A Study on Crowdsourcing for Multi-Label Affect Annotation

Lei DUAN

A Dissertation

Submitted to the Division of Synergetic Information Science

in partial fulfillment of the requirements for the

Doctor of Philosophy

at the

Graduate School of Information Science and Technology

HOKKAIDO UNIVERSITY

January 2016

A Study on Crowdsourcing for Multi-Label Affect Annotation

Lei DUAN

Abstract

The issue of learning from multi-label data is an emerging and promising research topic and has attracted significant attention from a lot of researchers. In this dissertation, we especially focus on a typical referent in multi-label learning: affect. Generally, an enormous amount of multi-label affect annotations is needed to form a multi-label affective learning technique. Moreover, the quality of the annotations directly affects the performance of the technique. Although high-quality affect annotations can be obtained from both experts and large crowds, the process is both expensive and time-consuming in practice. One way to obtain affect annotations is to use online crowdsourcing services, which are being used more frequently in the labeling community. We thus investigated ways to obtain at low cost reliable multi-affect datasets from variable-quality crowdsourced annotations for use with affect-related techniques.

Although multi-label affect annotations can be obtained from a crowdsourcing service at very low cost, in most cases, crowdsourcing annotators are rarely trained and generally do not have the abilities needed to accurately perform the offered tasks. Therefore, ensuring the quality of the results is one of the biggest challenges in crowdsourcing. A promising approach to improving the quality of crowdsourced annotations is to introduce redundancy by aggregating annotations provided by several annotators to produce one reliable annotation. Generally, the more the number of annotators is, the more reliable the aggregated annotation can be. However, hiring more annotators needs higher cost. Moreover, in subjective multi-label annotation tasks, a larger number of annotators are necessary to obtain the reliable annotation than those in objective or single-label annotation tasks. Given that the categories of "affects" are different from those of other kinds of labels, we incorporated the characteristics of affect annotation into estimation process. The purpose of the study is to ensure the quality of the aggregated multi-label affect annotation for each instance, from annotations provided by a limited number of annotators, to reduce the cost of data collection. Experimental results on real-world crowdsourcing data showed that by processing crowdsourced annotations using the proposed methods, the obtained multi-label affect datasets are with quality approaching that of ones obtained from the general consensus of large crowds or from human experts. Our work provided a promising way to reduce the cost creating high-quality multi-affect datasets, with minimal degradation in the quality of the results. We envision that by leveraging proper statistical quality control strategies, a crowdsourcing setting could be a good candidate to the problem of insufficient annotation data in affective learning techniques.

In Chapter 1 the background of this study, including the evolving research area of multi-label affective learning and crowdsourcing, is introduced. Then the challenge of collecting high-quality multi-affect datasets from crowdsourced annotations is discussed. After that, the purpose and the threefold contribution of this study are presented.

Chapter 2 presents the fundamentals necessary for the remaining chapters, including the characterization of affect, techniques in multi-label learning, and challenges of human computation and crowdsourcing. Then it is followed by the description of employed datasets and related work.

In multi-label affective learning, candidate labels are interrelated. In Chapter 3, we investigated estimation of multiple affect labels from crowdsourced annotations, with flexible incorporation of label dependency into the label-generation process.

In view of that emotive expressions are inevitably restricted by "consistency" and "context", in Chapter 4 we propose incorporating information of *emotional consistency* and *contextual cues* among instances into the label-generation process. This is based on the multi-affect estimation proposed in Chapter 3.

Different affect taxonomies are commonly used in the affective learning domain, which results in complications. Given that different affect taxonomies are generally founded on the same latent semantic space, where each possible label set in a taxonomy denotes a single semantic concept, in Chapter 5, we proposed a novel approach for establishing a semantic matching function of label sets in two taxonomies in a crowdsourcing setting.

In Chapter 6, the methods proposed in this dissertation are summarized and some future work are discussed.

Acknowledgments

I offer my sincerest gratitude to my respected academic supervisor, Associate Professor Satoshi Oyama, for all his kind help, inspiring suggestions, patient guidance, and encouraging words that I will never forget. The opportunities that he gave me to review papers for and submit papers to excellent journals and conferences make me more mature as a researcher in the academic activities. I feel so lucky that I could be one of his students for the past three years. I sincerely wish to thank Professor Masahito Kurihara for all his kind directions for my research work and support for attending excellent international academic conferences held in Hong Kong (the International MultiConference of Engineers and Computer Scientists, IMECS 2015), Knoxville, USA (the 15th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL 2015), and Buenos Aires, Argentina (the 24th International Joint Conference on Artificial Intelligence, IJCAI 2015). Such wonderful experiences bring me much broader perspective. I am also grateful to Prof. Haruhiko Sato, a friend-like assistant professor of our laboratory, for all his kind support. In addition, I would like to thank Prof. Tetsuo Ono, Prof. Masahito Yamamoto and Prof. Hidenori Kawamura for their helpful comments and suggestions on the improvements of this dissertation.

I would like to give my special thanks to my fiancee Ms. Le Huang for encouraging me to finish my Ph.D. degree and for understanding and patience while I was frustrated with job hunting. I feel so lucky that I can meet with you at Hokkaido University. I will never forget the night when I went to the dining hall in the building of school of engineering and met "unexpectedly" a pretty girl with her beautiful eyes filling with tears, who would change my whole future life.

I intend to thank Mr. He Chen, Dr. Hao Liu, Mr. and Mrs. Zhang, Ms. Yingbin Hu for all their warm help and support during the period when I first came to Japan. Thank you all my friends: Ms. Jing Song, Mr. Chengcheng Ji, Mr. Jingshan Zhang, Mr. Masahiro Suzuki, Mr. Keiki Zen, and all the members in my laboratory and all the members in the Orchestra of Hokkaido University. Without you, my life in Japan cannot be so happy and colorful.

I also feel deeply grateful to my Alma Mater, Beijing Normal University and Northeastern University, who led me to the world of being a researcher when I was an ignorant stripling. In addition,I would like to take this opportunity to express my sincere appreciation to Hokkaido University for financial support and free Japanese classes. These make possible for me to study in this beautiful country. The days that I shared with her are my precious memories.

Finally, to my caring, loving, and supportive parents:my deepest gratitude. I always owe a big “thank you” to you for your unconditional support.

Contents

Abstract **i**

Acknowledgments **iii**

List of Figures **ix**

List of Tables **xi**

List of Algorithms **xii**

1 Introduction **1**

2 Fundamentals **5**

2.1 Characterization of Affect 5

2.2 Multi-Label Learning 8

2.3 Human Computation, Crowdsourcing and Quality Control 10

2.4 Dataset Descriptions 14

2.4.1 Aozora Library 14

2.4.2 Emotive Expression Dictionary 15

2.5 Related Work 15

2.5.1 Emotion Detection 16

2.5.2 Intention Inference 17

2.5.3 Semantic Matching 19

3 Multi-Affect Estimation Considering Relationships among Labels **21**

3.1 Introduction 21
3.2 Statistical Models 24
3.2.1 Conventional Method: Original Dawid-Skene (*O-DS*) Model . 24
3.2.2 Label-Dependent Dawid-Skene (*D-DS*) model 25
3.2.3 Label Pairwise Dawid-Skene (*P-DS*) model 28
3.3 Inference Algorithm 29
3.4 Discussion: Bayesian Network Label-dependent DS (*D-DS+*) Model . 33
3.5 Empirical Study 35
3.5.1 Emotion Annotation for Narratives 35
3.5.2 Intention Annotation for Tweets 40
3.6 Summary 43

4 Multi-Emotion Estimation Considering Relationships among Instances 47
4.1 Introduction 47
4.2 Statistical Models 48
4.2.1 Multi-Emotion Estimation Considering Emotional Consistency 49
4.2.2 Multi-Emotion Estimation Considering Emotional Consistency with Contextual Cues 50
4.3 Inference Algorithm 52
4.4 Empirical Study 52
4.5 Summary 56

5 Crowdsourced Semantic Matching of Multi-label Emotion Annotations 58
5.1 Introduction 58
5.2 Statistical Models 62
5.2.1 Vector Space Model (VSM)-Based Approach 63
5.2.2 Probability-Based Approach 67
5.3 Empirical Study 71
5.4 Summary 82

6 Conclusion & Future work **84**

6.1 Conclusion 84

6.2 Future work 85

A Proof of optimal label pairing pattern **88**

List of Figures

2-1 Differentiating factors between affect, emotion, and intention. 6

2-2 Three central aspects of a human computation system. 11

3-1 Triangle of Reference. 22

3-2 Multi-label estimation models: (a) *O-DS*, (b) *D-DS*, and (c) *P-DS*. . . 29

3-3 Example Bayesian network for emotion annotation experiment. 34

3-4 Example task input screen (translated from Japanese). Annotators were native Japanese language speakers. Both candidate emotion labels and sentences in the two narratives were presented to annotators in their original Japanese form. 37

3-5 Experimental strategy. 40

3-6 Average emotion annotation accuracies for "Although we are in love" narrative. 41

3-7 Average emotion annotation accuracies for "Little Masa and a red apple" narrative. 42

3-8 Average accuracies for intention annotation. 44

4-1 Graphical model representation for multi-affect estimation: (a) domain-independent model, (proposed in Chapter 3) (b) model considering *emotional consistency*, and (c) model considering *emotional consistency* with *contextual cues*. 49

4-2 Average emotion annotation accuracies for "Although we are in love" narrative. 54

4-3 Average emotion annotation accuracies for "Little Masa and a red apple" narrative. 55

5-1 Taxonomy mismatch between emotion-oriented systems. 60

5-2 Graphical representation of semantic matching: (a) joint maximum likelihood estimation and (b) proposed cascaded method. 68

5-3 Average matching accuracies for function "Ekman → Nakamura", Training narrative: "Although We Are in Love". 78

5-4 Average matching accuracies for function "Nakamura → Ekman", Training narrative: "Although We Are in Love". 79

5-5 Average matching accuracies for function "Ekman → Nakamura", Training narrative: "Little Masa and A Red Apple". 80

5-6 Average matching accuracies for function "Nakamura → Ekman", Training narrative: "Little Masa and A Red Apple". 81

6-1 The relationship between the contributions of this dissertation. 86

List of Tables

2.1 Definitions of subjectivity terminologies provided by Merriam-Webster Online Dictionary. 6

2.2 Some applications of multi-label learning. 9

2.3 Example of a multi-label data set. 10

2.4 Estimation procedures constructed by the *BR* method. 10

2.5 Estimation procedures constructed by the *LP* method. 10

2.6 Distribution of separate expressions across emotions in Nakamura's dictionary, ordered by number of expressions per emotion. 16

2.7 Some emotion detection research. 18

3.1 Annotation frequencies of emotion labels and *neutral*, ordered by total annotation frequencies of per label. 37

3.2 Statistics for emotion annotation experiment. 37

3.3 Average emotion annotation accuracies for "Although we are in love" narrative. 41

3.4 Average emotion annotation accuracies for "Little Masa and a red apple" narrative. 42

3.5 Annotation frequencies of intention labels, ordered by annotation frequencies of per label. 44

3.6 Statistics for intention annotation experiment. 44

4.1 Average emotion annotation accuracies for "Although we are in love" narrative. 54

4.2 Average emotion annotation accuracies for "Little Masa and a red apple" narrative. 55

5.1 Annotation frequencies of emotion labels in Ekman's taxonomy and *neutral*, ordered by total frequency. 75

5.2 Statistics for experiments. 75

5.3 Average matching accuracies for function "Ekman → Nakamura", Training narrative: "Although We Are in Love". 78

5.4 Average matching accuracies for function "Nakamura → Ekman", Training narrative: "Although We Are in Love". 79

5.5 Average matching accuracies for function "Ekman → Nakamura", Training narrative: "Little Masa and A Red Apple". 80

5.6 Average matching accuracies for function "Nakamura → Ekman", Training narrative: "Little Masa and A Red Apple". 81

List of Algorithms

1 Multi-label Estimation from Crowdsourced Annotations. 32

2 Joint Maximum Likelihood Estimation. 72

3 Cascaded Maximum Likelihood Estimation. 72

4 Cascaded Estimation with Dawid-Skene Model. 73

Chapter 1

Introduction

Affect conveyed in digital contents is essential for content such as literature, music, fine art, etc. So affect annotations are important meta-data for affective media (such as narrative sentences) in affect-related techniques and digital libraries. Such annotations are necessary to be used as input for expressive text-to-speech conversion of narratives, reference material for affective education support, training data for machine learning algorithms supporting automatic affect detectors, etc.

Take affective learning for example. Given the complexity of human subjectivity, artificial intelligence techniques aimed at learning human affect experiences, including emotion prediction and intention inference, have one thing in common: They more naturally fit the paradigm of multi-label classification than that of single-label classification since one instance may evoke more than one "affect" at the same time. Generally, an enormous amount of multi-label affect annotations ({<instance, associated affect labels>} pairs) is necessary to form a multi-label affective learning technique, e.g., emotion detection or intention inference. Moreover, the annotation quality directly affects the performance of learning techniques.

Affect annotations are usually provided by human annotators. Annotation acquisition varies according to the number of assigned instances per annotator (single expert vs. multiple crowdsourcing annotators). One naïve way is that one expert annotates all instances. The expert annotations are expected to be accurate. However, this approach leads to very expensive cost and comparatively long time. On-line

crowdsourcing services provide a means for outsourcing various kinds of tasks to hundreds of thousands of people, and labeling is one of the main crowdsourcing tasks. The state-of-the-art is for each instance to be annotated by one crowdsourcing annotator. Although this can get access to multi-label data collection 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 tasks. Moreover, some annotators may simply submit random responses as a means to earn easy money. Therefore, crowdsourced annotations may be not as reliable as expert does, due to the lack of knowledge and experience. From this viewpoint, ensuring the quality of the results submitted by annotators is one of the biggest challenges in crowdsourcing.

Affect annotation tasks is more subjective than most other labeling tasks. There are different tendencies and substantial variations among individuals when detecting affects. Although there is no denying that minority opinions are also important to depict the distribution in affect comprehension, to provide a compromise between individuals, it is still necessary to find one interpretation (i.e., high-quality affect annotation) agreed by the majority. This means that the high-quality affect annotation should be in accordance with the general consensus of large crowds. In other words, the annotation quality greatly depends on the judgment of an annotator, and collecting annotations from only one annotator per instance is actually problematic in most cases. In view of this consideration, obtaining high-quality affect annotations is a challenging problem due to the subjectivity of affect.

A promising approach to improving the quality of crowdsourced annotations is to introduce redundancy, which involves asking several annotators (a sub-set of the crowd) to work on each task, and then aggregating their responses (crowdsourced annotations) to obtain a more reliable annotation (crowd's opinion). This is also called "approximating the crowd" [1]. 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 annota-

tors equally produces poor quality results. However, collecting responses from a large number of annotators is almost impossible due to the extremely high cost (time and expense). To alleviate this problem, a number of sophisticated statistical techniques (e.g., [2, 3, 4, 5]) have been proposed for obtaining a reliable annotation from annotations provided by a limited number of crowdsourcing annotators[1]. However, most techniques simply handle the problem of estimating a single associated label for each single-labeled instance.

In multi-label affective learning, if there are n candidate labels, the number of possible label sets is 2^n. This means that in multi-label affect annotation tasks, a larger number of annotators are necessary to obtain the reliable annotation than those in single-label annotation tasks. To the best of our knowledge, the problem of multi-affect estimation has not been effectively solved. In this dissertation, we propose multi-affect estimation methods by incorporating affect-specific information. The aim is to determine the best way to estimate associated labels for each instance from annotations provided by a handful of crowdsourcing annotators. The final purpose is to reduce the cost of creating high-quality annotations for further multi-label affective learning techniques with minimal degradation in the quality of the results.

There are two kinds of raw datasets. One is instances without any associated labels, the other is instances with associated labels selected from an undesired taxonomy.

1. For the first kind of datasets, we extended the Dawid-Skene model (Section 3.2.1), which is originally proposed for single-label estimation, to handle the problem of multi-affect estimation. The proposed estimation approaches (Sections 3.2.2 and 3.2.3) not only consider annotator bias, but also take relationships among labels into account, to simultaneously estimate the multiple associated labels for each instance given crowdsourced multi-label annotations. In the proposed approaches, label dependency is flexibly incorporated into the label-generation process.

[1]For a detailed discussion, see Section 2.3

2. Especially, for the instances that implicitly contain information for "consistency" and "context", such as narrative sentences, we proposed incorporating relationships (*emotional consistency*) across instances into the aggregation process (Chapter 4). This enable the cost of preparing high-quality multi-label affect annotations to be further reduced.

3. For the second kind of datasets, we proposed a vector space-based approach (Section 5.2.1) and a probability-based approach (Section 5.2.2) for establishing the semantic matching function from label sets in the undesired (source) affect taxonomy to label sets in the desired (target) affect taxonomy, where each possible label set in any of the two taxonomies is viewed as denoting a single semantic concept.

The remainder of this article is organized as follows. First, we summarize the research and literature which is inspired or related to the work described later in this dissertation in Chapter 2. Then we introduce the approach aimed at handling the problem of multi-affect estimation in a crowdsourcing setting in Chapter 3, and discuss how we extended it to handle the problem implicitly containing information for "consistency" and "context" in Chapter 4. After that, we present a novel approach to bridge the gap between different affect taxonomies in multi-label affective learning in Chapter 5. Finally, we give the conclusion and discuss about our future work in Chapter 6.

Chapter 2

Fundamentals

2.1 Characterization of Affect

Affect, feeling, emotion, sentiment, mood, intention, ..., such subjectivity terminologies are often interchangeably used in affect-related studies. To analyze the meanings of terminologies, standard dictionaries are an initial resource. Table 2.1 provides the definitions of affect, emotion, and intention, the focuses of this dissertation, based on the Merriam-Webster Online Dictionary[1].

From the definitions given in Table 2.1, among these three subjectivity terminologies, affect seems to be the most abstract and difficult to be fully realized in language since it is always prior to or outside of consciousness [6]. Affect is the "body's way of preparing itself for action in a given circumstance by adding a quantitative dimension of intensity to the quality of an experience" [7]. Affective states may be real or faked, as well as closely connected to interpersonal interaction, context, rituals or social stratifications, patterns, and expectations [8]. Figure 2-1 serves to illustrate the factors of the subjective terminologies, differentiating according to the perspective of "conscious". Affect within the psychological literature has been thought of as the umbrella terminology. It is usually used for fuzzy affective notions such as emotion, feeling, sentiment, mood, intention [9]. A detailed discussion of the concept of affect is beyond the scope of this dissertation but can be found in Munezero *et al.* [10].

[1] http://www.merriam-webster.com

Table 2.1: Definitions of subjectivity terminologies provided by Merriam-Webster Online Dictionary.

Subjectivity Terminologies	**Definition**	**Reference Place**
Affect	1. The conscious subjective aspect of an emotion considered apart from bodily changes; 2. A set of observable manifestations of a subjectively experienced emotion.	One of the topics through this dissertation.
Emotion	A conscious mental reaction (as anger or fear) subjectively experienced as strong feeling usually directed toward a specific object and typically accompanied by physiological and behavioral changes in the body	Related work summarized in Section 2.5.1; Fundamental of experiments discussed in Sections 3.5.1, 4.4, and 5.3.
Intention	2. A concept considered as the product of attention directed to an object of knowledge.	Related work summarized in Section 2.5.2; Fundamental of experiment discussed in Section 3.5.2.

