in \mathcal{K}_i, i \in I)$ denotes the label vector corresponding to the label set assigned by annotator k, for instance i. Let $E = \{s_i, t_{ik} : k \in \mathcal{K}_i, i \in I\}$ be the set of obtained examples. The goal is to establish a semantic matching function f: $\{0,1\}^{|S|} \rightarrow \{0,1\}^{|T|}$ from E such that $t = f(s)$ has the semantic concept most similar to that of s. Using the established function f, the associated label vector t_i for instance $i \in \mathbf{I}$ can be directly detected from s_i without any extra effort.

3.1 Conventional Method: Joint Maximum Likelihood Estimation (J-MLE)

In a probabilistic framework, the target label vector having the semantic concept most similar to that of source label vector s is the one that has the maximum likelihood:

$$f(s) = \underset{t \in \{0,1\}^{|T|}}{\arg\max} \Pr(t \mid s) \,. \tag{1}$$

The naïve solution for establishing the semantic matching function is to estimate the maximum likelihood using annotation frequencies of label vectors:

$$\Pr(t \mid s) = \frac{\sum_{i \in I} \sum_{k \in \mathcal{K}_i} [[s_i = s]] \cdot [[t_{ik} = t]]}{\sum_{i \in I} [[s_i = s]] \cdot |\mathcal{K}_i|} \,. \tag{2}$$

Here we define the notation $[[\cdot]]$ as

$$[[\textit{true}]] = 1, \; [[\textit{false}]] = 0 \,.$$

Pseudo-code for *J-MLE* is given in Algorithm 1 .

Algorithm 1 Joint Maximum Likelihood Estimation

Input: $\{s_i, t_{ik} : k \in \mathcal{K}_i, i \in I \subset \mathbf{I}\}$.
Output: $\{t_i : i \in \mathbf{I}\}$.
for each $\langle s, t \rangle \in \left\{ \left\langle \{0,1\}^{|S|}, \{0,1\}^{|T|} \right\rangle \right\}$ **do**
$\Pr(t \mid s) = \frac{\sum_{i \in I} \sum_{k \in \mathcal{K}_i} [[s_i = s]] \cdot [[t_{ik} = t]]}{\sum_{i \in I} [[s_i = s]] \cdot |\mathcal{K}_i|}$.
end for
$t_i = \arg\max \Pr(t_i \mid s_i)$ for each $i \in \mathbf{I}$.

Because the states of labels in both source taxonomy S and target taxonomy T are binary-valued (presence or absence), with *J-MLE*, a target label vector needs to be estimated for each of the $2^{|S|}$ possible source label vectors in taxonomy S. Therefore, at least $2^{|S|}$ instances must be selected to cover all possible subsets in source taxonomy S (in the extreme case that each instance is associated with a unique source label vector). In contrast, only the target label vectors that have been assigned to at least one instance can be considered as candidate output in the codomain of the function, which means source label vectors will never be mapped to the unassigned target label vectors. At the practical level, it is too expensive and nearly impossible to select a sufficient number of instances for every perspective and expect annotators to annotate them. Furthermore, *J-MLE* simply treats annotations given by different annotators equally under the assumption that all annotators are equally good. However, in crowdsourcing, it is safe to assume that annotators have a wide range of expertise. Therefore, it is necessary to adopt more robust methods that take into account annotator expertise.

3.2 Proposed Method: Cascaded Method

Our proposed probabilistic method for establishing a semantic matching function overcomes the weakness of *J-MLE* by transferring semantic concepts from the source taxonomy to the target taxonomy in a cascaded way. Figure 1(b) shows the causal structure of the proposed method. The proposed method can robustly solve Equation (2), i.e., the posterior probability distribution of a target label vector given a source label vector. We start by assuming that labels in the target taxonomy are statistically independent. This means that the completely general joint distribution in Equation (2) can be calculated as the product of the independent distributions of the candidate target labels:

$$\Pr(t \mid s) = \prod_{m=1}^{|T|} \Pr\left(t^{(m)} \mid s\right) \,. \tag{3}$$

Using Bayes' theorem, we have

$$\Pr\left(t^{(m)} \mid s\right) = \frac{\Pr\left(t^{(m)}\right) \cdot \Pr\left(s \mid t^{(m)}\right)}{\Pr(s)} \,. \tag{4}$$

We assume that the labels in the source taxonomy are statistically independent as well. Therefore, we can obtain the prior joint distribution and the posterior joint distribution over the source label vectors:

$$\Pr(s) = \prod_{n=1}^{|S|} \Pr\left(s^{(n)}\right) \,, \tag{5}$$

$$\Pr\left(s \mid t^{(m)}\right) = \prod_{n=1}^{|S|} \Pr\left(s^{(n)} \mid t^{(m)}\right) \,. \tag{6}$$

Again using Bayes' theorem, we have

$$\Pr\left(s^{(n)}\right) = \frac{\Pr\left(t^{(m)}\right) \cdot \Pr\left(s^{(n)} \mid t^{(m)}\right)}{\Pr\left(t^{(m)} \mid s^{(n)}\right)} \,. \tag{7}$$

Next we substitute Equation (7) for $\Pr\left(s^{(n)}\right)$ in Equation (5):

$$\begin{aligned} \Pr(s) &= \prod_{n=1}^{|S|} \frac{\Pr\left(t^{(m)}\right) \cdot \Pr\left(s^{(n)} \mid t^{(m)}\right)}{\Pr\left(t^{(m)} \mid s^{(n)}\right)} \\ &= \left[\Pr\left(t^{(m)}\right)\right]^{|S|} \cdot \prod_{n=1}^{|S|} \frac{\Pr\left(s^{(n)} \mid t^{(m)}\right)}{\Pr\left(t^{(m)} \mid s^{(n)}\right)} \,. \end{aligned} \tag{8}$$

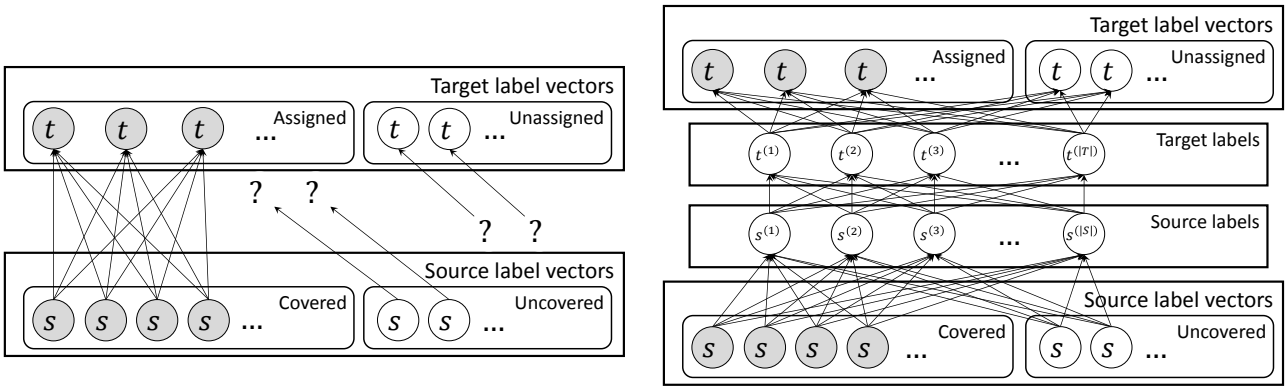

Figure 1: Graphical representation of semantic matching: (a) joint maximum likelihood estimation and (b) proposed cascaded method.