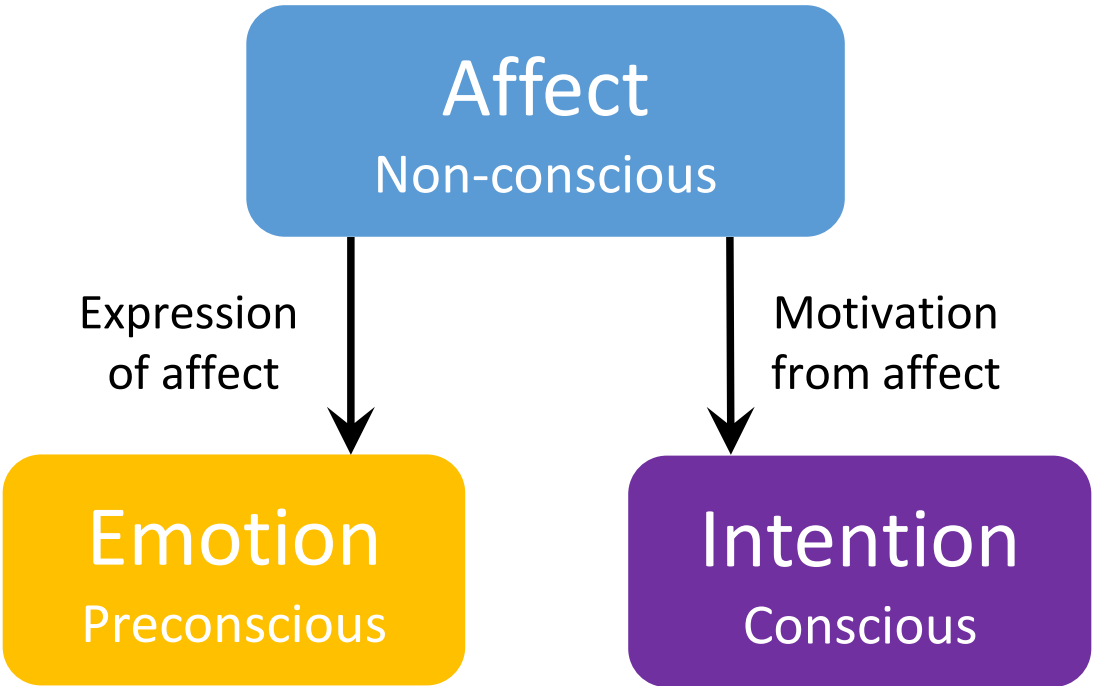


Figure 2-1: Differentiating factors between affect, emotion, and intention.

Affect has been given a great deal of attention in psychology, and turning to a fast-growing new area of computer science in resent years. Studies of affect annotation, recognition, generation, or labeling by humans or computational agents is generally based on one or more of the presented perspectives and ideas. The paradigms used for affect annotation depend on affect applications. For simple applications, it is sufficient to annotate whether an instance (such as a narrative sentence, a movie clip, or a music piece) is affective or the instance's affect valence (positive or negative). Some researchers (e.g., [11, 12, 13, 14]) have considered affect state as a single category, with only one particular affect category (e.g., *happiness* or *sadness*) appearing at a time. However, such annotations obviously simplify the complexity of human affect and are thus not effective for more complicated applications such as expressive text-to-speech synthesis [15, 16] and therapeutic education of children with communication disorders [17]. Therefore, the naïve assumption has been undermined by the results of psychology studies.

It has been demonstrated that a single affect category is unable to represent all possible emotional manifestations [18] and that some affect manifestations are a combination of several emotion categories [19]. For example, Alm [20] observed that the following sentence from H. C. Andersen's fairy tale "The Ugly Duckling[2]" 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 affect category would fail to represent this multiplicity. Therefore, this dissertation focus on the multiplicity of affect, where each instance can be associated with multiple affect categories.

[2]http://www.surlalunefairytales.com/uglyduckling/index.html

2.2 Multi-Label Learning

Traditional multi-class classification is aimed at categorizing instances into a set of candidate labels, in which each instance is associated with a single label j from a set of disjoint labels J. Multi-label classification is a generalization of multi-class classification. In multi-label learning, each instance can be simultaneously associated with multiple labels $\mathcal{J} \subseteq J$.

Learning from multi-label data is an emerging and promising research topic and has attracted significant attention. In addition to affective learning introduced in Section 2.1, many techniques are developed on the basis of multi-label learning. This is mainly motivated by an increasing number of new applications, such as topic categorization of news articles [21] and web pages [22], email analysis [23], semantic categorization of images [24] and videos [25], and emotion prediction in music [26] and narratives [27]. Table 2.2 presents a variety of typical multi-label learning applications. A good survey on multi-label learning was presented by Tsoumakas *et al.* [28].

Two widely used methods for multi-label classification are the binary relevance (BR) method and the label combination or label power-set (LP) method [28]. The methods will be exemplified through the multi-label data set showed by Table 2.3.

1. The BR method decomposes the multi-label estimation problem into several independent binary-label estimation 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. Table 2.4 shows the three estimation procedures (one for each column) constructed by the BR method when applied to the data set of Table 2.3. Obviously, the BR method does not consider dependency among candidate labels. This is reasonable only in the extreme case that labels are mutually independent.

2. The LP method 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. Table

Table 2.2: Some applications of multi-label learning.

Data type	Application	Resource	Label Description	Label Examples	References
text	categorization	news article	Reuters topics	agriculture, fishing	[21]
		web page	Yahoo! directory	health, science	[22]
		patent	WIPO	paper-making, fibreboard	[29, 30]
		email	R&D activities	delegation	[23]
		legal document	Eurovoc	software, copyright	[31]
		medical report	MeSH	disorders, therapies	[32]
		radiology report	ICD-9-CM	diseases, injuries	[33]
		research article	Heart conditions	myocarditis	[34]
		research article	ACM classification	algorithms	[35]
		bookmark	Bibsonomy tags	sports, science	[36]
		reference	Bibsonomy tags	ai, kdd	[36]
		adjectives	semantics	object-related	[37]
	emotion prediction	narrative sentence	Nakamura's emotion taxonomy	fondness, shame	[27]
image	semantic annotation	pictures	concepts	trees, sunset	[38, 39, 24]
video	semantic annotation	news clip	concepts	crowd, desert	[25]
audio	noise detection	sound clip	type	speech, noise	[40]
	emotion detection	music clip	emotions	relaxing-calm	[41, 26]
structured	functional genomics	gene	functions	energy, metabolism	[42, 43, 44]
	proteomics	protein	enzyme classes	ligases	[30]
	directed marketing	person	product categories		[45]

Table 2.3: Example of a multi-label data set.

Instance 1	{a}
Instance 2	{b,c}
Instance 3	{a,c}

Table 2.4: Estimation procedures constructed by the *BR* method.

	a	**b**	**c**
Instance 1	**Y**	N	N
Instance 2	N	**Y**	**Y**
Instance 3	**Y**	**Y**	N

Table 2.5: Estimation procedures constructed by the *LP* method.

	{}	**{a}**	**{b}**	**{c}**	**{a,b}**	**{a,c}**	**{b,c}**	**{a,b,c}**
Instance 1	N	**Y**	N	N	N	N	N	N
Instance 2	N	N	N	N	N	N	**Y**	N
Instance 3	N	N	N	N	N	**Y**	N	N

2.5 shows the three estimation procedures (one for each column) constructed by the *LP* method when applied to the data set of Table 2.3. Although the *LP* method takes label dependency into account, a large number of classes has to be dealt with when the number of candidate labels is large. Therefore, the *LP* method can easily suffer from the sparsity of high-dimensional annotations.

2.3 Human Computation, Crowdsourcing and Quality Control

As a relatively young field, the idea of *human computation* and *crowdsourcing* is not yet well defined. This study mostly uses these two terminologies interchangeably. For the term *crowdsourcing*, Howe [46] 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

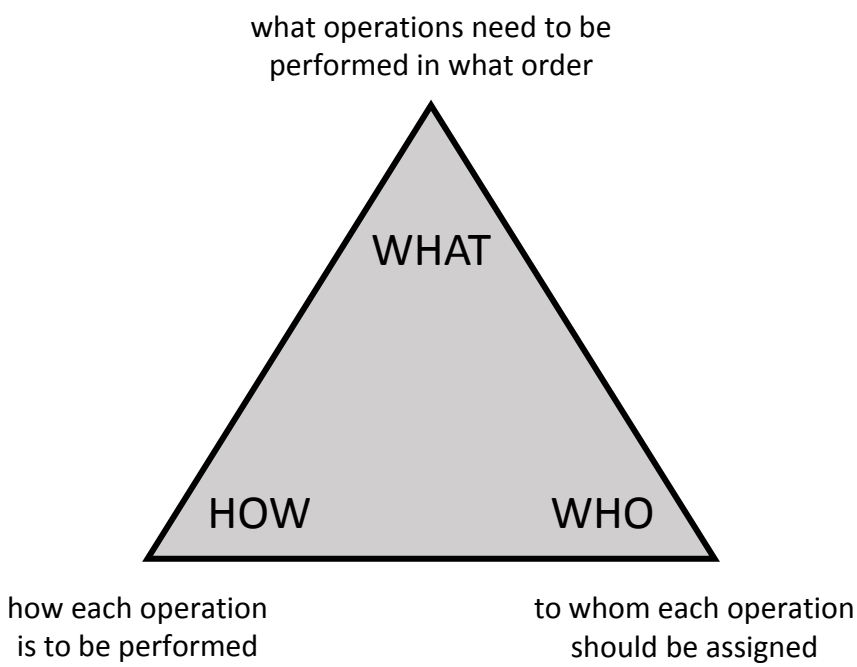


Figure 2-2: Three central aspects of a human computation system.

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 that enables individuals or businesses to request and manage works that are difficult to accomplish by computers but relatively easy for humans.

Crowdsourcing is refered as the act of outsourcing tasks, traditionally performed by employees or contractors, to an undefined, large group of people or community (crowd) through an open platform. Figure 2-2 shows the three central aspects of a human computation system [47], which can also be seen as follows:

- Who (Owner/Sponsor + Sector)
- What (Task being Outsourced)
- How (Crowdsourcing Process - Passive, Collaborative, Competitive)

On the basis of the definition, crowdsourcing can be considered as a novel employment between employees and employers, with payment or not. In contrast to

crowdsourcing tasks with payment, Wikipedia[3] is a typical free-payment crowdsourcing application. There has been over 30 million articles written collaboratively by communities and individuals in this its free, web-based encyclopaedia since 2001.

With the recent expansion of crowdsourcing platforms such as Amazon Mechanical Turk[4](MTurk), CrowdFlower[5], and Lancers[6], the concept of crowdsourcing has been successfully leveraged to solve problems that are hard to solve in old days, especially in various areas of computer science research, including natural language processing [48] and computer vision [49].

Treating the annotations collected by crowdsourcing as training data is the most common way in Machine Learning area. Although annotations can be obtained from a crowdsourcing service at very low cost (time and expense), as defined by Howe [46], there is no guarantee that all crowdsourcingannotators are sufficiently competent to complete the offered tasksaccurately. 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 of the results 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.* [2] presented a method for inferring the unknown health

[3]https://www.wikipedia.org/
[4]http://www.mturk.com
[5]http://crowdflower.com
[6]http://www.lancers.jp

state of a patient given diagnostic tests by several clinicians, where the biases of the annotators (clinicians) were modeled by a confusion matrix. This model is the basis of the idea described in Chapter 3 and is used as a learning tool in the method described in Section 5.2.2. Whitehill *et al.* [3] presented a model for simultaneously estimating the associated label for each repeatedly labeled instance, the expertise of each annotator, and the difficulty of each question. Welinder *et al.* [4] incorporated into their bird image classification model all the above factors, along with a normalized weight vector for each annotator, where each weight indicates relevance to the annotator. Snow *et al.* [48] 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.* [50] 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.* [1] 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.* [5] investigated the use of not only crowdsourced annotations, but also annotators' self-reported confidence scoresto estimate the associated label for each single-labeled instance. Although the accuracy rate could be improved by combining confidence scores with annotations, assigning confidence scores for each response would still be an additional burden for annotators. Jung *et al.* [51] propose using Probabilistic Matrix Factorization to address the missing data problem that an annotator is generally not required to annotate all the instances. Baba *et al.* [52] applied quality control techniques to the detection of crowdsourcing tasks considered to be improper by a crowdsourcing service. They found that the accuracy of detecting improper tasks could be improved by combining non-expert judgments provided by crowdsourcing annotators with expert judgments. In addition to explore statistical models, Mao *et al.* [53] investigated how different voting rules perform in human computation problems. More detailed discussions can be found in benchmark studies [54].

The aforementioned works focused on single-label estimation. There has also been some research on the problem of multi-label estimation, the focus of this dissertation.

Nowak *et al.* [55] 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. Bragg *et al.* [56] presented a decision-theoretic approach to taxonomy creation that implements the *BR* method. They reported that, with their approach, 16 annotators per instance are sufficient to achieve quality comparable to the general consensus of large crowds.

Some research in the machine learning community addressed the problem of supervised learning directly from crowdsourced annotations, such as the semantic-matching problem discussed in Chapter 5. Sheng *et al.* [57] explored several methods for choosing which instances should get more labels and how to include label uncertainty information when training classifiers. Donmez *et al.* [58] proposed simultaneously estimating annotator accuracies and training a classifier using annotator responses to actively select the next instance for annotating. Raykar *et al.* [59] extended the work of Dawid and Skene [2] by introducing a logistic classification model to incorporate domain-specific information (the features of a medical image in the area of computer-aided diagnosis) about the task (predicting whether a suspicious region on the medical image is malignant or benign).

2.4 Dataset Descriptions

2.4.1 Aozora Library

The Aozora (Blue Sky) Library[7] is a Japanese digital repository created on the Internet in 1997, containing freely available books (online library). It contains over 10,000 books of various genres (fiction, philosophy, history, art, etc.) published in Japanese for which copyrights have expired (50 years after the death of the copyright holder). In Japan, Aozora Library is sometimes compared to Project Gutenberg.

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

One of the genres in Aozora Library is the children's book. This genre includes sub-categories such as *history books*, *beautiful arts and crafts books*, and *literature*. Children's narratives and fairy tales are included in the *literature* category, which contained 1217 books in December 2014. From this category we chose two narratives at random for experiments in Sections 3.5.1, 4.4, and 5.3.

2.4.2 Emotive Expression Dictionary

The Emotive Expression Dictionary [60] is a dictionary developed by Akira Nakamura over a period exceeding 20 years. It is a collection of over 2000 expressions describing emotional states collected manually from a wide range of literature. It is not a tool for emotion prediction *per se* but was converted into an emotive expression database by Ptaszynski *et al.* [27] in their research on emotion prediction of utterances in Japanese. Since the chosen narratives in this dissertation were in Japanese, it is necessary to use a candidate emotion category set proven to be appropriate for the Japanese language. Nakamura's dictionary is a state-of-the-art example of a hand-crafted lexicon of emotive expressions. In particular, it uses ten emotion categories proven to be appropriate for the Japanese language and culture [27]. 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 2.6. These ten emotion categories are employed as the candidate emotion labels for experiments in Section 3.5.1, 4.4, and 5.3.

2.5 Related Work

Emotion detection (Section 2.5.1) is the fundamental of experiments discussed in Sections 3.5.1, 4.4, and 5.3. Intention inference (Section 2.5.2) is the fundamental of experiment discussed in Section 2.5.2. Semantic matching (Section 2.5.3) is the fundamental of Chapter 5.

Table 2.6: Distribution of separate expressions across emotions in Nakamura's dictionary, ordered by number of expressions per emotion.

Emotion (in Japanese)	English translation	No. of expressions
en/iya	**disgust**, dislike, detestation	532
kō/takaburi	**excitement**, excitation	269
ai/aware	**sadness**, sorrow, gloom	232
ki/yorokobi	**happiness**, joy, delight	224
dō/ikari	**anger**, wrath	199
kō/suki	**fondness**, liking	197
fu/kowagari	**fear**, scare, terror	147
kyō/odoroki	**surprise**, amazement	129
an/yasuragi	**relief**, slack	106
chi/haji	**shame**, shyness, bashfulness	65
Total		2100

2.5.1 Emotion Detection

Humans, by nature, can be emotionally affected by literature, music, fine art, etc, so detecting "emotion" is an important access point in digital libraries and online repositories. Analyzing how we are affected is a vital research direction in artificial intelligence and digital media processing as it is potentially applicable to many further emotion-related applications. Many researchers have thus concentrated on this area. For example, Alm *et al.* [11] 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. Neviarouskaya *et al.* [61] tested their emotion analysis model for English on various kinds of data, including fairy tales (the same narratives as Alm *et al.* [11]). Mohammad [62] proposed a visualization method for particular emotion related words in stories based on the idea of emotion word density. However, they did not aim at precise affect detection, but rather visualization of presence of emotion words within certain childrens stories. Kim *et al.* [12] modeled emotion as a continuous *manifold* in sentiment detection research and constructed a statistical model connecting it to documents and to a discrete set of emotions. Li *et al.* [63] proposed a method for identifying emotions in micro-blog posts by using "emotion cause extraction". A

number of machine learning algorithms have been proposed for classifying music by mood in the music digital library domain [64, 14]. Similar to the model introduced in Section 5.2.1, Danisman *et al.* [65] 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. The emotion detection research described above is summarized in Table 2.7.

There have also been several attempts to leverage crowdsourcingin the emotion detection domain. Alm [20] 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.* [13] 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.

As described in Section 2.1, due to the multifaceted nature of affect, an instance is more naturally to be associated with a combination of multiple affect categories. Much of the recent emotion detection research has concentrated on exploiting the complexity of emotion experiences. Trohidis *et al.* [26] modeled emotion detection in music pieces as a multi-label classification task. Liu *et al.* [66] proposed an implicit video multi-emotion tagging method. Ptaszynski *et al.* [27] did an experiment on multi-emotion analysis of certain characters in narratives. A complete discussion of emotion detection research is beyond the scope of this dissertation but can be found in Calvo *et al.* [67] and Pelachaud *et al.* [68].

2.5.2 Intention Inference

In today's Web 2.0 era, people post descriptions of their various real-world experiences such as visiting places, participating in activities, and shopping to social networking services, such as Twitter[8] and Facebook[9]. Extracting such information from the huge

[8]https://twitter.com
[9]https://www.facebook.com

Table 2.7: Some emotion detection research.

Domain	**Approach**	**Language**	**Level**	**Dataset**	**Techniques**	**Emotion categories**	**References**
Fairy tales	Fusion-based	English	Sentence	Children narratives (185 tales)	Winnow	Anger/Disgust/Fear /Happiness/Sadness /Positively Surprised /Negatively Surprised /*Neutral*, Valence	[11, 20]
Fairy tales, News, Blogs	Corpus-based	English	Sentence, Document	Children stories (185 tales), News headlines (1890 sentences), Blog entries (700 sentences)	Affect analysis model	Anger/Fear /Happiness/Sadness /Surprize/*Neutral*, Valence	[61]
Fairy tales	Corpus-based	English	Word, Document	The Brothers Grimm fairy tales (192 stories)	Keyword matching	Angry/Disgust/Fear /Happiness/Sadness /Surprise/Trust /Anticipation, Valence	[62]
Blogs	Corpus-based	English	Sentence	Blog entries (1,346,937 sentences)	Manifold	32 emotions	[12]
Micro-Blogs	Corpus-based	Chinese	Sentence	Micro-Blog entries (16485 sentences)	SVM	Happiness/Anger /Disgust/Fear /Sadness/Surprise /*Neutral*	[63]
Music	Corpus-based	English	Song	songs in AllMusic.com (13,948 songs))	Naïve Bayes	178 emotions	[64]
Music	Corpus-base	English	Song	MIREX 2007 & 2008 (296 songs)	SVM	18 emotions	[14]
Text	Corpus-base	English	Sentence	ISEAR, Wordnet-Affect, WPARD, SemEval	SVM, Naïve Bayes, Vector Space	Anger/Disgust/Fear /Joy/Sad/Shame /Guilt	[65]

amounts of real-time updated text corpora is important for estimating the popularity of places, activities, and products, and is of great value to navigation and recommendation systems. Lee *et al.* [69] developed a method for detecting geo-social events, such as local festivals, by monitoring crowd behaviors indirectly via Twitter. Liao *et al.* [70] investigated whether and how micro-messaging technologies could be used to predict attendance trends at the World Expo 2010 in Shanghai. Xu *et al.* [71] proposed using a *semidefinite programming* optimization technique for identifying valuable customers from social network services in terms of profit maximization.