By substituting Equations (6) and (8) into Equation (4), we have

$$\Pr\left(t^{(m)} \mid s\right) = \frac{\prod_{n=1}^{|S|} \Pr\left(t^{(m)} \mid s^{(n)}\right)}{\left[\Pr\left(t^{(m)}\right)\right]^{|S|-1}}. \tag{9}$$

Finally, substituting Equation (9) into Equation (3) enables the semantic matching function in Equation (1) to be established using

$$f(s) = \underset{t \in \{0,1\}^{|T|}}{\arg\max} \prod_{m=1}^{|T|} \frac{\prod_{n=1}^{|S|} \Pr\left(t^{(m)} \mid s^{(n)}\right)}{\left[\Pr\left(t^{(m)}\right)\right]^{|S|-1}}. \tag{10}$$

As long as we can obtain the prior distributions and the posterior distributions over the target labels, i.e.,

$$\left\{\Pr\left(t^{(m)}\right) : m \in \{1, ..., |T|\}\right\} \text{ and}$$
$$\left\{\Pr\left(t^{(m)} \mid s^{(n)}\right) : m \in \{1, ..., |T|\}, n \in \{1, ... |S|\}\right\},$$

the semantic matching function $f(s)$ can be easily established using Equation (10). In other words, it is only necessary to estimate $|T| + |S| \cdot |T|$ probabilities. This is more robust than *J-MLE* because estimating the distribution over label vectors using the distributions over individual labels is easier than directly estimating the distribution over label vectors.

The proposed cascaded method is essentially a meta-strategy because distributions $\Pr\left(t^{(m)}\right)$ and $\Pr\left(t^{(m)} \mid s^{(n)}\right)$ can be obtained using existing learning tools. The simplest strategy is to estimate the maximum likelihoods using the annotation frequencies of labels. Pseudo-code for this strategy is given in Algorithm 2 .

Similar to *J-MLE*, this strategy treats annotations given by different annotators equally. To take annotator bias into consideration, we also adopted the Dawid-Skene model [Dawid and Skene, 1979]. This model was originally aimed at inferring the unknown health state of a patient given diagnostic tests by several clinicians, where the biases of the clinicians were modeled by a confusion matrix. Crowdsourcing annotators, the state of $s^{(n)}$, and the state of $t^{(m)}$ in this problem are the counterparts of clinicians, patients, and health states in the Dawid-Skene model. Pseudo-code for this strategy is given in Algorithm 3 .

Algorithm 2 Cascaded Maximum Likelihood Estimation

Input: $\{s_i, t_{ik} : k \in \mathcal{K}_i, i \in I \subset \mathbf{I}\}$.
Output: $\{t_i : i \in \mathbf{I}\}$.
for each $\langle s, t \rangle \in \left\{\left\langle \{0,1\}^{|S|}, \{0,1\}^{|T|} \right\rangle\right\}$ **do**
  for each $\langle n, m \rangle \in \{\langle \{0, 1, ..., |S|\}, \{0, 1, ..., |T|\} \rangle\}$ **do**
    - $\Pr\left(t^{(m)}\right) = \frac{\sum_{i \in I} \sum_{k \in \mathcal{K}_i} [[t_{ik}^{(m)} = t^{(m)}]]}{\sum_{i \in I} |\mathcal{K}_i|}$.
    - $\Pr\left(t^{(m)} \mid s^{(n)}\right) = \frac{\sum_{i \in I} \sum_{k \in \mathcal{K}_i} [[t_{ik}^{(m)} = t^{(m)}]] \cdot [[s_i^{(n)} = s^{(n)}]]}{\sum_{i \in I} \sum_{k \in \mathcal{K}_i} [[t_{ik}^{(m)} = t^{(m)}]]}$.
  end for
  $\Pr(t \mid s) = \prod_{m=1}^{|T|} \frac{\prod_{n=1}^{|S|} \Pr\left(t^{(m)} \mid s^{(n)}\right)}{\left[\Pr\left(t^{(m)}\right)\right]^{|S|-1}}$.
end for
$t_i = \arg\max \Pr(t_i \mid s_i)$ for each $i \in \mathbf{I}$.

Note that we do not have to be concerned about whether there is any label the two taxonomies have in common because, if $s^{(n)}$ and $t^{(m)}$ are semantically similar, $\Pr\left(t^{(m)} \mid s^{(n)}\right)$ is automatically assigned a high value no matter whether they are the same or different in form.

The major difference between *J-MLE* and the cascaded method lies in how semantic concepts are transmitted from the source label vector layer to the target label vector layer. *J-MLE* treats each unique label vector in both the source taxonomy and the target taxonomy as an atomic "label." Therefore, the exchangeability of semantic concepts between two taxonomies is restricted by the *stubborn* treatment. On the other hand, the cascaded method captures the exchangeability in a more flexible way: the semantic concepts are transmitted through the layers of individual labels between the layers of label vectors.

Algorithm 3 Cascaded Estimation with Dawid-Skene model

Input: $\{s_i, t_{ik} : k \in \mathcal{K}_i, i \in I \subset \mathbf{I}\}$.
Output: $\{t_i : i \in \mathbf{I}\}$.
for each $\langle s, t\rangle \in \left\{\left\langle \{0,1\}^{|S|}, \{0,1\}^{|T|}\right\rangle\right\}$ **do**
  for each $\langle n, m\rangle \in \{\langle \{0,1,...,|S|\}, \{0,1,...,|T|\}\rangle\}$ **do**
    initialize $k = 0$, $\Pr\left(t^{(m)} \mid s^{(n)}\right)_0 = \frac{\sum_{i\in I}\sum_{k\in\mathcal{K}_i}[[t_{ik}^{(m)}=t^{(m)}]]\cdot[[s_i^{(n)}=s^{(n)}]]}{\sum_{i\in I}\sum_{k\in\mathcal{K}_i}[[t_{ik}^{(m)}=t^{(m)}]]}$.
    while the converge condition of the Dawid-Skene model is not satisfied **do**
      - compute $\Pr\left(t^{(m)} \mid s^{(n)}\right)_{k+1}$, $\Pr\left(t^{(m)}\right)_{k+1}$ using Dawid-Skene model with $\Pr\left(t^{(m)} \mid s^{(n)}\right)_k$.
      - $k = k + 1$.

    end while
    $\Pr\left(t^{(m)} \mid s^{(n)}\right) = \Pr\left(t^{(m)} \mid s^{(n)}\right)_{k+1}$.
    $\Pr\left(t^{(m)}\right) = \Pr\left(t^{(m)}\right)_{k+1}$.
  end for
  $\Pr(t \mid s) = \prod_{m=1}^{|T|} \frac{\prod_{n=1}^{|S|}\Pr\left(t^{(m)}|s^{(n)}\right)}{\left[\Pr\left(t^{(m)}\right)\right]^{|S|-1}}$.
end for
$t_i = \arg\max \Pr(t_i \mid s_i)$ for each $i \in \mathbf{I}$.

4 Empirical Study

To test the efficiency of the proposed method, we performed an experimental evaluation in the domain of emotion-oriented research, a typical example in multi-label domains, as discussed in Section 1. To collect real-world data, we used two Japanese children's narratives, "Although we are in love"[6] ("Love" for short) and "Little Masa and a red apple"[7] ("Apple" for short), from the Aozora Library[8] as the texts to be annotated. We conducted the experiments using the Lancers crowdsourcing service[9].

We used two typical emotion taxonomies as the source taxonomy and the target taxonomy. One was Ekman's taxonomy (six emotion labels), which is the most commonly used emotion taxonomy in emotion-oriented research. The other was Nakamura's taxonomy (ten emotion labels), which was taken from the "Emotive Expression Dictionary" [Nakamura, 1993] and proven to be appropriate for Japanese language and culture [Ptaszynski *et al.*, 2013]. To enable *mutual validation* to be performed between the two taxonomies (Ekman→Nakamura, Nakamura→Ekman), both were used to annotate the sentences in the two narratives.

An example task input screen is shown in Figure 2 (with Nakamura's taxonomy shown below the sentences). The annotators were native Japanese language speakers. Both Nakamura's taxonomy and the sentences in the two narratives were presented in their original Japanese form. Ekman's taxonomy was presented in its original English form with Japanese explanations.