2.5.3 Semantic Matching

Interoperability among people of different cultures and languages, having different viewpoints and using different terminology, has always been a huge problem. This is also 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 lightweight ontologies, like classifications, XML schemas, and label sets in taxonomies (the scenario in this dissertation), to identify those objects in two structures that semantically correspond to one another [72]. This technique is often used to identify information which is semantically related. For example, applied to file systems it can identify that a folder labeled car is semantically equivalent to another folder automobile because they are synonyms in English. This information can be taken from a linguistic resource like WordNet [73].

Semantic matching represents a fundamental technique in many applications in areas such as query translation, data migration, data integration, and resource discovery. It has been proposed as a valid solution to the semantic heterogeneity problem, such as managing the diversity in knowledge. This problem seems to be emphasized especially with the advent of the consequential information and the Web explosion. People face the concrete problems to integrate, disambiguate, and retrieve information coming from a wide variety of sources.

In the recent years many semantic matching operators has been offered. S-Match [74] is an example of a semantic matching operator. It works on lightweight ontolo-

gies [75], namely graph structures where each node is labeled by a natural language sentence, for example in English. Semantic matching is also being investigated in other areas such as event processing [76]. A good survey is presented by Shvaiko and Jérôme [77].

Chapter 3

Multi-Affect Estimation Considering Relationships among Labels

3.1 Introduction

In this and the next chapters, we leverage the aggregation strategy to estimate the associated labels for each instance given noisy crowdsourced annotations. The proposed models enable the cost of preparing high-quality multi-label affect annotations for use with multi-label learning techniques to be reduced, with minimal degradation in the quality of the results.

On-line crowdsourcing services provide a means for outsourcing labeling tasks to a large group of people. 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 [78] shown in Figure 3-1. The interpreter perceives the sign (e.g., a word, a sound, an image, a sentence, etc) and through some cognitive process attempts to find the referent (e.g., an object, an idea, a class of things, etc) of that sign. As discussed in section 2.3, most state-of-the-art quality control techniques include latent factors related to these three components,

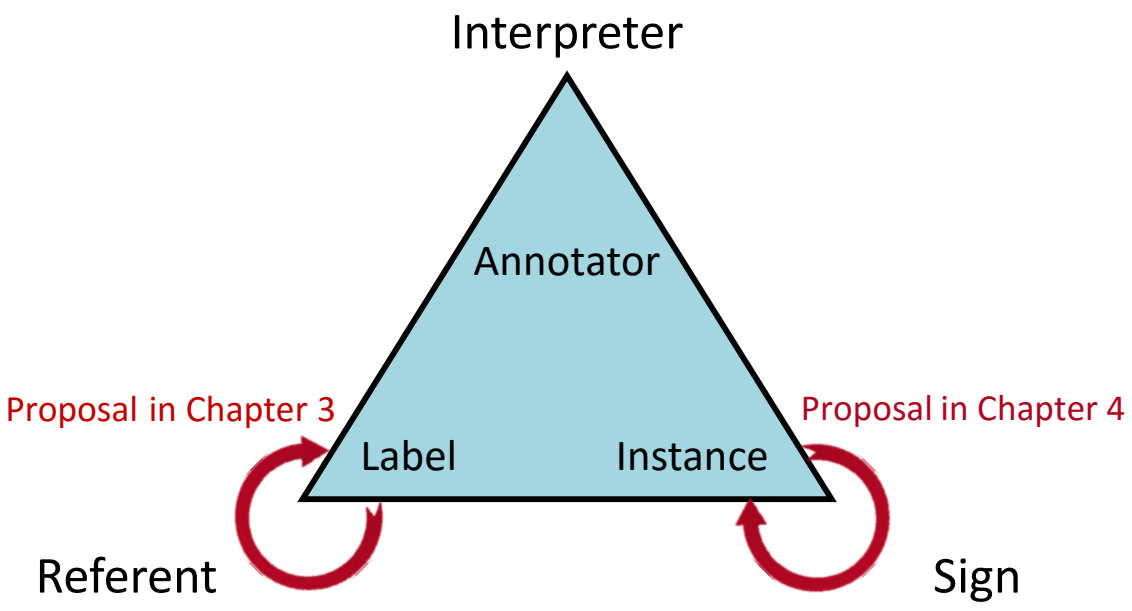


Figure 3-1: Triangle of Reference.

such as annotator bias [2], annotator expertise [3], and instance difficulty [4, 3]. However, they ignore the internal relationships among labels and among instances.[1] In other words, for each repeatedly annotated instance, the reliable annotation is produced separately.

The categories of "affect" have characteristics different from those of other kinds of labels. For example, some affect labels may reveal clues about other affect labels. For instance, in the emotion annotation experiment described in Section 3.5.1, a sentence expressing *fear* may also express a certain degree of *anger* and/or *surprise*, *fondness* and *anger* are rarely co-true, and *shame* or *anger* may be false when *relief* is true.

Annotation acquisition varies according to the number of labels per example (single vs. multiple). Aiming to address limitations in multi-label affective learning techniques, in this chapter, we propose flexibly incorporating label dependency into the label-generation process. In particular, we propose three statistical quality control models based on the model of Dawid and Skene [2] (DS), a well-known unsupervised single-label classification algorithm:

[1] 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.

- **Label-dependent *DS* (*D-DS*) model** (Section 3.2.2)

  The *D-DS* model, which is an implementation of the *LP* method, simply incorporates dependency relationships among all candidate labels into account.

- **Label pairwise *DS* (*P-DS*) model** (Section 3.2.3)

  The *P-DS* model groups candidate labels into pairs, and then separately estimates the states of the two labels within each pair, thereby preventing interference from uncorrelated labels.

- **Bayesian network label-dependent *DS* (*D-DS+*) model** (Section 3.4)

  The *D-DS+* model depicts the conditional independence properties of the joint distribution over candidate labels as a Bayesian network and approximates the underlying high-dimensional joint distribution by using the product of the conditional distributions of the candidate labels.

Traditional affect detection research is aimed at detecting single or multiple affect label(s) from an instance using a trained detector. The work in this chapter aimed at estimating multiple affect labels for each instance directly from crowdsourced multi-label annotations. This can be seen as an unsupervised multi-label classification problem. The expectation maximization (EM) algorithm [79] is widely used to solve unsupervised classification problem. Therefore, we exploited an EM-based incremental algorithm (described in Section 3.3) to estimate the gold standard multi-label annotations together with the parameters of the proposed models.

To evaluate the effectiveness of the proposed models, we conducted two experiments using Lancers crowdsourcing service. Affect might occur at different levels, such as word, sentence, paragraph and chapter in scenario of narrative. In our experiments, we only focused on the sentence level. In one experiment (in Section 3.5.1), crowdsourcing annotators were tasked with annotating the emotion labels of sentences in narratives, and in the other experiment (in Section 3.5.2) they were tasked with annotating the intentions of tweet posters. The results of two experiments show that, with annotations provided by a handful of crowdsourcing annotators,

1. in most cases, the *D-DS+* model most effectively handles the multi-affect estimation problem with annotations provided by only about five annotators per instance;

2. the *P-DS* model is best if there are pairwise comparison relationships among candidate labels.

3.2 Statistical Models

It is obvious that affect labels are interrelated. To take into account the dependency relationship among affect labels, we first introduce the concept of *conjoint-affect*. A conjoint-affect represents a subset of the set of candidate affect labels. For example, in the scenario of emotion annotation, the two conjoint-affects {*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 instances, J be the set of candidate affect labels. K be the set of crowdsourcing annotators, and $\mathcal{K}_i \subseteq K$ $(i \in \mathcal{I})$ be the set of annotators who annotated instance i. (Note that it is not necessary to ask every annotator to annotate all the instances.) The number of times that annotator k annotated instance i with conjoint-affect $\mathcal{L}$ is given by $n_{i\mathcal{L}}^{(k)} \in \mathbb{N}$ $(i \in I, k \in \mathcal{K}_i, \mathcal{L} \subseteq J)$, which can be directly calculated from the crowdsourced annotations. The true conjoint-affect for instance i is denoted by $\mathcal{T}_i \subseteq J$ $(i \in I)$. The objective is to aggregate the set of annotations $\left\{n_{i\mathcal{L}}^{(k)} : i \in I, k \in \mathcal{K}_i, \mathcal{L} \subseteq J\right\}$ to estimate the set of true conjoint-affects $\{\mathcal{T}_i : i \in I\}$.

3.2.1 Conventional Method: Original Dawid-Skene (*O-DS*) Model

Our work is based on the well-known Dawid-Skene model [2], which is aimed at inferring the unknown health state of a patient given the assessments of several clinicians.

In the *O-DS* model, patients and clinicians are the counterparts of instances I and annotators K. Let J be the possible health states. $t_i \in J\,(i \in I)$ is the true state of patient i, which is determined by the maximum a posteriori (MAP) principal:

$$t_i = \arg\max_{\jmath \in J} \Pr\left[t_i = \jmath \mid \left\{n_{il}^{(k)} : k \in \mathcal{K}_i, l \in J\right\}\right], \tag{3.1}$$

where $n_{il}^{(k)} \in \mathbb{N}\,(k \in K, i \in I, l \in J)$ denotes the number of times that clinician k declared patient i to be in state l.

In our research, the state (*true* or *false*) of a particular label for an instance can be considered as the health state of a patient. On the basis of this, the *O-DS* model can be directly used to estimate the state of a particular label for each instance. Let $t_i^{(j)} \in \{\text{true}, \text{false}\}\,(i \in I, j \in J)$ denote whether label j is true or false for instance i, and let $n_{i\imath}^{(jk)} \in \mathbb{N}\,(j \in J, k \in K, i \in I, \imath \in \{\text{true}, \text{false}\})$ be the number of times that annotator k annotated instance i with ($\imath$ = true) or without ($\imath$ = false) the label. Similar to equation (3.1), whether label j is true for instance i can be estimated using

$$t_i^{(j)} = \arg\max_{\jmath \in \{-1,1\}} \Pr\left[t_i^{(j)} = \jmath \mid \left\{n_{i\imath}^{(k)} : k \in \mathcal{K}_i, \imath \in \{\text{true}, \text{false}\}\right\}\right]. \tag{3.2}$$

Then, the true conjoint-affect for instance i is determined as

$$j \begin{cases} \in \mathcal{T}_i, & \text{if:} \quad t_i^{(j)} = \text{true}\,, \\ \notin \mathcal{T}_i, & \text{if:} \quad t_i^{(j)} = \text{false}\,. \end{cases}$$

Simply put, the *O-DS* model is an implementation of the *BR* method.

3.2.2 Label-Dependent Dawid-Skene (*D-DS*) model

As described in Section 3.2.1, the states of different labels for each instance must be estimated separately using different *O-DS* models. This is suitable for multi-label estimation only in the extreme case that labels are mutually independent. We have discussed in Section 3.1 that affect labels are interrelated To take advantage of this insight, we extended the *O-DS* model so that it takes label dependency into account to

simultaneously estimate multiple associated labels for each instance given multi-label annotations.

As a first step, we focus on the assumption that affect labels are completely dependent on each other. Similar to that of the *O-DS* model, the true conjoint-affect for instance i is the one that achieves the maximum expectation:

$$\mathcal{T}_i = \arg\max_{\mathcal{J} \subseteq J} \mathrm{E}\left[\mathcal{T}_i = \mathcal{J}\right], \tag{3.3}$$

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

$$\mathrm{E}\left[\mathcal{T}_i = \mathcal{J}\right] = \Pr\left[\mathcal{T}_i = \mathcal{J} \mid \left\{n_{i\mathcal{L}}^{(k)} : k \in \mathcal{K}_i, \mathcal{L} \subseteq J\right\}\right]. \tag{3.4}$$

Therefore, the *D-DS* model is an implementation of the *LP* method.

It is self-evident that $2^{|J|}$ conjoint-affects can be derived from J. We now describe the estimation of the $2^{|J|}$ posterior probabilities in equation (3.4) for each instance in I . Using Bayes' Theorem, we obtain the expectation as

$$\mathrm{E}\left[\mathcal{T}_i = \mathcal{J}\right] = \frac{\Pr\left[\left\{n_{i\mathcal{L}}^{(k)} : k \in \mathcal{K}_i, \mathcal{L} \subseteq J\right\} \mid \mathcal{T}_i = \mathcal{J}\right] \cdot \Pr\left[\mathcal{T}_i = \mathcal{J}\right]}{\Pr\left[\left\{n_{i\mathcal{L}}^{(k)} : k \in \mathcal{K}_i, \mathcal{L} \subseteq J\right\}\right]}. \tag{3.5}$$

In the *O-DS* model, each clinician's predilections, which are called *error rates*, are captured in confusion matrix π, where $\pi_{jl}^{(k)}$ specifies how likely clinician k declares a patient to be in state l when the patient is actually in state j. In the *D-DS* model, the annotator bias $\pi_{\mathcal{J}\mathcal{L}}^{(k)}$ $(k \in K, \mathcal{J} \subseteq J, \mathcal{L} \subseteq J)$ is defined as the probability that annotator k annotated an instance with conjoint-affect $\mathcal{L}$ when the true conjoint-affect for the instance is $\mathcal{J}$:

$$\pi_{\mathcal{J}\mathcal{L}}^{(k)} := \frac{\sum_{i \in I} \gamma_{i\mathcal{J}} \cdot n_{i\mathcal{L}}^{(k)}}{\sum_{\mathcal{L} \subseteq J} \sum_{i \in I} \gamma_{i\mathcal{J}} \cdot n_{i\mathcal{L}}^{(k)}}, \tag{3.6}$$

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

$$\gamma_{i\mathcal{J}} := \Pr\left[\mathcal{T}_i = \mathcal{J}\right]. \tag{3.7}$$

The numbers of times that annotator k annotated instance i with different conjoint-affects $\mathcal{L} \subseteq J$ when $\mathcal{J}$ is the true conjoint-affect are distributed according to a multinomial distribution, i.e.,

$$\Pr\left[\left\{n_{i\mathcal{L}}^{(k)} : \mathcal{L} \subseteq J\right\} \mid \mathcal{T}_i = \mathcal{J}\right] = \frac{\left(\sum_{\mathcal{L}\subseteq J} n_{i\mathcal{L}}^{(k)}\right)!}{\prod_{\mathcal{L}\subseteq J}\left(n_{i\mathcal{L}}^{(k)}!\right)} \cdot \prod_{\mathcal{L}\subseteq J}\left(\pi_{\mathcal{J}\mathcal{L}}^{(k)}\right)^{n_{i\mathcal{L}}^{(k)}}.$$

Since the annotators make their decisions independently, we assume that given the true conjoint-affect, each annotator's capability to make the correct judgment is conditionally independent of that of other annotators, i.e.,

$$\begin{aligned}\Pr\left[\left\{n_{i\mathcal{L}}^{(k)} : k \in \mathcal{K}_i, \mathcal{L} \subseteq J\right\} \mid \mathcal{T}_i = \mathcal{J}\right] &= \prod_{k\in\mathcal{K}_i} \Pr\left[\left\{n_{i\mathcal{L}}^{(k)} : \mathcal{L} \subseteq J\right\} \mid \mathcal{T}_i = \mathcal{J}\right] \\ &= \prod_{k\in\mathcal{K}_i}\left(\frac{\left(\sum_{\mathcal{L}\subseteq J} n_{i\mathcal{L}}^{(k)}\right)!}{\prod_{\mathcal{L}\subseteq J}\left(n_{i\mathcal{L}}^{(k)}!\right)} \cdot \prod_{\mathcal{L}\subseteq J}\left(\pi_{\mathcal{J}\mathcal{L}}^{(k)}\right)^{n_{i\mathcal{L}}^{(k)}}\right).\end{aligned} \tag{3.8}$$

That different conjoint-affects are true for instance i are mutually exclusive events. From the law of total probability, we have

$$\begin{aligned}&\Pr\left[\left\{n_{i\mathcal{L}}^{(k)} : k \in \mathcal{K}_i, \mathcal{L} \subseteq J\right\}\right] \\ &= \sum_{\mathcal{J}\subseteq J}\left(\Pr\left[\left\{n_{i\mathcal{L}}^{(k)} : k \in \mathcal{K}_i, \mathcal{L} \subseteq J\right\} \mid \mathcal{T}_i = \mathcal{J}\right] \cdot \Pr\left[\mathcal{T}_i = \mathcal{J}\right]\right) \\ &= \sum_{\mathcal{J}\subseteq J}\left(\prod_{k\in\mathcal{K}_i}\left(\frac{\left(\sum_{\mathcal{L}\subseteq J} n_{i\mathcal{L}}^{(k)}\right)!}{\prod_{\mathcal{L}\subseteq J}\left(n_{i\mathcal{L}}^{(k)}!\right)} \cdot \prod_{\mathcal{L}\subseteq J}\left(\pi_{\mathcal{J}\mathcal{L}}^{(k)}\right)^{n_{i\mathcal{L}}^{(k)}}\right) \cdot \gamma_{i\mathcal{J}}\right).\end{aligned} \tag{3.9}$$

Finally, by substituting Equations (3.7), (3.8), and (3.9) into Equation (3.5), we can

estimate the expectation in Equation (3.3) using

$$\begin{aligned}\mathrm{E}\left[\mathcal{T}_i=\mathcal{J}\right]=&\frac{\left(\prod_{k\in\mathcal{K}_i}\left(\frac{\left(\sum_{\mathcal{L}\subseteq J}n_{i\mathcal{L}}^{(k)}\right)!}{\prod_{\mathcal{L}\subseteq J}\left(n_{i\mathcal{L}}^{(k)}!\right)}\cdot\prod_{\mathcal{L}\subseteq J}\left(\pi_{\mathcal{J}\mathcal{L}}^{(k)}\right)^{n_{i\mathcal{L}}^{(k)}}\right)\right)\cdot\gamma_{i\mathcal{J}}}{\sum_{\mathcal{J}\subseteq J}\left(\prod_{k\in\mathcal{K}_i}\left(\frac{\left(\sum_{\mathcal{L}\subseteq J}n_{i\mathcal{L}}^{(k)}\right)!}{\prod_{\mathcal{L}\subseteq J}\left(n_{i\mathcal{L}}^{(k)}!\right)}\cdot\prod_{\mathcal{L}\subseteq J}\left(\pi_{\mathcal{J}\mathcal{L}}^{(k)}\right)^{n_{i\mathcal{L}}^{(k)}}\right)\cdot\gamma_{i\mathcal{J}}\right)}\\=&\frac{\gamma_{i\mathcal{J}}}{z_i}\cdot\prod_{k\in K}\prod_{\mathcal{L}\subseteq J}\left(\pi_{\mathcal{J}\mathcal{L}}^{(k)}\right)^{n_{i\mathcal{L}}^{(k)}},\end{aligned}\tag{3.10}$$

where z_i is defined as the normalization constant:

$$z_i:=\sum_{\mathcal{J}\subseteq J}\left(\gamma_{i\mathcal{J}}\cdot\prod_{k\in K}\prod_{\mathcal{L}\subseteq J}\left(\pi_{\mathcal{J}\mathcal{L}}^{(k)}\right)^{n_{i\mathcal{L}}^{(k)}}\right).\tag{3.11}$$

This ensures that the expectation $\mathrm{E}\left[\mathcal{T}_i=\mathcal{J}\right]$ is a valid probability density and can be integrated to one.