The crowdsourcing annotators were asked to read the narrative sentences and spontaneously indicate the emotion felt by the character as expressed in each sentence. They were told to check *neutral* if none of the candidate emotions was felt. The two taxonomies were presented separately to arbitrary annotators. The emotions in the two taxonomies are shown in Tables 1 and 2 with their annotation frequencies.

[6]http://www.aozora.gr.jp/cards/001475/files/52111_47798.html
[7]http://www.aozora.gr.jp/cards/001475/files/52113_46622.html
[8]http://www.aozora.gr.jp
[9]http://www.lancers.jp

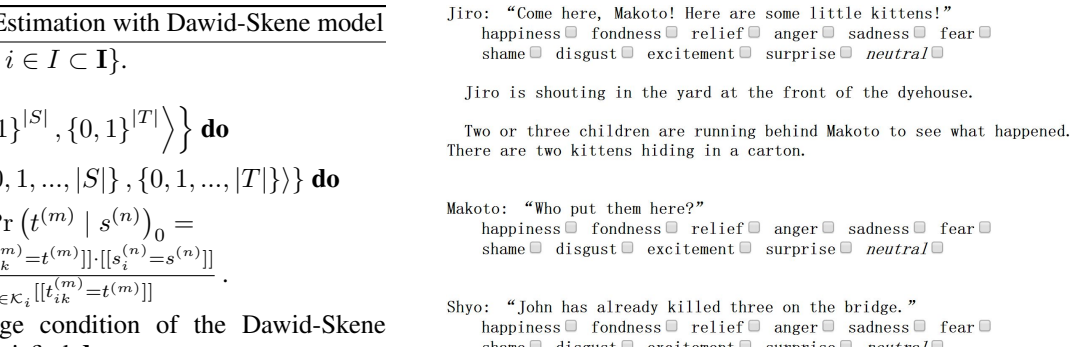

Figure 2: Example task input screen (translated from Japanese).

Table 1: Annotation frequencies of emotions in Ekman's taxonomy and *neutral*, ordered by total frequency

Emotion	"Love"	"Apple"	Total
Sadness	459	500	959
Anger	242	713	955
Neutral	*482*	*450*	*932*
Happiness	519	397	916
Disgust	298	578	876
Surprise	209	529	738
Fear	259	261	520
Total (except *Neutral*)	1986	2978	4964

Table 2: Annotation frequencies of emotions in Nakamura's taxonomy and *neutral*, ordered by total frequency

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

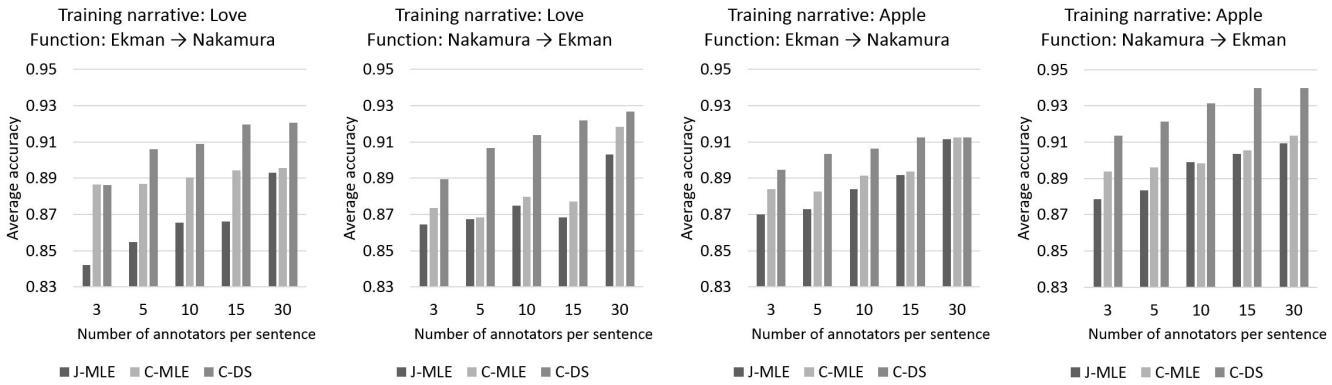

Figure 3: Experimental results.

Table 3: Statistics for experiments

	"Love"	"Apple"	Total
No. of sentences	63	78	141
No. of characters	12	9	21
No. of annotators for Ekman	54	68	93
No. of annotators for Nakamura	30	57	84
No. of annotations for Ekman/Nakamura	1890	2340	4230
Avg. no. of annotations per sentence for Ekman/Nakamura	30	30	30
Avg. no. of assigned emotions per annotation for Ekman	1.05	1.27	1.17
Avg. no. of assigned emotions per annotation for Nakamura	1.75	1.18	1.43

Other statistics for the experiments are shown in Table 3.

For the emotion annotations to be reliable, they should be in accordance with the general consensus of a large crowd. The majority vote strategy most objectively reflects the general consensus if the number of annotators is large enough. Therefore, for each taxonomy, we obtained the gold-standard associated label set for each sentence by having each sentence annotated 30 times using each taxonomy and then taking the majority vote. That is, the most often annotated label set was used as the gold-standard label set.

To determine the effect of the number of annotators on accuracy, we randomly split the 30 annotators who annotated a particular sentence using the target taxonomy 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). Every group of annotators were used to generate the matching function respectively. The accuracy of a certain group size is measured as the average accuracy of the functions generated by all groups in the group size.

The empirical results were actually tested using a kind of cross-validation. In the training step, we used the sentences in one narrative with their aggregated gold-standard source label sets and assigned target label sets in a group to establish the semantic matching function. Then, in the test step, we used the established function and the gold-standard source label set for each sentence to predict the associated target label set for each sentence in both narratives. The semantic matching function was established given the target annotations within each group using the following three models:

- *J-MLE*: **J**oint **M**aximum **L**ikelihood **E**stimation;
- *C-MLE*: **C**ascaded **M**aximum **L**ikelihood **E**stimation;
- *C-DS*: **C**ascaded estimation with **D**awid-**S**kene model.

Since, to the best of our knowledge, there is no work comparable to our proposed method, we used the *J-MLE* model as the baseline for comparison. Because there is no guarantee that the source label sets presented during the test step were also presented during the training step, which would cause the *zero-shot* problem for the *J-MLE* model (as shown in Figure 1(a)), we forced those uncovered label sets to map to *neutral*.

Since it is unreasonable to check whether two binary vectors (label sets) match exactly, e.g., $\{$*anger*, *excitement*$\}$ is closer to $\{$*anger*, *sadness*$\}$ than $\{$*happiness*, *surprise*$\}$, we used the *simple matching coefficient* to evaluate the performance of the function, i.e., the average proportion of state-consistent emotion labels between the predicted target label set and the aggregated gold-standard target label set over all sentences.

Figure 3 shows the experimental results. No matter which of the two narratives was used to establish the semantic matching function, the (*C-MLE* and *C-DS*) models with the cascaded method consistently outperformed the *J-MLE* model although the accuracies of the three models increased with the group size. This means that estimation using the proposed cascaded method is more accurate than ordinary maximum likelihood estimation. Moreover, in most cases, the *C-DS* model achieved better accuracies than the *C-MLE* model and had accuracies greater than 90% for five or more annotators per instance. This demonstrates that annotator bias should be considered in variable-quality crowdsourced anno-

tations and that five is a reasonable number of annotators to achieve satisfactory performance. Finally, the accuracies of the mapping from Nakamura's taxonomy to Ekman's taxonomy were higher than (second histogram) or comparable to (fourth histogram) the accuracies of the reversed matching. This is because there are ten labels in Nakamura's taxonomy and six in Ekman's taxonomy. The mapping from a taxonomy with more labels to one with less labels tends to be more accurate than in the opposite case.