3.2.3 Label Pairwise Dawid-Skene (*P-DS*) model

It is generally agreed that if two affect labels are similar or opposite, they are strongly correlated. Let us consider an attendance intention annotation task with four candidate labels, *have attended*, *plan to attend*, *want to attend*, and *no intention of attending*. It is obvious that the first two labels are strongly correlated, as are the last two, because someone who has already attended an activity (like an annual festival) would not likely plan to attend again, and someone who has no intention of attending would also be unlikely to want to attend. However, the four affect labels are not so generally correlated. Unfortunately, neither the "independent assumption" of the *O-DS* model nor the "dependent assumption" of the *D-DS* model can properly depict the relationships among these four labels. In view of this, we propose grouping candidate labels into pairs and then estimating the states of the two labels within each pair separately using $|J|/2$ mutually independent models, each of which can be seen as a "two-label version of the *D-DS* model", in order to prevent interference from uncorrelated labels. The differences among the *O-DS*, *D-DS*, and *P-DS* models

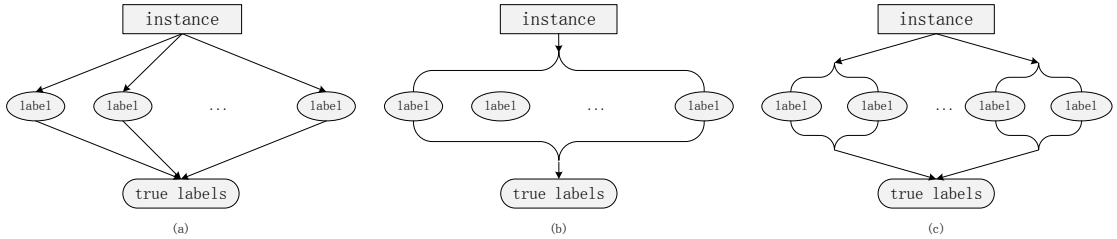


Figure 3-2: Multi-label estimation models: (a) *O-DS*, (b) *D-DS*, and (c) *P-DS*.

are illustrated in Figure 3-2.

The crucial problem with the *P-DS* model is how to pair the candidate labels. Let $\mathrm{H}[a,b]\,(a \in J, b \in J, a \neq b)$ be the joint entropy of labels a and b. The optimal pairing pattern S is the one that minimizes the sum of the joint entropies of all label pairs:[2]

$$\arg\min_{S} \sum_{\{a,b\} \in S} \mathrm{H}[a,b].$$

Our experiments demonstrated that most label pairs contain synonymous or antonymous labels. Especially, the pairing pattern described above for the four intention labels is the optimal one for handling the intention annotation experiment, as discussed in Section 3.5.2.

3.3 Inference Algorithm

Let us take the *D-DS* model as an example because the *P-DS* model can be considered to consist of several two-label *D-DS* models. Inference of the true labels and the parameters can be greatly simplified if we use the EM-based 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

[2]For a detailed proof of this, see Appendix A.

research [2, 3, 5]. We treat the maximum-likelihood estimates of the parameters

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

as the hidden data, with the expectations of the true conjoint-affects for each instance

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

as unobserved variables (missing data). The observed variables

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

can be directly calculated from the obtained crowdsourced annotations.

We then proceed as follows:

(1) Initialization[3]

Obtain the initial estimates of unobserved variables $\{\mathrm{E}\left[\mathcal{T}_i = \mathcal{J}\right] : i \in I, \mathcal{J} \subseteq J\}$.

(2) Maximization (M-Step)

Estimate the maximum-likelihood estimates of parameters $\left\{\pi^{(k)}_{\mathcal{J}\mathcal{L}} : k \in K, \mathcal{J} \subseteq J, \mathcal{L} \subseteq J\right\}$, $\{\gamma_{i\mathcal{J}} : i \in I, \mathcal{J} \subseteq J\}$, and $\{z_i : i \in I\}$ by using Equations (3.6), (3.7), and (3.11), with the current estimates of $\{\mathrm{E}\left[\mathcal{T}_i = \mathcal{J}\right] : i \in I, \mathcal{J} \subseteq J\}$, which means

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

[3]This step is discussed in more detail in Section 3.4.

(3) Expectation (E-Step)

Estimate the expected values of $\{\mathrm{E}[\mathcal{T}_i = \mathcal{J}] : i \in I, \mathcal{J} \subseteq J\}$ using Equation (3.10) 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 $\Pr\left[\left\{n_{i\mathcal{L}}^{(k)} : i \in I, k \in \mathcal{K}_i, \mathcal{L} \subseteq J\right\}\right]$ converge. Since that all instances are annotated independently, and that the multinomial coefficient $\frac{\left(\sum_{\mathcal{L}\subseteq J} n_{i\mathcal{L}}^{(k)}\right)!}{\prod_{\mathcal{L}\subseteq J}\left(n_{i\mathcal{L}}^{(k)}!\right)}$ is a constant for a certain $< i, k >$ pair, from Equation (3.9), we have

$$\Pr\left[\left\{n_{i\mathcal{L}}^{(k)} : i \in I, k \in \mathcal{K}_i, \mathcal{L} \subseteq J\right\}\right] \propto C\,,$$

where

$$C = \prod_{i \in I} z_i\,. \tag{3.12}$$

At this point, the $\mathcal{J}$ with the maximum $\mathrm{E}[\mathcal{T}_i = \mathcal{J}]$ is the estimated true conjoint-affect for instance i, as specified in Equation (3.3). Pseudo-code for this strategy is given in Algorithm 1 .

To avoid the "zero frequency problem" in the M-Step, π is estimated using Lidstone's smoothing raw [80]. Note that if annotator k annotated only a certain instance with conjoint-affect $\mathcal{L}$ one time and did not annotate any other instances, for k's *error rate* matrix, $\pi_{\mathcal{J}\mathcal{L}}^{(k)} = 1$ and $\pi_{\mathcal{J}\mathcal{L}'}^{(k)} = 0$ ($\mathcal{J} \subseteq J, \mathcal{L}' \subseteq J, \mathcal{L}' \neq \mathcal{L}$) constantly within iterations. Therefore, to estimate an annotator's *error rate* matrix, at least two annotations provided by that annotator must be collected. This requirement may decrease the flexibility of crowdsourcing somewhat.

One of the characteristics of the EM algorithm is that, after alternately performing steps (2) and (3), only one of the unobserved variables for an instance has a probability converging towards 1 while the other unobserved variables for the instance have a probability converging towards 0. In other words, it is unlikely that any of the expectations $\{\mathrm{E}[\mathcal{T}_i = \mathcal{J}] : i \in I, \mathcal{J} \subseteq J\}$ is between 0.1 and 0.9.

Algorithm 1 Multi-label Estimation from Crowdsourced Annotations.

Input: $\left\{n_{i\mathcal{L}}^{(k)} : i \in I, k \in \mathcal{K}_i, \mathcal{L} \subseteq J\right\}$.
Output: $\{\mathcal{T}_i : i \in I\}$.

for each $\langle i \in I, \mathcal{J} \subseteq J\rangle$ **do**

initialize $a = 0$, $\mathrm{E}\left[\mathcal{T}_i = \mathcal{J}\right]_0$ (discussed in Section 3.4)

end for

while ($a = 0$ or $C_a \neq C_{a-1}$) **do**

- $a = a + 1$
- **M-Step: for each** $\langle i \in I, \mathcal{J} \subseteq J\rangle$ **do**

  calculate $\Pr\left[\mathcal{T}_i = \mathcal{J}\right]_a$ using Equation (3.3) with $\mathrm{E}\left[\mathcal{T}_i = \mathcal{J}\right]_{a-1}$.

  end for
- **M-Step: for each** $\langle i \in I, \mathcal{J} \subseteq J\rangle$ **do**

  calculate $\left[\gamma_{i\mathcal{J}}\right]_a$ using Equation (3.7) with $\Pr\left[\mathcal{T}_i = \mathcal{J}\right]_a$.

  end for
- **M-Step: for each** $\langle k \in K, \mathcal{J} \subseteq J, \mathcal{L} \subseteq J\rangle$ **do**

  calculate $\left[\pi_{\mathcal{J}\mathcal{L}}^{(k)}\right]_a$ using Equation (3.6) with $\left[\gamma_{i\mathcal{J}}\right]_a$.

  end for
- **M-Step: for each** $\langle i \in I\rangle$ **do**

  calculate $\left[z_i\right]_a$ using Equation (3.11) with $\left[\gamma_{i\mathcal{J}}\right]_a$ and $\left[\pi_{\mathcal{J}\mathcal{L}}^{(k)}\right]_a$.

  end for
- **E-Step: for each** $\langle i \in I, \mathcal{J} \subseteq J\rangle$ **do**

  calculate $\mathrm{E}\left[\mathcal{T}_i = \mathcal{J}\right]_a$ using Equation (3.10) with $\left[\gamma_{i\mathcal{J}}\right]_a$, $\left[\pi_{\mathcal{J}\mathcal{L}}^{(k)}\right]_a$, and $\left[z_i\right]_a$.

  end for
- calculate C_a using Equation (3.12) with $\left[z_i\right]_a$.

end while

output $\mathcal{T}_i$ using Equation (3.3) with $\mathrm{E}\left[\mathcal{T}_i = \mathcal{J}\right]_a$ for each $i \in I$.

3.4 Discussion: Bayesian Network Label-dependent DS (*D-DS+*) Model

Recall that there is an unsolved problem in the first step of the EM algorithm described in Section 3.3: how to initialize the estimates of unobserved variables $\{\mathrm{E}[\mathcal{T}_i = \mathcal{J}] : i \in I, \mathcal{J} \subseteq J\}$. Let $x_{i\mathcal{J}}^{(j)} \in \{0,1\}$ $(j \in J, i \in I, \mathcal{J} \subseteq J)$ be the state of the jth label in conjoint-affect $\mathcal{J}$ for instance i. One possible and intuitive way to initialize the estimates is to assign

$$\mathrm{E}[\mathcal{T}_i = \mathcal{J}] = \Pr\left[\left\{x_{i\mathcal{J}}^{(j)} : j \in J\right\}\right] = \frac{\sum_{k \in \mathcal{K}_i} n_{i\mathcal{J}}^{(k)}}{\sum_{\mathcal{L} \subseteq J} \sum_{k \in \mathcal{K}_i} n_{i\mathcal{L}}^{(k)}}, \tag{3.13}$$

which is an intuitive way to assign the maximum-likelihood estimates. This is indeed used in the *D-DS* model. This approach can be computationally demanding because it is equivalent to estimating a $|J|$-dimensional joint distribution for each instance over the candidate labels. Because the label 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. 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. This means that the *D-DS* model can easily suffer from the sparsity of high-dimensional annotations. An effective strategy to overcome this problem is to represent the underlying joint distribution more compactly, and to approximate the distribution from a finite number of samples by using the conditional independence properties of the joint distribution.

To motivate our discussion, we first assume that all candidate affect labels are statistically independent. That is, the completely general joint distribution in Equation (3.13) can be approximated as the product of the independent distributions of the candidate labels:

$$\mathrm{E}[\mathcal{T}_i = \mathcal{J}] = \prod_{j \in J} \Pr\left[x_{i\mathcal{J}}^{(j)}\right]. \tag{3.14}$$

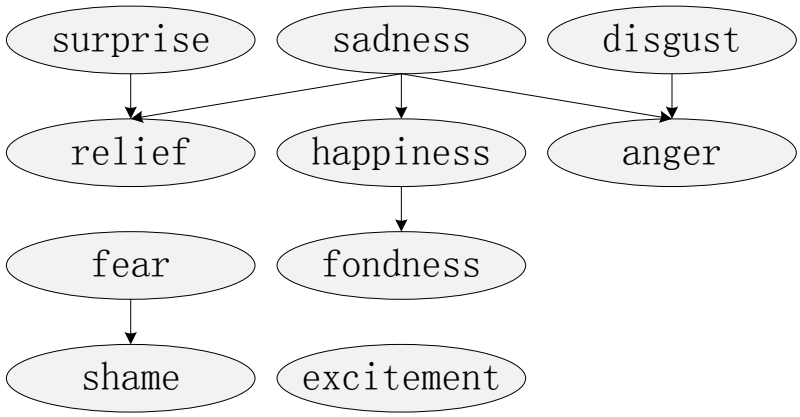


Figure 3-3: Example Bayesian network for emotion annotation experiment.

Intuitively, this simple assumption of ignoring the dependency relationships among affect labels is unreasonable in most cases, as we explained at the beginning of Section 3.2.2. There have been several proposals for approximating high-dimensional joint distributions. Chow and Liu [81], for example, addressed this problem by approximating an n-dimensional joint distribution as the product of $n-1$ second-order component distributions, where the relationships among random variables are represented by a *dependence tree*. Here we represent label dependency as a Bayesian network and call this extended *D-DS* model the "Bayesian network *D-DS* (*D-DS+*) model". Figure 3-3 shows an example Bayesian network for the emotion annotation experiment described in Section 3.5.1. The corresponding approximate product of the joint distribution is

$$\begin{aligned}\Pr[\mathcal{J}] =& \Pr[\text{surprise}] \cdot \Pr[\text{sadness}] \cdot \Pr[\text{disgust}] \cdot \Pr[\text{relief} \mid \text{surprise}, \text{sadness}] \\ & \cdot \Pr[\text{happiness} \mid \text{sadness}] \cdot \Pr[\text{anger} \mid \text{sadness}, \text{disgust}] \cdot \Pr[\text{fear}] \\ & \cdot \Pr[\text{fondness} \mid \text{happiness}] \cdot \Pr[\text{shame} \mid \text{fear}] \cdot \Pr[\text{excitement}] .\end{aligned}$$

Since the number of annotations for one instance is not sufficient for learning a Bayesian network, in the *D-DS+* model, all instances are assumed to share an identical Bayesian network, which is learned from the annotations for all instances provided by all annotators. This is reasonable because the relationships among candidate labels are independent of instances and annotators. We build the network structure of

candidate labels using the "PC" algorithm [82], which is based on hypothesis testing. To test whether two labels x_a and x_b are conditionally independent given a subset of other labels $\bar{X}$, we compute the conditional mutual information of these two labels,

$$\mathrm{CMI}\left[x_a; x_b \mid \bar{X}\right] = \sum_{\bar{X}} \Pr\left[\bar{X}\right] \sum_{x_a, x_b} \Pr\left[x_a, x_b \mid \bar{X}\right] \log \frac{\Pr\left[x_a, x_b \mid \bar{X}\right]}{\Pr\left[x_a \mid \bar{X}\right] \Pr\left[x_b \mid \bar{X}\right]},$$

by using the maximum-likelihood estimates on annotations for all instances. Under the independence assumption, $2m * \mathrm{CMI}\left[x_a; x_b \mid \bar{X}\right]$ follows a χ^2 distribution with degrees of freedom equal to $2^{|\bar{X}|}$, where m is the sample size $|I| * |K|$.

Although we take the p-value for rejecting the null hypothesis that two labels are conditionally dependent as 0.1, it is worth mentioning that if the p-value is 1, all labels are determined to be unconditionally independent of each other, and the approximation strategy of the *D-DS+* model is the same as Equation (3.14). Likewise, the *D-DS+* model is equivalent to the *D-DS* model if the p-value is 0, which means that the network structure is a complete directed acyclic graph, and the depicted approximate product of the joint distribution is the chain rule for Equation (3.13).

In summary, we proposed two models, *D-DS* and *P-DS*, for estimating associated labels for each instance given crowdsourced multi-label affect annotations. Moreover, we extended the *D-DS* model to create the *D-DS+* model, using the Bayesian network to approximate the joint distribution over the candidate labels.

3.5 Empirical Study

3.5.1 Emotion Annotation for Narratives

To create a first test bed for the proposed models containing actual annotations obtained from the Lancers crowdsourcing service, we asked crowdsourcing annotators to read some narrative sentences and spontaneously indicate the character's emotions expressed in each sentence. The associated (gold standard) labels for each sentence are then estimated by aggregating the obtained multi-label emotion annotations.

To simplify the task, 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 tint, distinct character personalities, as the aim is to better attract the attention of children. The proportion of speaking sentences in children's narratives is also higher than that in other narrative genres. Therefore, children's narratives and fairy tales are commonly used in emotion detection research (as shown in Table 2.7). We thus chose two Japanese children's narratives, "Although we are in love"[4] ("Love" for short) and "Little Masa and a red apple"[5] ("Apple" for short), as the texts to be annotated. As the source of the annotated texts we used the Aozora Library[6].

Due to different aspects that emotion-oriented research looks to capture, the set of candidate labels used differs among research efforts. Generally, simple emotion label sets give better performance than expanded sets of emotion labels which require cognitive information and deeper understanding of the subject. While "the Big Six" emotion labels [83] (i.e., *happiness*, *fear*, *anger*, *surprise*, *disgust*, and *sadness*) and the related emotion sets are typically used in emotion-oriented research (e.g., [11, 20, 26]), we used ten emotion labels as the candidate labels in order to provide more choices for the annotators and thereby enable us to perform a more in-depth study on multi-affect estimation. The emotions were taken from the "Emotive Expression Dictionary"[7] [60]. We conducted the experiments using the Lancers crowdsourcing service. An example task input screen is shown in Figure 3-4. The annotators were told to check *neutral* if none of the listed emotions was felt. The annotation frequencies of the labels are shown in Table 3.1, and other statistics about the datasets are shown in Table 3.2.

People have different tendencies when detecting subjective feelings, so two people may be affected differently by the same sentence. This means that 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. Although there are statistical

[4]http://www.aozora.gr.jp/cards/001475/files/52111_47798.html
[5]http://www.aozora.gr.jp/cards/001475/files/52113_46622.html
[6]For a detailed description, see Section 2.4.1.
[7]For a detailed description, see Section 2.4.2

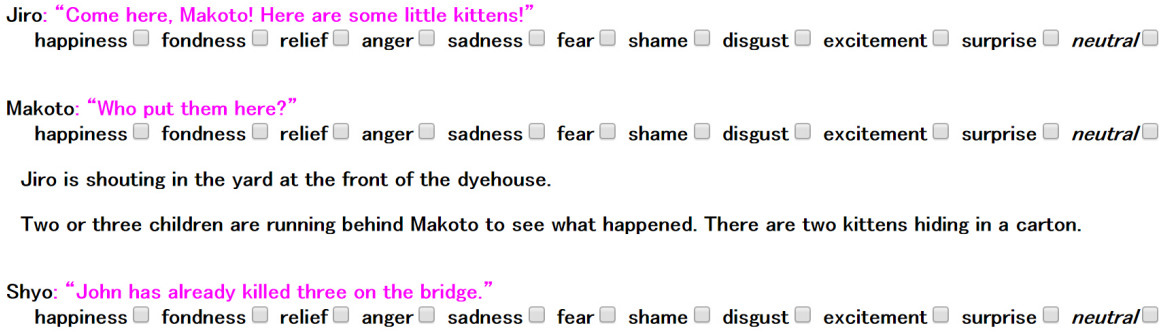


Figure 3-4: Example task input screen (translated from Japanese). Annotators were native Japanese language speakers. Both candidate emotion labels and sentences in the two narratives were presented to annotators in their original Japanese form.

Table 3.1: Annotation frequencies of emotion labels and *neutral*, ordered by total annotation frequencies of per label.

Emotion label	Freq. in "Love"	Freq. in "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

Table 3.2: Statistics for emotion annotation experiment.

	"Love"	"Apple"	Total
No. of annotations	1890	2340	4230
No. of annotators	30	57	84
No. of sentences	63	78	141
Avg. no. of checked labels per annotation	1.75	1.18	1.43
Avg. no. of annotations per sentence	30	30	30

theories (e.g. CrowdSense.Bin [1]) discussing the lower bound on how many voters (subcrowd) could agree with the crowd's majority vote, the discussion is beyond the scope of this dissertation. Under the budget limitation constraints, we obtained gold standards by having each sentence annotated 30 times and then taking the *majority vote*. That is, the most often annotated conjoint-affect for a sentence was used as the gold standard for that sentence.

For the "Love" narrative, we asked each of 30 annotators to annotate each sentence one time, which ensured that each annotator annotated the complete set of sentences. 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 sentence were provided by arbitrary annotators, and few, if any, of them annotated the complete set of the sentences. 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 large 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.

Moreover, although our proposed models can handle a sentence 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 sentence differently at different times. Therefore, in our experiments, all the annotations for a sentence were obtained from different annotators. This means that the values of the observed variables $n_{ii}^{(k)}$ in equation (3.2) and $n_{i\mathcal{L}}^{(k)}$ in equation (3.4) are either 0 or 1.

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

- *MV*: **M**ajority **V**ote;
- *O-DS*: **O**riginal **D**awid-**S**kene model;
- *D-DS*: Label-**D**ependent **D**awid-**S**kene model;
- *P-DS*: Label **P**airwise **D**awid-**S**kene model;
- *D-DS+*: Bayesian Network label-**D**ependent **D**awid-**S**kene model.

Both the estimation result and the gold standard for a sentence can be regarded as a binary vector: if a label is present in a label set, the value of the corresponding element in the vector is 1, and -1 otherwise. It is unreasonable to check whether the two binary vectors match exactly. For example, {*anger*, *excitement*} is closer to {*anger*, *sadness*} than {*happiness*, *surprise*}. Therefore, the average *Simple Matching Coefficient* (defined in Equation (3.15)) is used to evaluate the performance of the proposed models, i.e., the average proportion of state-consistent labels between the estimation results t_{result} and the gold standards for all sentences t_{gold} within a group:

$$\text{similarity}\left(t_{\text{result}}, t_{\text{gold}}\right) = \frac{\sum_{\imath \in T}\left(2 - |t^{(\imath)}_{\text{result}} - t^{(\imath)}_{\text{gold}}|\right)}{2 \cdot |T|}. \tag{3.15}$$

The *MV* results for 30 annotators represent the accuracy rate (100%) of the gold standard group. Figure 3-5 shows the experimental strategy.

As shown in Table 3.3 and Figure 3-6, for the "Love" narrative, the statistical models achieved better or comparable accuracy rates than the *majority vote* strategy when the group size was 3, 5, or 10. As shown in Table 3.4 and Figure 3-7, for the "Apple" narrative, the statistical models achieved better accuracy rates when the group size was 3 or 5, and two of them achieved better or comparable accuracy rates when the group size was 10. This means that ten annotators at most for each sentence would be a reasonable number. Moreover, the *D-DS+* model consistently outperformed the *D-DS* model. This means that a learned Bayesian network is effective for approximating the high-dimensional joint distribution over ten labels from a limited number of annotations. Although the *O-DS*, *P-DS*, and *D-DS+* models

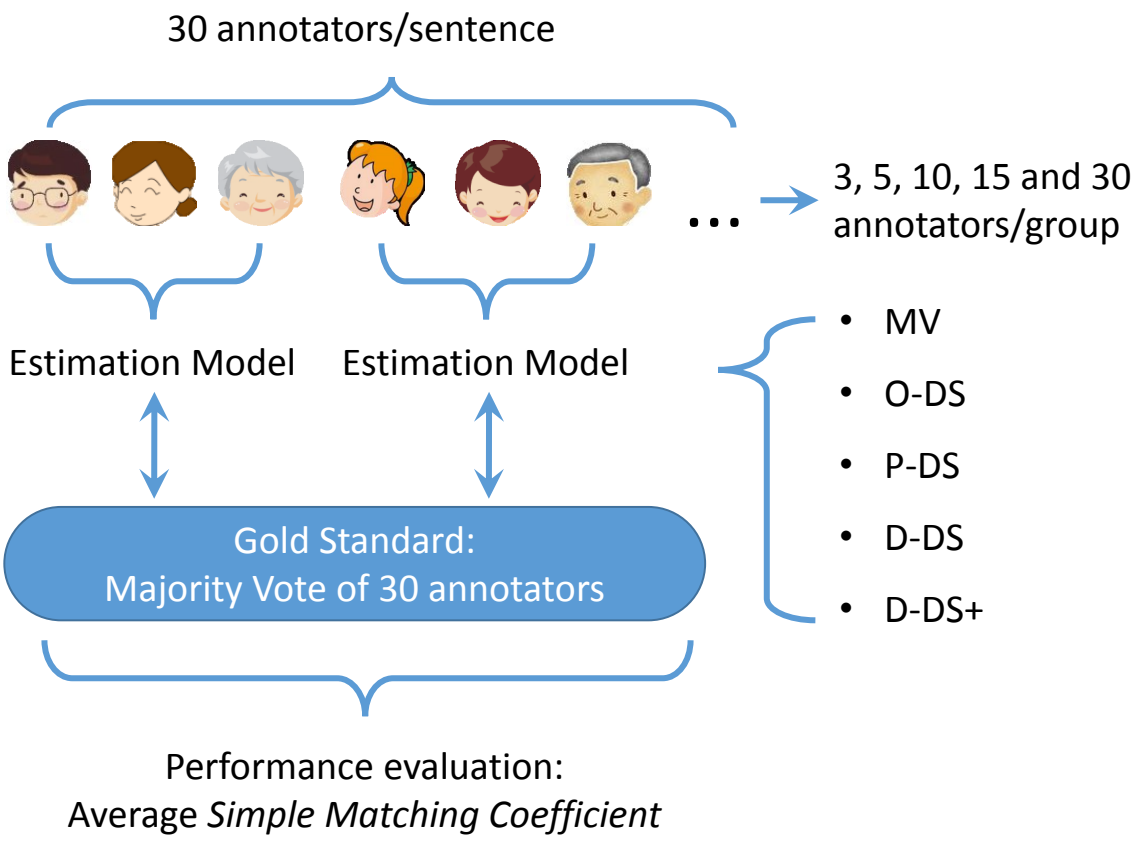


Figure 3-5: Experimental strategy.

had virtually the same average accuracy rate for three annotators per sentence, the *D-DS+* model had significantly better accuracy rates (greater than 90%) for five or more annotators per sentence. In other words, the *D-DS+* model can most effectively handle the multi-affect estimation problem in this experiment, with annotations provided by only about five crowdsourcing annotators per instance. One noteworthy result is that the average accuracies of the *O-DS* and *P-DS* models remained basically unchanged as the group size increased for the "Love" task while they decreased for the "Apple" task. This could be because these two models are more sensitive to the effects of "arbitrary annotator interference".

3.5.2 Intention Annotation for Tweets

In the second experiment, intention annotation for tweets, we collected 1398 tweets on the Twitter micro-blogging service related to the Sapporo Snow Festival. We again

Table 3.3: Average emotion annotation accuracies for "Although we are in love" narrative.

Model	**No. of annotators per sentence**				
	3	**5**	**10**	**15**	**30**
MV	0.8168	0.8354	0.8968	0.9238	1 (Gold Standard)
O-DS	0.8936	0.8981	0.8962	0.8841	0.8809
P-DS	0.8917	0.8955	0.8968	0.8880	0.8968
D-DS	0.8747	0.8783	0.9116	0.9277	0.9523
D-DS+	0.8906	0.9103	0.9317	0.9404	0.9587

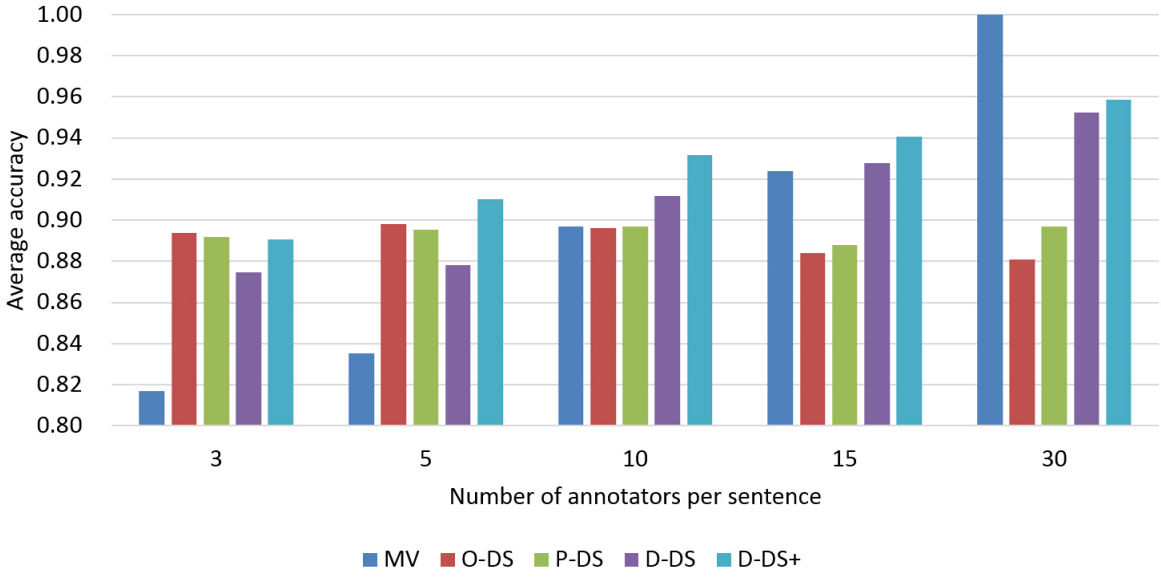


Figure 3-6: Average emotion annotation accuracies for "Although we are in love" narrative.

Table 3.4: Average emotion annotation accuracies for "Little Masa and a red apple" narrative.

Model	**No. of annotators per sentence**				
	3	**5**	**10**	**15**	**30**
MV	0.8653	0.9027	0.9500	0.9782	1 (Gold Standard)
O-DS	0.9375	0.9252	0.9102	0.9051	0.8910
P-DS	0.9283	0.9239	0.9175	0.9141	0.9166
D-DS	0.9266	0.9384	0.9512	0.9570	0.9692
D-DS+	0.9382	0.9521	0.9632	0.9673	0.9743

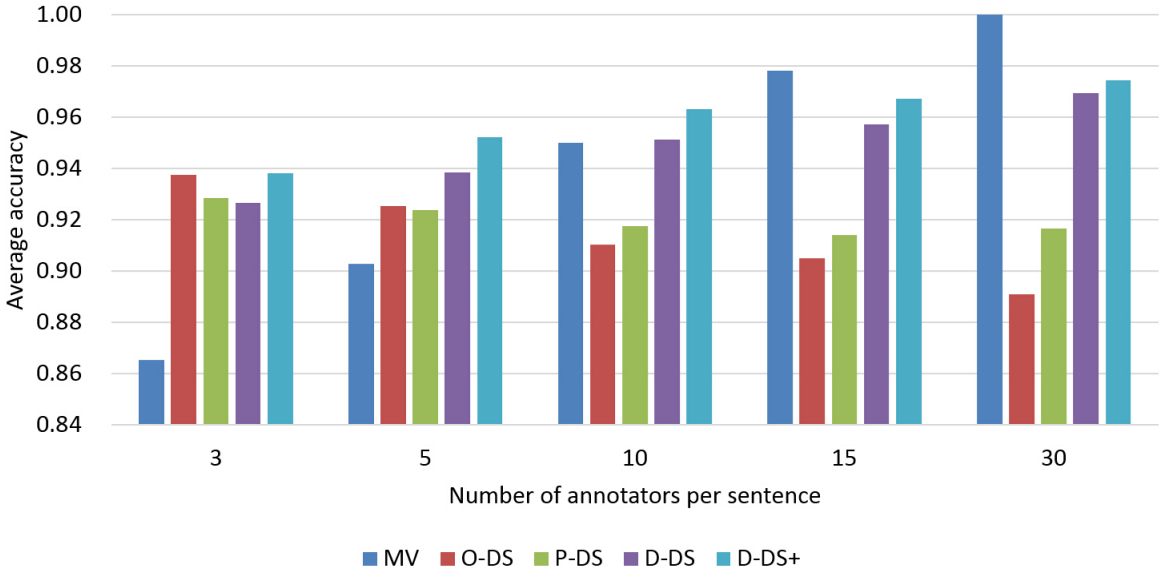


Figure 3-7: Average emotion annotation accuracies for "Little Masa and a red apple" narrative.

used the Lancers crowdsourcing service and asked annotators to infer the attendance intentions of the tweet poster and then select appropriate ones from four intention labels: *have no intention of attending*, *want to attend*, *plan to attend*, and*have attended*. Each tweet was annotated by five arbitrary annotators. The annotation frequencies of the intention labels are shown in Table 3.5, and other statistics about the dataset are shown in Table 3.6.

We manually assigned reliable labels to each tweet and used them as the gold standards. The performance of the proposed models was measured in the same way as in the affect annotation experiment. As shown in Figure 3-8, all the statistical models as well as the *majority vote* strategy performed well due to the simplicity of the task. It is particularly noteworthy that the *P-DS* model had the highest accuracyrate, followed by the *D-DS+* model. The superior performance of the *P-DS* model is attributed to the fact that the four intention labels have a typical pairwise characteristic, as explained in Section 3.2.3.

3.6 Summary

We focused on crowdsourcing tasks fitting the paradigm of multi-label affect annotation, which means an instance can be associated with one or more label(s). The objective was to determine how many crowdsourcing annotators have to provide annotations in the affect labeling task in order for the aggregated annotation to be accurate. The statistical quality control models we proposed for the multi-affect estimation problem incorporated label dependency into the label-generation process. An EM-based incremental algorithm was used to estimate the associated labels for each instance as well as the maximum-likelihood estimates of the model parameters. Two experiments using Lancers crowdsourcing service showed that two of the models showed promising performance: in most cases, the *D-DS+* model most effectively handled the annotations provided by about five crowdsourcing annotators per instance. The *P-DS* model was the most effective when there were pairwise comparison relationships among candidate labels.

Table 3.5: Annotation frequencies of intention labels, ordered by annotation frequencies of per label.

Intention label	Frequency
have no intention of attending	2521
want to attend	2417
plan to attend	1365
have attended	1200
Total	7503

Table 3.6: Statistics for intention annotation experiment.

No. of annotators	94
No. of tweets	1398
No. of annotations	6990
Avg. no. of checked labels per annotation	1.07
Avg. no. of annotations per tweet	5

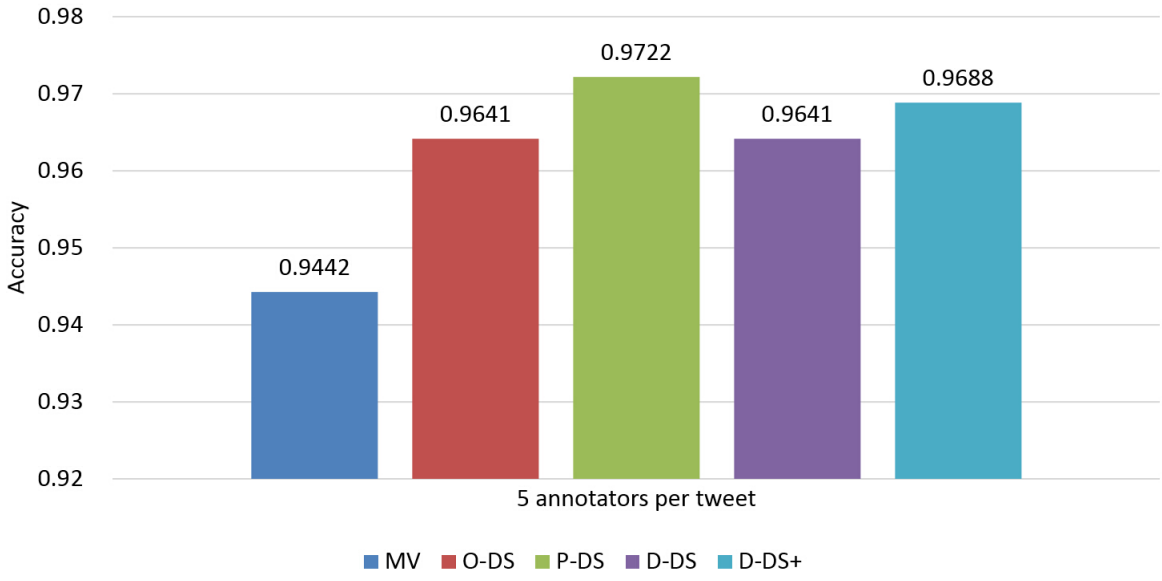


Figure 3-8: Average accuracies for intention annotation.

Two widely used methods for multi-label classification are the *BR* and the *LP* methods [28], which are the counterparts of the *O-DS* and the *D-DS* models in this chapter. The *O-DS* model simply decomposes the multi-label estimation problem into several independent binary-label estimation problems, one for each label in the set of candidate labels, and final associated labels for each instance are determined by aggregating the predictions from all binary estimators. A significant limitation of this model is that they do not take into account any dependency among candidate affect labels. Since multi-label tasks often have many candidate labels, if we simply incorporate dependency relationships among candidate affect labels into the label-generation process, as the *D-DS* model does, we may get data sets with a large number of classes and few samples per class. This means that the *D-DS* model can easily suffer from the sparsity of high-dimensional annotations, which makes the learning process difficult. Therefore, the *D-DS* model performs poorly for high-dimensional data sets.

To address these limitations, we proposed two approaches that flexibly use label dependency. In the first approach, the *P-DS* model is used to group candidate labels into pairs. The states of the two labels within each pair are then estimated separately in order to prevent interference from uncorrelated labels. The crucial problem is how to recognize pairwise comparison relationships among candidate affect labels. If the labels are pairwise correlated, the optimal pairing pattern should be the one that minimizes the sum of the joint entropies of all label pairs. The reason for this is explained in detail in A.

In the second approach, the underlying high-dimensional joint distribution over candidate labels is represented more compactly to enable it to be approximated from a finite number of annotations. The *D-DS+* model compactly represents label dependency using conditional independence properties to overcome the data sparsity problem. It depicts the properties as a Bayesian network, enabling the joint distribution to be approximated using the product of the conditional distributions of the candidate labels. Because the dependency relationships among candidate labels are independent of instances and annotators, the network is learned from the annota-

tions provided by all annotators for all instances. Experimental results showed the superiority of these two approaches.

Besides annotator bias and dependency relationships among candidate affect labels, for instances implicitly contain information for "consistency" and context, such as narrative sentences, such information is also important for the label-generation process. This will be discussed in Chapter 4.

Chapter 4

Multi-Emotion Estimation Considering Relationships among Instances

4.1 Introduction

In addition to the subjectivity characteristic discussed Chapter 3, "emotions" expressed by instances across the same context have their unique characteristics in comparison with other affect labels:

• Emotional consistency

Emotive expressions can not be divorced from their context [84, 85]. 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 chapter, 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.

• Contextual Cues

It is intuitively understandable that emotions 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.

Multi-emotion estimation in narratives is a domain-specific aspect of multi-affect estimation. To improve the accuracy of the domain-independent multi-affect estimation described in Chapter 3, in this Chapter, we explore the domain-specific viability of crowdsourced emotion annotations in narratives. We propose incorporating the internal relationships among sentences (instances) into the estimation process. We use the same data of the experiment discussed in Section 3.5.1. The relationships are specified as the aforementioned information for *emotional consistency* and *contextual cues* across sentences. The experimental results demonstrate that, from a limited number of crowdsourced annotations, incorporating relationships among instances into the estimation process enables gold standards to be more effectively estimated than the majority vote and the models only considering relationships among affect labels.

4.2 Statistical Models

The problem formulation and symbol declaration are the same as those in Chapter 3, where narrative sentences and the set of candidate emotion labels are the counterparts of instances I and the set of candidate affect labels J. The graphical representation of the based domain-independent model is illustrated in Figure 4-1 (a)

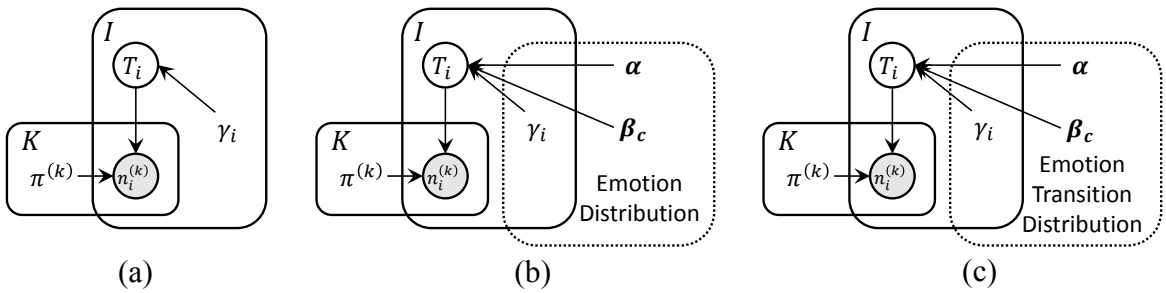


Figure 4-1: Graphical model representation for multi-affect estimation: (a) domain-independent model, (proposed in Chapter 3) (b) model considering *emotional consistency*, and (c) model considering *emotional consistency* with *contextual cues*.