5 Conclusion

We have proposed a novel probabilistic cascaded method for mapping label sets in a source taxonomy to label sets in a target taxonomy in terms of the semantic distances between them in the latent semantic space. This can be useful for addressing the problems that occur when different taxonomies are used in a multi-label domain. Different from multi-label "classification" learning, the proposed method is aimed to enable the associated label set in one (source) taxonomy for an instance to be detected directly from its associated label set in another (target) taxonomy without looking into the content of the instance. Our another objective was to determine how many crowdsourcing annotators have to provide annotations in order for the established matching function to be accurate. The experimental results on real-world crowdsourcing data demonstrated that using the proposed method enables a matching function to be robustly and accurately established from the responses provided by a limited number of annotators.

The results of this research provide several benefits. For example, (1) multi-label applications can use classifiers that have already been vetted, (2) the classifier that comes with a large annotated corpus can be used to train other classifiers or simply supplement the original dataset if allowed by the usage restrictions on the corpus, and (3) the performance of systems using different taxonomies can be uniformly benchmarked. However, notice that the proposed method provides an ideal solution for compromise since instances associated with the same label set in one taxonomy are not necessarily associated with the same label set in another taxonomy.

Our experiments were conducted on a small dataset, two children's narratives, using two emotion taxonomies. We will extend our research across different multi-label domains, such as film genre classification, and explore whether the proposed method is also accurate for larger datasets. We expect that semantic matching will be more difficult if the number of source labels is much smaller than that of target labels because of the difference in the granularity of information. We plan to extend our experiments for such cases to see whether the proposed method is still effective. To simplify the problem of establishing the semantic matching function, we made a preliminary assumption that labels in both the source taxonomy and the target taxonomy are statistically independent. However, this is not the case in reality—some labels may indirectly reveal clues about other labels. We thus plan to design an effective mechanism for automatically incorporating label correlations into the estimation process to further improve the accuracy.

Acknowledgments

This work was supported in part by JSPS KAKENHI 24650061.

References

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

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

[Duan *et al.*, 2014] Lei Duan, Satoshi Oyama, Haruhiko Sato, and Masahito Kurihara. Separate or joint? estimation of multiple labels from crowdsourced annotations. *Expert Systems with Applications*, 41(13):5723–5732, 2014.

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

[Kim *et al.*, 2013] Seungyeon Kim, Fuxin Li, Guy Lebanon, and Irfan Essa. Beyond sentiment: The manifold of human emotions. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 360–369, 2013.

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

[Ptaszynski *et al.*, 2013] Michal Ptaszynski, Hiroaki Dokoshi, Satoshi Oyama, Rafal Rzepka, Masahito Kurihara, Kenji Araki, and Yoshio Momouchi. Affect analysis in context of characters in narratives. *Expert Systems with Applications*, 40(1):168–176, 2013.

[Russell and Fernández-Dols, 1997] James A Russell and José Miguel Fernández-Dols. *The Psychology of Facial Expression*. Cambridge university press, 1997.

[Strapparava and Valitutti, 2004] Carlo Strapparava and Alessandro Valitutti. Wordnet affect: an affective extension of wordnet. In *LREC*, volume 4, pages 1083–1086, 2004.

[Tellegen *et al.*, 1999] Auke Tellegen, David Watson, and Lee Anna Clark. On the dimensional and hierarchical structure of affect. *Psychological Science*, 10(4):297–303, 1999.

[Trohidis *et al.*, 2008] Konstantinos Trohidis, Grigorios Tsoumakas, George Kalliris, and Ioannis P Vlahavas. Multi-label classification of music into emotions. In *International Society for Music Information Retrieval (ISMIR)*, volume 8, pages 325–330, 2008.

[Tsoumakas *et al.*, 2010] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Mining multi-label data. In *Data Mining and Knowledge Discovery Handbook*, pages 667–685. Springer, 2010.

[Widen *et al.*, 2004] Sherri C Widen, James A Russell, and Aimee Brooks. Anger and disgust: Discrete or overlapping categories? In *2004 APS Annual Convention, Boston College, Chicago, IL*, 2004.