4.2.1 Multi-Emotion Estimation Considering Emotional Consistency

As an extension to our previous work on domain-independent multi-affect estimation described in Chapter 3, we first propose incorporating *emotional consistency* into the estimation process. As discussed in Section 4.1, there are *emotional consistency* across sentences in narrative emotional tendency and in character personalities. 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 of the proportion of sentences expressing conjoint-emotion $\mathcal{J}$ over all sentences, namely the prior probability that a sentence drawn at random has conjoint-emotion $\mathcal{J}$.

Let $c(i)$ $(i \in I)$ denote the character speaking in sentence i and $I_{c(i)}$ $(i \in I)$ denote the sentences 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 maximum-likelihood estimate of the proportion of the sentences expressing conjoint-emotion $\mathcal{J}$ over all the sentences spoken by character $c(i)$.

As a domain-specific extension of Equation (3.4), the expectation of the true

conjoint-emotion for sentence i is then estimated considering *emotional consistency* in the narrative emotional tendency and in the character personality:

$$\mathrm{E}\left[\mathcal{T}_i = \mathcal{J}\right] = \Pr\left[\mathcal{T}_i = \mathcal{J} \mid \left\{n_{i\mathcal{L}}^{(k)} : k \in \mathcal{K}_i, \mathcal{L} \subseteq J\right\}; \alpha_{\mathcal{J}}, \beta_{c(i)\mathcal{J}}\right] . \tag{4.1}$$

This means that the true conjoint-emotion of each sentence 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 4-1 (b). In a manner similar to that for Equation (3.10), using Bayes' Theorem, we obtain the expectation in Equation (4.1):

$$\mathrm{E}\left[\mathcal{T}_i = \mathcal{J}\right] = \frac{\alpha_{\mathcal{J}} \cdot \beta_{c(i)\mathcal{J}} \cdot \gamma_{i\mathcal{J}}}{z_i} \cdot \prod_{k \in \mathcal{K}_i} \prod_{\mathcal{L} \subseteq J} \left(\pi_{\mathcal{J}\mathcal{L}}^{(k)}\right)^{n_{i\mathcal{L}}^{(k)}} , \tag{4.2}$$

where z_i is the normalization constant:

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

Equation (4.2) demonstrates that the domain-specific model considering *emotional consistency* 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.

4.2.2 Multi-Emotion Estimation Considering Emotional Consistency with Contextual Cues

As mentioned in Section 4.1, 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*. 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 4.2.1, we propose estimating *emotional consistency* across instances by using *contextual cues*. We use the idea of *bi-gram* to learn the transition distribution over conjoint-emotions. This means that the conjoint-emotion expressed by a sentence is conditional on the conjoint-emotion expressed by the previous sentence. Let $i-1$ be the sentence before sentence i and $\bar{\mathcal{J}}$ be the true conjoint-emotion of sentence $i-1$, 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 4.2.1. They are estimated considering the true conjoint-emotions for two consecutive sentences:

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

The annotations for each sentence 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 sentences:

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

The differences among the domain-independent multi-affect estimation model, the model considering *emotional consistency*, and the model considering *emotional consistency* with *contextual cues* are illustrated in Figure 4-1. The proposed domain-specific multi-emotion estimation models, (b) and (c), extend the domain-independent

multi-affect model (a) by statistically analyzing *emotional consistency* in narrative emotional tendency and character personality, which are estimated as probabilistic variables following different distributions.

4.3 Inference Algorithm

The inference algorithm is similar to the approach for the domain-independent multi-affect estimation problem described in Section 3.3. The differences are listed as follows.

Maximization (M-Step)

Parameters representing not only domain-independent information, but also domain-specific information are estimated:

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

Expectation (E-Step)

Estimate the expected values of $\{\mathrm{E}\left[\mathcal{T}_i = \mathcal{J}\right] : i \in I, \mathcal{J} \subseteq J\}$ using Equation (4.2) with the current estimates of the parameters calculated in the M-Step.

4.4 Empirical Study

We used the data of "emotion annotation for children's narratives" in Section 3.5.1 since the characteristics of children's narratives are also the focal points of our research. However, at this time, the associated labels for each sentence was estimated

given the annotations within each group using the following four models:

- *MV*: **M**ajority **V**ote;
- *D-DS+*: the best performed domain-independent multi-affect 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 *D-DS+* are the baselines to which we compare the results for *EC* and *EC+*. The performance evaluation strategy was also the same as the one in Section 3.5.1.

As shown in Tables 4.1, 4.2 and Figures 4-2, 4-3, 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 models increased with the group size, the *EC* and *EC+* models almost consistently outperformed the *D-DS+* model and had accuracies greater than 90 % for five or more annotators per sentence. This means that considering *emotional consistency* 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 sentence to achieve satisfactory performance. Moreover, the average accuracy rates 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 consistency* with *contextual cues*, became evident. Finally, none of the models was particularly sensitive to the effect of "arbitrary annotator interference" in the "apple" narrative, as the average accuracies remained basically unchanged as the group size increased.

Table 4.1: Average emotion annotation accuracies for "Although we are in love" narrative.

Model	**No. of annotators per sentence**				
	3	**5**	**10**	**15**	**30**
MV	0.8168	0.8354	0.8968	0.9238	1 (Gold Standard)
D-DS+	0.8906	0.9103	0.9317	0.9404	0.9587
EC	0.8984	0.9116	0.9318	0.9452	0.9623
EC+	0.8901	0.9229	0.9389	0.9570	0.9696

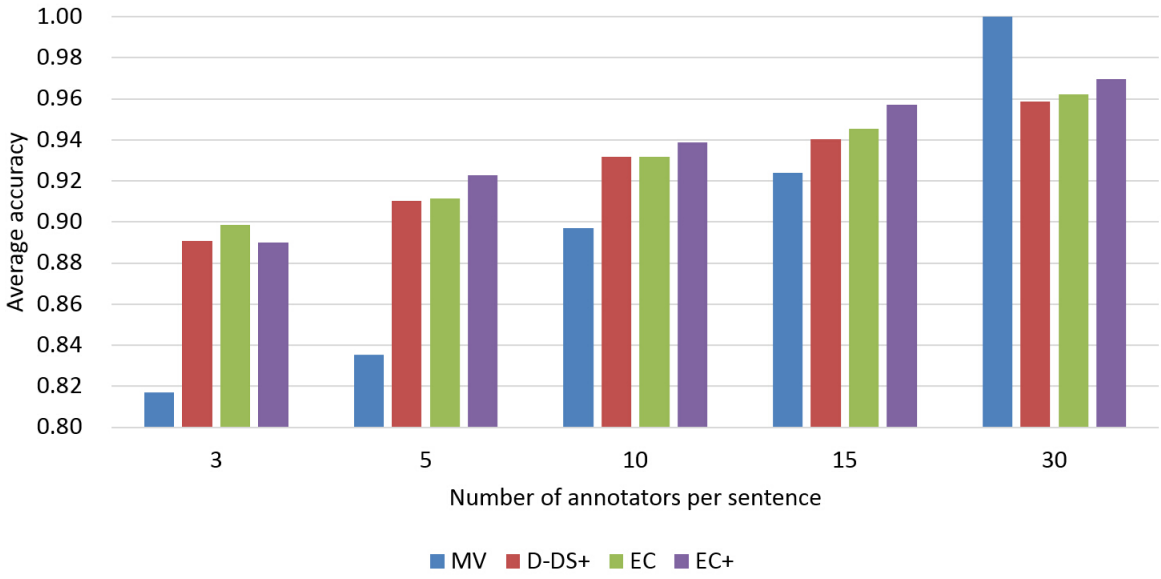


Figure 4-2: Average emotion annotation accuracies for "Although we are in love" narrative.

Table 4.2: Average emotion annotation accuracies for "Little Masa and a red apple" narrative.

Model	**No. of annotators per sentence**				
	3	**5**	**10**	**15**	**30**
MV	0.8653	0.9027	0.9500	0.9782	1 (Gold Standard)
D-DS+	0.9382	0.9521	0.9632	0.9673	0.9743
EC	0.9439	0.9566	0.9634	0.9719	0.9818
EC+	0.9321	0.9581	0.9706	0.9786	0.9851

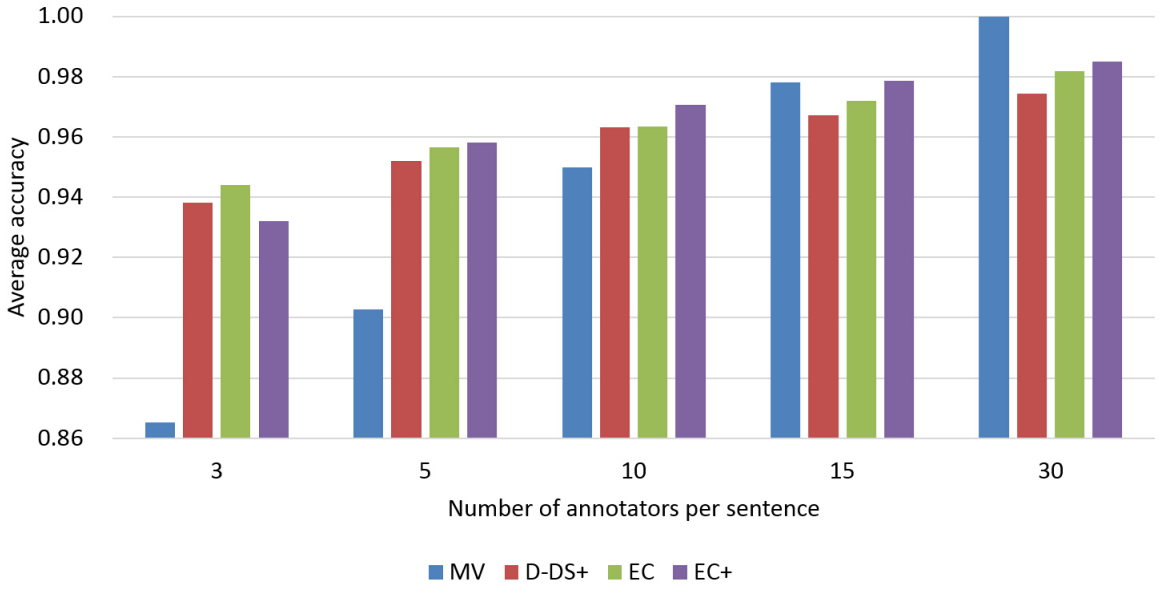


Figure 4-3: Average emotion annotation accuracies for "Little Masa and a red apple" narrative.

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 sentences, the number of annotators, and the number of candidate labels and the dataset used was not so large. The other is that we did not estimate the expectations $\{\mathrm{E}[\mathcal{T}_i = \mathcal{J}]\}$ 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.

4.5 Summary

The domain-independent multi-affect estimation model proposed in Chapter 3 ignored the internal relationships among instances. To improve the accuracy for narrative sentences, we extended it to specify those relationships as the domain-specific information. The information is extended by making use of the characteristics of emotions in narratives. An EM-based incremental algorithm was devised for estimating the gold standard emotion annotation for each narrative sentence together with the model parameters. The experimental results demonstrated that incorporating *emotional consistency* 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 demonstrated that incorporating *contextual cues* further improved the accuracy.

In this research, *bi-grams* were used to learn the transition distribution over conjoint-emotions. It would be worth investigating whether *tri-grams* or higher *n-grams* yield better results, especially when dealing with larger contexts. The 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 sent by the poster 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.

Chapter 5

Crowdsourced Semantic Matching of Multi-label Emotion Annotations

5.1 Introduction

In a multi-label domain, each instance is associated with a 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, as shown in table 2.2, the taxonomy used may differ among systems in the same multi-label 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 emotion domain. Before starting on an emotion-related technique, the first question is "Which emotions should be addressed?" There are many different emotion taxonomies exists in the literature in-

cluding basic emotions, universal emotions, primary and secondary emotions, neutral vs. emotional, and for some cases the problem is reduced to a two class classification problem (Sentiment Analysis) using the Positive and Negative values as affect labels. Due to different aspects that emotion-related techniques look to capture or just inconsistency in terminology usage, the taxonomy used differs among research efforts. Even though the taxonomy of Ekman's six basic emotions [83] (i.e., *happiness*, *fear*, *anger*, *surprise*, *disgust*, and *sadness*) and the related taxonomies has been used very broadly to cover a wide range of emotion-oriented research (e.g., [11, 20, 26]), other emotion taxonomies are also widely used. For example, Trohidis *et al.* [26] used six other emotions {*amazed-surprised*, *happy-pleased*, *relaxing-calm*, *quiet-still*, *sad-lonely*, and *angry-fearful*} based on the Tellegen-Watson-Clark taxonomy [86] to conduct the automated detection of emotion in music. The taxonomy used by the *manifold* emotion analyzer [12] consists of 32 emotions while the WordNet-Affect thesaurus [87] 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., [27], and the experiment discussed in Section 3.5.1) which tends to use the taxonomy of Nakamura's ten emotions [60], represented in Table 2.6. A complete discussion of the taxonomies used for emotion-oriented research is beyond the scope of this dissertation but can be found in Calvo and D'Mello [67].

The lack of an authoritative emotion taxonomy means that no general emotion taxonomy has yet been agreed on [88]. Therefore, emotion-related research is faced with the problem of inconsistency in taxonomy usage. For example, as shown in Figure 5-1, 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., {angry, sadness, disgust}) in the Ekman's taxonomy, while a text-to-speech synthesis system requires sentences with associated emotions in Nakamura's taxonomy as the input for affective pronunciation. This means that the output of the emotion detector cannot be used as the input to the expressive text-to-speech synthesis system since

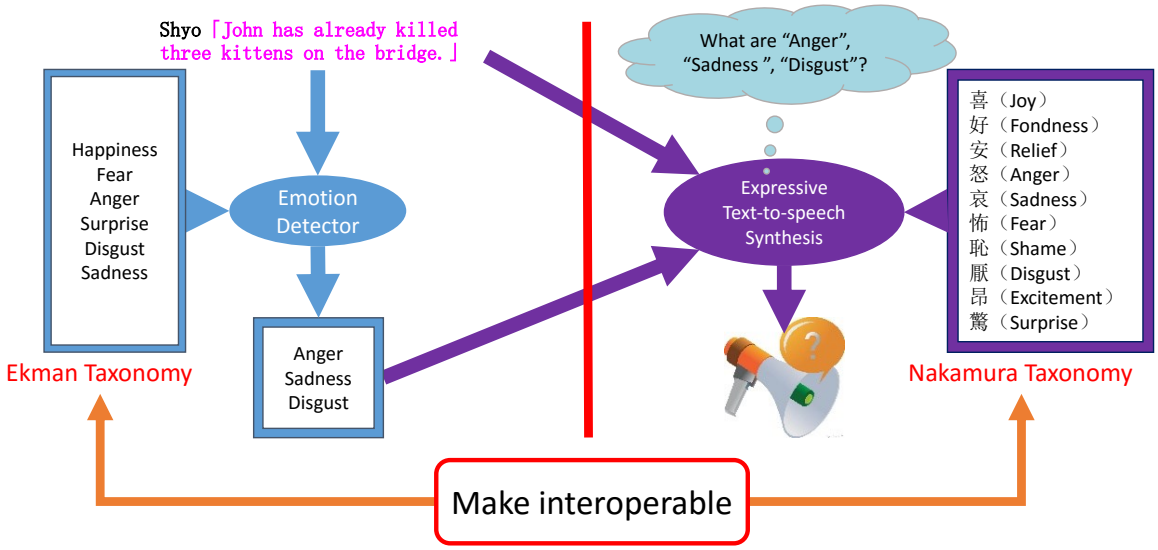


Figure 5-1: Taxonomy mismatch between emotion-oriented systems.

the taxonomy used for the output does not match the taxonomy used for the input.

Although different emotion taxonomies are founded on different psychological theories and fit specific purposes of particular emotion-related research in various fields, 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 emotion 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.

Although this situation occurs frequently, there has been little study of it using a principled statistical approach. Given all this, it is both important and necessary to bridge the gap between taxonomies in the emotion domain.

A taxonomy is constructed for a certain latent semantic space where each possible label set in the taxonomy denotes a unique semantic concept. Darwin [89] observed that "the same state of mind is expressed throughout the world with remarkable

uniformity". In a similar vein, although different emotion taxonomies are proposed on the basis of different psychological 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, so that multi-label affective learning techniques using different taxonomies can become interoperable.

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. (Such as those shown in Figure 3-4, the taxonomy T is listed under the instances. The associated label sets in taxonomy S have already been assigned to the sentences but the annotators cannot see them.). 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 contribution 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). As discussed in the former Section 2.3, ensuring the quality of the responses is one of the biggest challenges in crowdsourcing. This chapter thus focused on how to exploit effective quality control strategy so that semantic matching functions can be accurately established in a crowdsourcing setting with a minimum of human effort.

We developed vector space model-based and probability-based approaches for establishing the semantic matching function. Experimental results on real-world data (emotion annotations for narrative sentences) demonstrated that the proposed cascaded probability-based model can robustly and accurately establish semantic matching functions exhibiting satisfactory performance from the responses provided by a limited number (about five) of crowdsourcing annotators. They also demonstrated that incorporating annotator bias further improves accuracy rates.

5.2 Statistical Models

Problem Formulation

Let I be the set of instances, S be the source taxonomy, and T be the target taxonomy. Two indicator (column) vectors $s = \left(s^{(1)}, s^{(2)}, ..., s^{(|S|)}\right)$ and $t = \left(t^{(1)}, t^{(2)}, ..., t^{(|T|)}\right)$ corresponding to $\mathcal{S} \subseteq S$ and $\mathcal{T} \subseteq T$ are defined as:

$$s^{(\imath)} = \begin{cases} 1, & \text{if: the } \imath\text{-th label} \in \mathcal{S} \\ -1, & \text{if: the } \imath\text{-th label} \notin \mathcal{S} \end{cases} \qquad t^{(\imath)} = \begin{cases} 1, & \text{if: the } \imath\text{-th label} \in \mathcal{T} \\ -1, & \text{if: the } \imath\text{-th label} \notin \mathcal{T} \end{cases} .$$

This means that s (or t) is a binary vector: if a label is present in a label set, the value of the corresponding element in the vector is 1, and -1 otherwise. s_i $(i \in I)$ denotes the associated label vector of instance i in taxonomy S. Let $\mathcal{I} \subset I$ be the set of instances annotated using taxonomy T, K be the set of crowdsourcing annotators, and $\mathcal{K}_i \subseteq K$ $(i \in \mathcal{I})$ be the set of annotators who annotated instance i

using taxonomy T. (Note that it is not necessary to ask every annotator to annotate all the instances.) t_{ik} ($k \in \mathcal{K}_i, i \in \mathcal{I}$) denotes the label vector corresponding to the label set assigned by annotator k, for instance i. Let $E = \{s_i, t_{ik} : k \in \mathcal{K}_i, i \in \mathcal{I}\}$ be the set of obtained examples. The goal is to establish a semantic matching function f: $\{-1, 1\}^{|S|} \rightarrow \{-1, 1\}^{|T|}$ 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 $i \in I$ can be directly detected from s_i without any extra effort.

5.2.1 Vector Space Model (VSM)-Based Approach

The Vector Space Model 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 semantic matching problem, multi-label annotations (label sets) and individual labels can be seen as the counterparts of documents and terms, and the relationships between two taxonomies are formalized as a transformation mapping. By leveraging the VSM, the problem of establishing a semantic matching function f: $\{-1, 1\}^{|S|} \rightarrow \{-1, 1\}^{|T|}$ can be solved by establishing a linear transformation mapping f:

$$t = f(s) = \mathbf{A}s\,, \tag{5.1}$$

where $\mathbf{A}$ is a real-valued $|T| \times |S|$ transformation matrix of mapping f. Using the VSM, the state of each label in target taxonomy T can be seen as a linear combination of the labels in source taxonomy S. We first focus on how to leverage the VSM to establish relationships (transformation mapping) between taxonomies from crowdsourced annotations.

1. Aggregated VSM (A-VSM)

The semantic matching function f can be chose from a hypothesis class F, such that a loss function: $F \times \{-1,1\}^{|S|} \times \{-1,1\}^{|T|} \rightarrow \mathbb{R}$ is minimized. We first 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:

$$\mathbf{A}^{\star} = \arg\min_{\mathbf{A}} \left\{ \sum_{i \in I} \operatorname{dis}\left(\mathbf{A} s_i, \bar{t}_i\right) \right\}, \tag{5.2}$$

where $\operatorname{dis}(\cdot,\cdot)$ denotes the distance between two vectors, $\bar{t}_i$ is the vector corresponding to the aggregated annotation of the annotation set $\{t_{ik} : k \in \mathcal{K}_i\}$. The $\imath$-th ($\imath \in \{1,2,...,|T|\}$) row in matrix $\mathbf{A}$ is the transformation vector from the states (exist or not) of the labels in the source label set s_i to the state of the $\imath$-th emotion in the aggregated emotion set $\bar{t}_i$. It is determined as the least-squares solution to the overdetermined system, where the number of equations is $|I|$ and the number of unknowns is $|S|$.

There are many aggregation strategies proposed by crowdsourcing researchers. To simplify the computation, we adopted the majority vote, the most commonly used strategy. This mean thats $\bar{t}_i$ is the corresponding vector of the most frequently annotated label set among annotations $\{t_{ik} : k \in \mathcal{K}_i\}$.

After constructing the optimal transformation matrix $\mathbf{A}^{\star}$, for any instance with its associated label vector in s in source taxonomy S, we can obtain its associated label vector t in target taxonomy T as

$$t^{(\imath)} \begin{cases} = \quad 1, & \text{if:} \quad \mathbf{A}^{\star} s^{(\imath)} > 0 \\ = -1, & \text{if:} \quad \mathbf{A}^{\star} s^{(\imath)} \leq 0\,, \end{cases}$$

where $\imath \in \{1,2,...,|T|\}$.

2. Combined VSM

Ordinary Combination of VSM (C-VSM)

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, i.e., the aggregated annotation in taxonomy T for each sentence are truncated to one binary-valued vector. We introduce a combination of multiple annotations to establish the mapping between two taxonomies in a crowdsourced setting directly from multiple annotators.

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

$$\mathbf{A}^{\star} = \arg\min_{\mathbf{A}} \left\{ \sum_{i \in I} \sum_{k \in \mathcal{K}_i} \mathrm{dis}\left(\mathbf{A} s_i, t_{ik}\right) \right\}, \tag{5.3}$$

where the $\imath$-th ($\imath \in \{1, 2, ..., |T|\}$) row in matrix $\mathbf{A}$ is the transformation vector from the states of the labels in the source label vector s_i to the state of the $\imath$-th label in the annotation t_{ik}. It is determined as the least-squares solution to the overdetermined system, where the number of equations is $|I| * |K|$ and the number of unknowns is $|S|$.

Weighted Combination of VSM (C-VSM+)

Both Equations (5.2) and (5.3) 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 (Equa-

tion (5.3)) is rewritten as

$$\mathbf{A}^{\star} = \arg\min_{\mathbf{A}} \left\{ \sum_{i \in I} \sum_{k \in \mathcal{K}_i} w_{ik} \cdot \mathrm{dis}\left(\mathbf{A} s_i, t_{ik}\right) \right\} . \tag{5.4}$$

At this time, the $\imath$-th ($\imath \in \{1, 2, ..., |T|\}$) row in matrix $\mathbf{A}$ is determined as the weighted least-squares solution to the overdetermined system, where the number of equations is $|I| * |K|$ and the number of unknowns is $|S|$.

Note that the a weight is specific to both a sentence and an annotator. This means that the mapped target vector ($\mathbf{A}^{\star}\mathbf{x}$) should be nearer to the vectors of 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

$$s_{ik} = \frac{\sum_{\hat{k} \in \mathcal{K}_i, \hat{k} \neq k} \mathrm{sim}\left(t_{ik}, t_{i\hat{k}}\right)}{|\mathcal{K}_i| - 1} ,$$

where the similarity between two indicator vectors is defined as the *Simple Matching Coefficient* defined by Equation (3.15). Finally, the weight of annotator k for sentence i in Equation (5.4) is defined as

$$w_{ik} = \frac{s_{ik}}{\sum_{\hat{k} \in \mathcal{K}_i} s_{i\hat{k}}} .$$

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 label $\imath$ in target taxonomy T is the one with the maximum value in the mapping vector:

$$\imath^{\star} = \arg\max_{\imath \in T} \mathbf{A}^{\star} s^{(\imath)} .$$

5.2.2 Probability-Based Approach

1. 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:

$$f(s) = \underset{t \in \{0,1\}^{|T|}}{\arg\max} \Pr[t \mid s] \ . \tag{5.5}$$

As the causal structure showed in Figure 5-2(b), the naïve solution for establishing the semantic matching function is to estimate the maximum likelihood using annotation frequencies of label vectors:

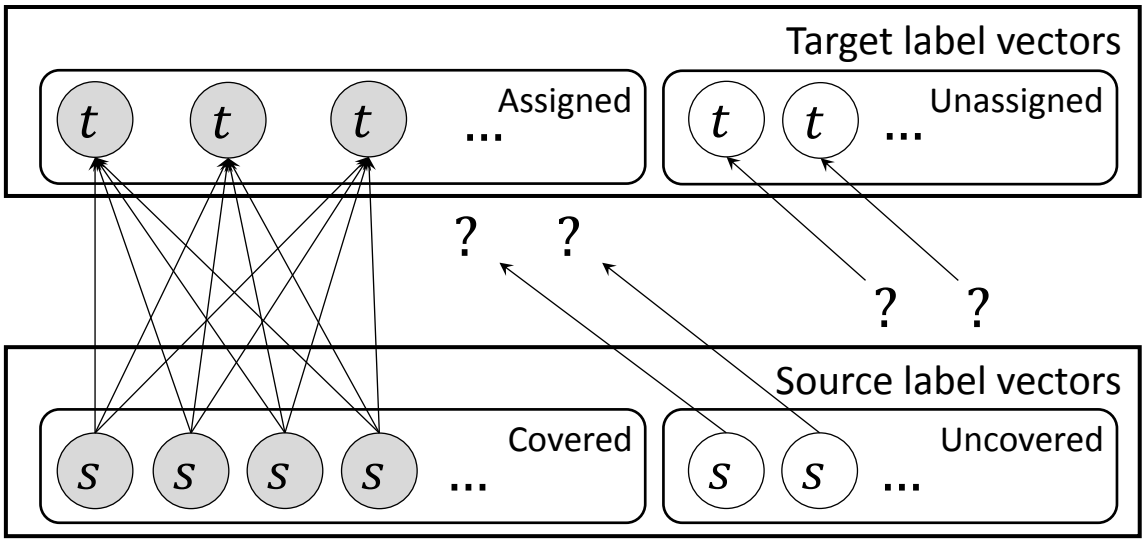
$$\Pr[t \mid s] = \frac{\sum_{i \in I} \sum_{k \in \mathcal{K}_i} [[s_i = s]] \cdot [[t_{ik} = t]]}{\sum_{i \in I} [[s_i = s]] \cdot |\mathcal{K}_i|} \ . \tag{5.6}$$

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

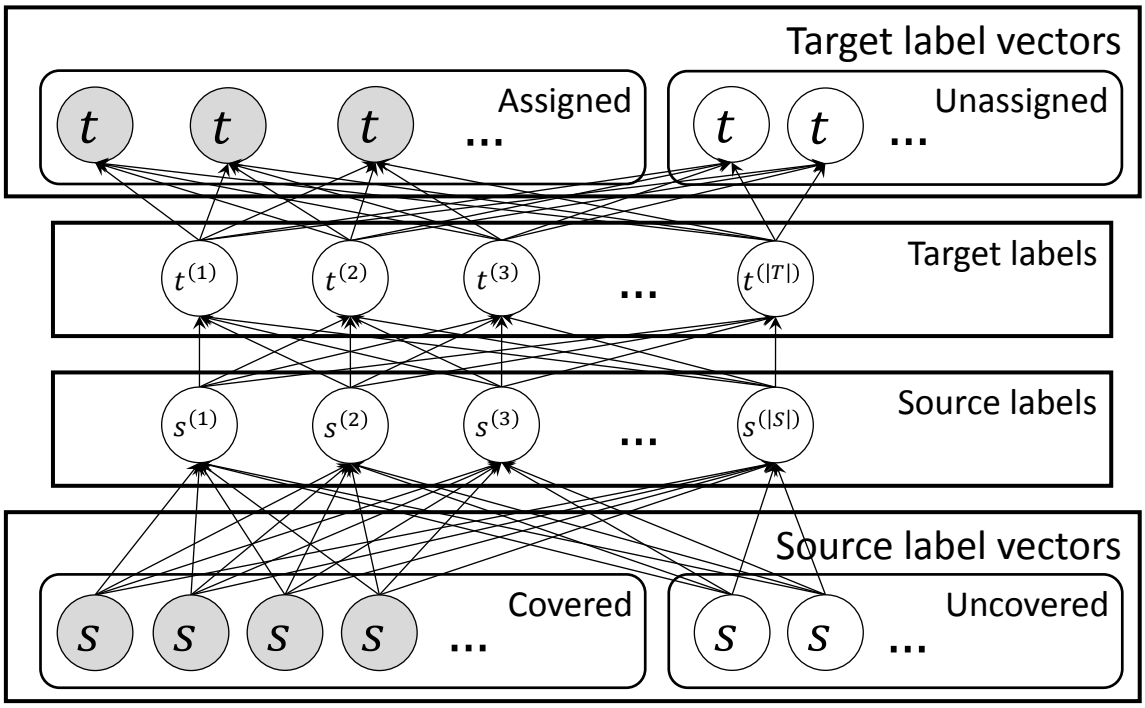
$$[[\mathit{true}]] = 1, \ [[\mathit{false}]] = 0 \, .$$

Pseudo-code for *J-MLE* is given in Algorithm 2 .

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



(a)



(b)

Figure 5-2: Graphical representation of semantic matching: (a) joint maximum likelihood estimation and (b) proposed cascaded method.

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.

2. Cascaded Maximum Likelihood Estimation (C-MLE)

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 5-2(b) shows the causal structure of the proposed method. The proposed method can robustly solve Equation (5.6), 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 (5.6) can be calculated as the product of the independent distributions of the candidate target labels:

$$\Pr\left[t \mid s\right] = \prod_{m=1}^{|T|} \Pr\left[t^{(m)} \mid s\right] . \tag{5.7}$$

Using Bayes' theorem, we have

$$\Pr\left[t^{(m)} \mid s\right] = \frac{\Pr\left[t^{(m)}\right] \cdot \Pr\left[s \mid t^{(m)}\right]}{\Pr\left[s\right]} . \tag{5.8}$$

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:

$$\Pr\left[s\right] = \prod_{n=1}^{|S|} \Pr\left[s^{(n)}\right] , \tag{5.9}$$

$$\Pr\left[s \mid t^{(m)}\right] = \prod_{n=1}^{|S|} \Pr\left[s^{(n)} \mid t^{(m)}\right] . \tag{5.10}$$

Again using Bayes' theorem, we have

$$\Pr\left[s^{(n)}\right] = \frac{\Pr\left[t^{(m)}\right] \cdot \Pr\left[s^{(n)} \mid t^{(m)}\right]}{\Pr\left[t^{(m)} \mid s^{(n)}\right]} . \tag{5.11}$$

Next we substitute Equation (5.11) for $\Pr\left[s^{(n)}\right]$ in Equation (5.9):

$$\begin{aligned} \Pr[s] &= \prod_{n=1}^{|S|} \frac{\Pr\left[t^{(m)}\right] \cdot \Pr\left[s^{(n)} \mid t^{(m)}\right]}{\Pr\left[t^{(m)} \mid s^{(n)}\right]} \\ &= \left[\Pr\left[t^{(m)}\right]\right]^{|S|} \cdot \prod_{n=1}^{|S|} \frac{\Pr\left[s^{(n)} \mid t^{(m)}\right]}{\Pr\left[t^{(m)} \mid s^{(n)}\right]} . \end{aligned} \tag{5.12}$$

By substituting Equations (5.10) and (5.12) into Equation (5.8), we have

$$\Pr\left[t^{(m)} \mid s\right] = \frac{\prod_{n=1}^{|S|} \Pr\left[t^{(m)} \mid s^{(n)}\right]}{\left(\Pr\left[t^{(m)}\right]\right)^{|S|-1}} . \tag{5.13}$$

Finally, substituting Equation (5.13) into Equation (5.7) enables the semantic matching function in Equation (5.5) to be established using

$$f(s) = \underset{t \in \{0,1\}^{|T|}}{\arg\max} \prod_{m=1}^{|T|} \frac{\prod_{n=1}^{|S|} \Pr\left[t^{(m)} \mid s^{(n)}\right]}{\left(\Pr\left[t^{(m)}\right]\right)^{|S|-1}} . \tag{5.14}$$

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

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

the semantic matching function $f(s)$ can be easily established using Equation (5.14). 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 3 .

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 [2]. 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 4 .

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.

5.3 Empirical Study

To test the efficiency of the proposed method, we performed an experimental evaluation in the emotion domain. To collect real-world data, we used two Japanese children's narratives, which were also adopted in Sections 3.5.1 and 4.4, as the texts

Algorithm 2 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

output $t_i = \arg\max \Pr[t_i \mid s_i]$ for each $i \in \mathbf{I}$.

Algorithm 3 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

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

end for

output $t_i = \arg\max \Pr[t_i \mid s_i]$ for each $i \in \mathbf{I}$.

Algorithm 4 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 $a = 0$, $\Pr\left[t^{(m)} \mid s^{(n)}\right]_0 = \frac{\sum_{i\in I}\sum_{k\in\mathcal{K}_i}[[t_{ik}^{(m)}=t^{(m)}]]\cdot[[s_i^{(n)}=s^{(n)}]]}{\sum_{i\in I}\sum_{k\in\mathcal{K}_i}[[t_{ik}^{(m)}=t^{(m)}]]}$.

while the converge condition of the Dawid-Skene model is not satisfied **do**

- compute $\Pr\left[t^{(m)} \mid s^{(n)}\right]_{a+1}$, $\Pr\left[t^{(m)}\right]_{a+1}$ using Dawid-Skene model with $\Pr\left[t^{(m)} \mid s^{(n)}\right]_a$.
- $a = a + 1$.

end while

$\Pr\left[t^{(m)} \mid s^{(n)}\right] = \Pr\left[t^{(m)} \mid s^{(n)}\right]_{a+1}$.

$\Pr\left(t^{(m)}\right) = \Pr\left[t^{(m)}\right]_{a+1}$.

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

output $t_i = \arg\max \Pr[t_i \mid s_i]$ for each $i \in \mathbf{I}$.

to be annotated.

We used two typical emotion taxonomies as the source taxonomy and the target taxonomy. One was Ekman's taxonomy (including six labels), which is the most commonly used emotion taxonomy in emotion-related research. The other was Nakamura's taxonomy (including ten labels), which is used in former experiments. To enable *mutual validation* to be performed between the two taxonomies (Ekman→Nakamura, Nakamura→Ekman), both of them were used to annotate the sentences in the two narratives. An example task input screen with Nakamura's taxonomy shown below the sentences is shown in Figure 3-4. Ekman's taxonomy was presented in its original English form with Japanese explanations. We still conducted the experiments using the Lancers crowdsourcing service. The two taxonomies were presented separately to arbitrary annotators, and few, if any, of them annotated sentences with both the taxonomies.

Similar to the experimental strategy in Sections 3.5.1 and 4.4, 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. The 30 annotators who annotated a particular sentence using the target taxonomy into 3 (ten groups), 5 (six groups), 10 (three groups), 15 (two groups), and 30 (one group), in order to determine the effect of the number of annotators on accuracy. Every group of annotators were used to generate the matching function respectively. The accuracy rate of a certain group size is measured as the average accuracy rate of the functions generated by all groups in the group size. The labels in the two taxonomies are shown in Tables 5.1 and 3.1 with their annotation frequencies. Other statistics for the experiments are shown in Table 5.2.[1]

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

[1]In fact, the annotations in the Nakamura's taxonomy are the same as the annotations used in Sections 3.5.1 and 4.4.

Table 5.1: Annotation frequencies of emotion labels 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 5.2: 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 annotation for Ekman/Nakamura	1890	2340	4230
Avg. no. of annotations per sentence for Ekman/Nakamura	30	30	30
Avg. no. of assigned labels per annotation for Ekman	1.05	1.27	1.17
Avg. no. of assigned labels per annotation for Nakamura	1.75	1.18	1.43

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 six models:

- *A-VSM*: **A**ggregated **V**ector **S**pace **M**odel;
- *C-VSM*: Ordinarily **C**ombined **V**ector **S**pace **M**odel;
- *C-VSM+*: Weighted **C**ombined **V**ector **S**pace **M**odel;
- *J-MLE*: **J**oint **M**aximum **L**ikelihood **E**stimation;
- *C-MLE*: **C**ascaded **M**aximum **L**ikelihood **E**stimation;
- *C-MLE+*: **C**ascaded **M**aximum **L**ikelihood **E**stimation with Dawid-Skene model.

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 4-1(a)), we forced those uncovered label sets to map to *Neutral*. We used the *Simple Matching Coefficient* (defined by Equation (3.15), also used in Sections 3.5.1 and 4.4) to evaluate the performance of the semantic matching function, i.e., the average proportion of state-consistent labels between the predicted target label set and the aggregated gold-standard target label set over all sentences.

The experiment results are shown in Tables 5.3, 5.4, 5.5, 5.6, and Figures 5-3, 5-4, 5-5, 5-4. No matter which of the two narratives was used to establish the semantic matching function, the results tends to maintain the same tendency. Accuracies of all the six models increased with the group size. For the three *VSM*-based models, both the *A-VSM* and *C-VSM+* models achieved better accuracies than the *C-VSM* model. This means that using unprocessed crowdsourced annotations to establish the function directly is ineffective since crowdsourcing annotators are not so reliable. The *C-VSM+* model performed better than the *A-VSM* model. This demonstrates that crowdsourcing annotators have variable levels of expertise. If proper quality control

techniques, such as weighting annotators' expertise in *C-VSM+* model, are adopted to treat the variable-quality crowdsourced annotations, *combination* strategy could achieve better accuracy rates rather than simply use the *Majority Vote* aggregation strategy. For the three *MLE* models, the (*C-MLE* and *C-MLE+*) models with the cascaded method consistently outperformed the *J-MLE* model. This means that estimation using the proposed cascaded method is more accurate than ordinary maximum likelihood estimation. The *J-MLE* model performed worst among all the six models due to that, in the *J-MLE* model, as a compromised solution to the *zero-shot* problem, we forced the uncovered label sets to map to *Neutral* Moreover, in most cases, the *C-MLE+* model achieved better accuracies than the *C-MLE* model. This demonstrates that annotator bias should be considered in variable-quality crowdsourced annotations. Among all models, *C-MLE+* performed best and had accuracies greater than 90% for five or more annotators per instance. This means 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 (Table 5.3, Figure 5-3) or comparable to (Table 5.5, Figure 5-5) 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.

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. Even the most complicated model, the *C-MLE+* model, could converge in less than 3 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 sentences, the number of annotators, and the number of candidate labels, while the used dataset was not so large. All models were non-iterative except the *C-MLE+* model, where the iterative Dawid-Skene model is used to consider annotator's bias on each label in the target taxonomy separately.

Table 5.3: Average matching accuracies for function "Ekman → Nakamura", Training narrative: "Although We Are in Love".

Model	**No. of annotators per sentence**				
	3	**5**	**10**	**15**	**30**
A-VSM	0.8582	0.8691	0.8789	0.8863	0.8984
C-VSM	0.8479	0.8615	0.8761	0.8777	0.8841
C-VSM+	0.8642	0.8717	0.8779	0.8903	0.9015
J-MLE	0.8421	0.8547	0.8654	0.8659	0.8929
C-MLE	0.8865	0.8867	0.8903	0.8943	0.8957
C-MLE+	0.8861	0.9060	0.9087	0.9195	0.9205

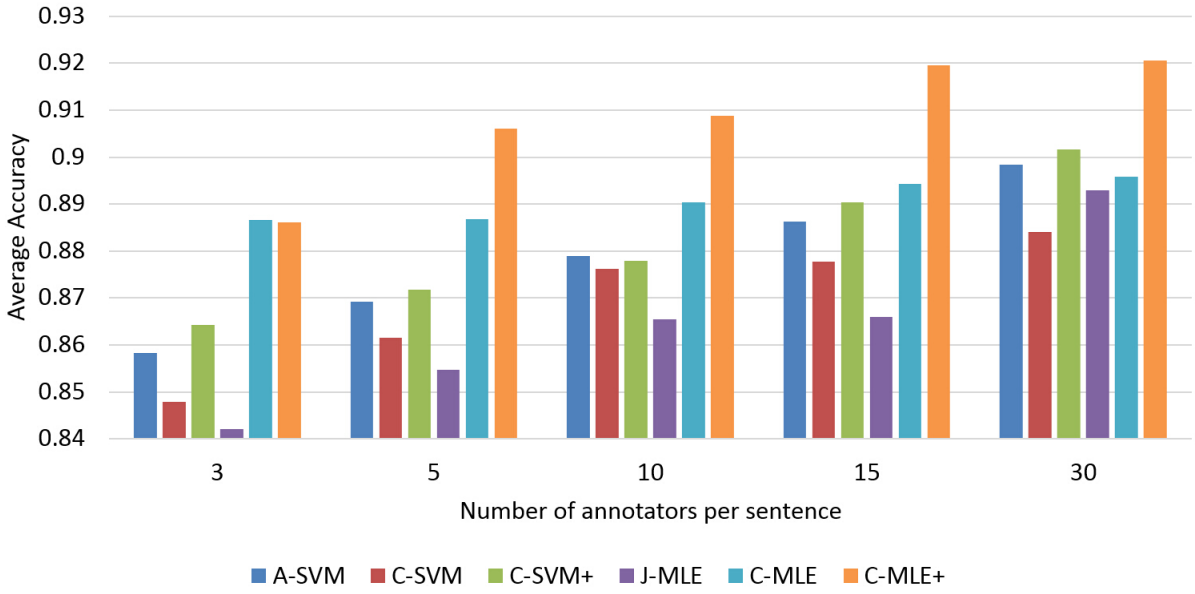


Figure 5-3: Average matching accuracies for function "Ekman → Nakamura", Training narrative: "Although We Are in Love".

Table 5.4: Average matching accuracies for function "Nakamura → Ekman", Training narrative: "Although We Are in Love".

Model	**No. of annotators per sentence**				
	3	**5**	**10**	**15**	**30**
A-VSM	0.8676	0.8776	0.8936	0.8968	0.9063
C-VSM	0.8655	0.8733	0.8821	0.8860	0.8947
C-VSM+	0.8703	0.8778	0.8937	0.8995	0.9079
J-MLE	0.8644	0.8672	0.8747	0.8682	0.9030
C-MLE	0.8734	0.8682	0.8798	0.8770	0.9184
C-MLE+	0.8893	0.9066	0.9137	0.9219	0.9267

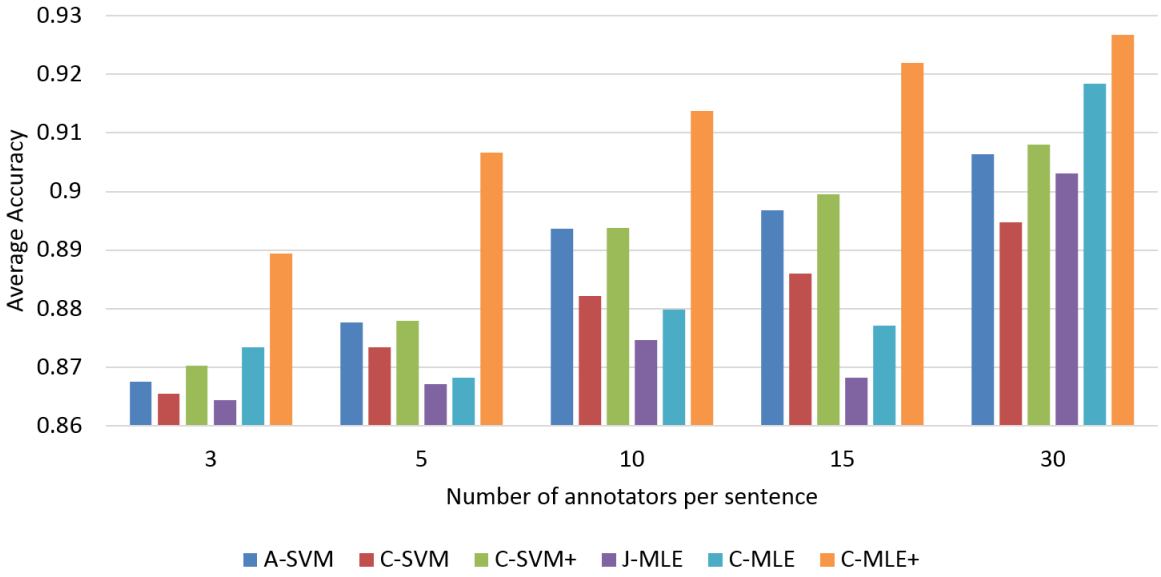


Figure 5-4: Average matching accuracies for function "Nakamura → Ekman", Training narrative: "Although We Are in Love".

Table 5.5: Average matching accuracies for function "Ekman → Nakamura", Training narrative: "Little Masa and A Red Apple".

Model	**No. of annotators per sentence**				
	3	**5**	**10**	**15**	**30**
A-VSM	0.8754	0.8759	0.8857	0.8943	0.8953
C-VSM	0.8708	0.8734	0.8781	0.8783	0.8941
C-VSM+	0.8773	0.8774	0.8909	0.8950	0.9005
J-MLE	0.8698	0.8729	0.8837	0.8918	0.9113
C-MLE	0.8840	0.8825	0.8912	0.8936	0.9125
C-MLE+	0.8945	0.9034	0.9062	0.9125	0.9125

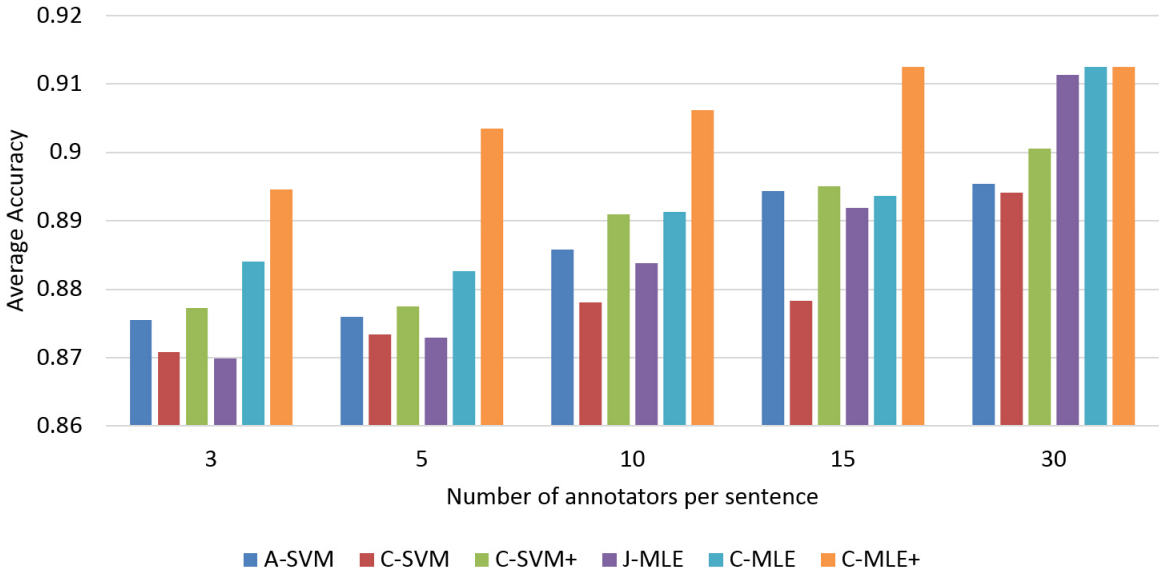


Figure 5-5: Average matching accuracies for function "Ekman → Nakamura", Training narrative: "Little Masa and A Red Apple".

Table 5.6: Average matching accuracies for function "Nakamura → Ekman", Training narrative: "Little Masa and A Red Apple".

Model	**No. of annotators per sentence**				
	3	**5**	**10**	**15**	**30**
A-VSM	0.8849	0.8938	0.9047	0.9179	0.9195
C-VSM	0.8825	0.8870	0.9026	0.9042	0.9079
C-VSM+	0.8911	0.9000	0.9064	0.9166	0.9211
J-MLE	0.8784	0.8834	0.8990	0.9035	0.9092
C-MLE	0.8939	0.8960	0.8983	0.9053	0.9134
C-MLE+	0.9136	0.9213	0.9314	0.9397	0.9397

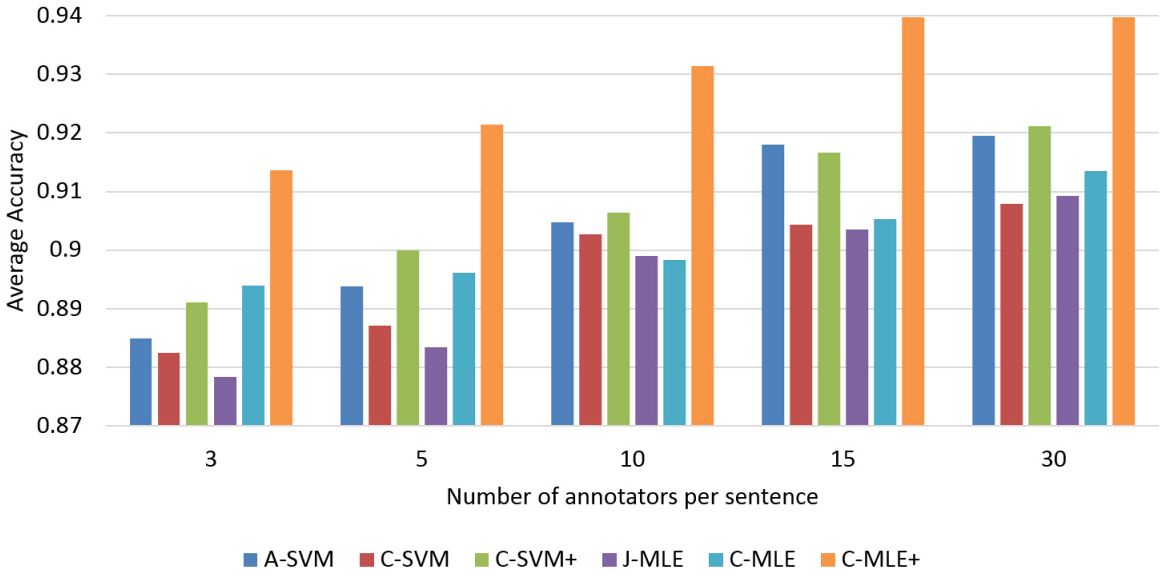


Figure 5-6: Average matching accuracies for function "Nakamura → Ekman", Training narrative: "Little Masa and A Red Apple".

5.4 Summary

Most multi-label domains lack an authoritative taxonomy. Therefore, different taxonomies are commonly used in the same domain, 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. This often results in complications. This chapter especially focused on leveraging crowdsourcing to establish relationships between different emotion taxonomies. VSM-based and Probability-based approaches are proposed for establishing a semantic matching function in a crowdsourcing setting that maps label sets in the source taxonomy to label sets in the target taxonomy, in terms of the semantic distances between them in the latent semantic space. To cover all possible label sets in both the two taxonomies, we extended the probability-based approach by transferring semantic concepts from the source taxonomy to the target taxonomy in a cascaded way. This can be useful for addressing the problems that occur when different taxonomies are used in the emotion domain.

Different from multi-label "classification" learning, the proposal 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. We conducted the experiment on real-world crowdsourcing data with two typical emotion taxonomies and the sentences in two narratives. Experimental results demonstrated that the semantic matching function established using the proposed models enabled the gold standard labels in the target taxonomy for an instance to be effectively estimated, directly from its associated labels in the source taxonomy without any extra effort.

The results of this research provided several benefits. For example, (1) multi-label affective applications can use multi-label affect detectors that have already been vetted, (2) the detector that comes with a large annotated corpus can be used to

train other detectors or simply supplement the original dataset if allowed by the usage restrictions on the corpus, and (3) the performance of affective learning techniques using different taxonomies can be uniformly benchmarked. However, note that the proposed method provided 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.

Although the proposed models are tested using two emotion taxonomies, the same problem can be found in other multi-label domains, such as film genre classification (using the list of genres from IMDB[2] or Netflix[3]) and text categorization (using Reuters Topics[4] or the Mozilla Directory[5]). We plan to extend our research across different multi-label domains. It is conceivable that semantic matching will be more difficult if the number of source labels is much smaller than that of target labels, due to the difference in the granularity of information. We also plan to extend our experiments for such cases to see whether the proposed models 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.

[2]http://www.imdb.com/genre
[3]http://dvd.netflix.com/AllGenresList
[4]http://vocab.org/reuters_topics/1.0/
[5]http://www.dmoz.org/

Chapter 6

Conclusion & Future work

6.1 Conclusion

Traditional classification learning deals with single-label data, while we focused on the problem of multi-label learning. Especially, we focused on multi-label affect annotations, where each instance can be associated with a combination of multiple affect labels. Such annotations is crucial for multi-label affective learning techniques. The annotation quality directly affects the performance of those techniques. Obtaining high-quality annotations from both experts and large crowds can be expensive and time-consuming. We thus investigated ways to obtain at low cost high-quality multi-label affect annotations for use with multi-label affective learning techniques.

Though crowdsourcing services allow us to get access to a large amount of annotations at very low cost (time and expense), the annotation quality varies from crowd annotators to crowd annotators. Some annotators are highly skilled, while some are not. Even as the skilled annotators, they can still provides invalid annotations, due to distraction or fatigue. Therefore, it is both important and necessary to find effective ways to extract valid annotations.

A common quality control strategy for crowdsourced labeling tasks is to aggregate the responses provided by multiple annotators to produce one reliable annotation. Experimental results on real-world crowdsourcing data showed that The proposed aggregation models enable associated affect labels to be effectively estimated, using

the annotations provided by a handful of crowdsourcing annotators. To sum up, our study provided a promising way to reduce the cost of affect data collection for future applications with minimal degradation in the quality of the results. We envision that by leveraging proper statistical quality control strategies, a crowdsourcing setting could be a good candidate to the problem of insufficient annotation data in multi-label affective learning.

The primary contribution of this dissertation is threefold:

1. For instances without any associated labels, we proposed flexibly incorporating affect label dependency into the label-generation process, making the estimation of associated labels for each instance more accurate.

2. For instances implicitly contain information for "consistency" and "context", we proposed considering internal relationships among them. This made estimation accuracy rates further improved.

3. For instances with associated labels selected from an undesired taxonomy, we proposed a novel and effective approach 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.

As a summary, we show the relationship between the contributions of this dissertation in Figure 6-1.

6.2 Future work

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.

1. Our experiments were conducted on small databases, especially two children's narratives using two affect taxonomies. We plan to explore whether the proposed models are also accurate for larger datasets.

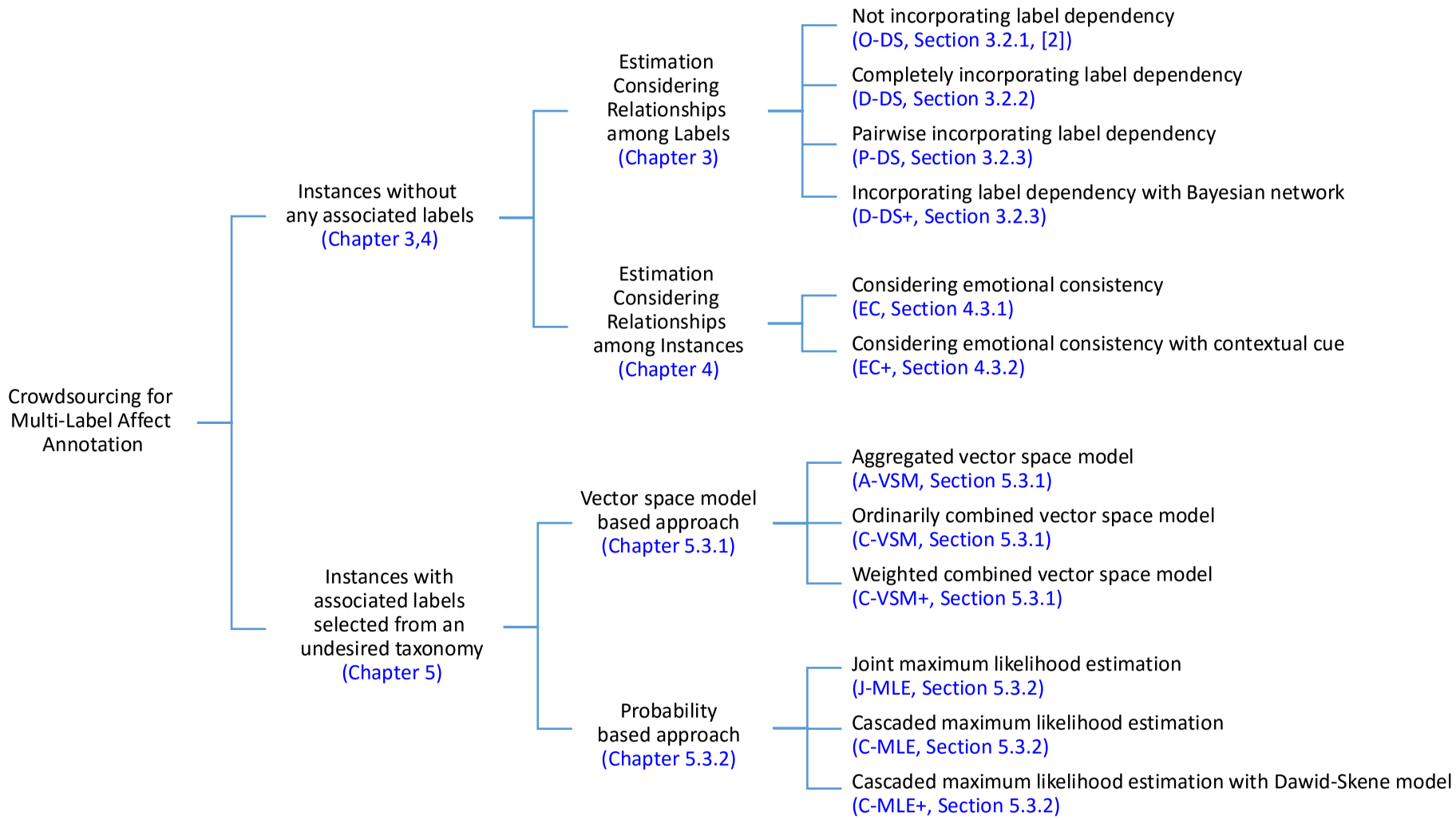


Figure 6-1: The relationship between the contributions of this dissertation.

2. In our research, 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 the difficulties of instances could further reduce annotation costs. 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.

3. Other information, such as annotators' self-reported confidence scores, which have shown an improvement recently [5], is also important for the label-generation process and worth studying.

4. In addition to affective learning, there are lots of aspects in multi-label learning, as discussed in Section 2.2. We will test effectiveness of the proposed models across different domains.

In the long-term, we plan to extend our work to other human computation issues, such as social network analysis, music comprehension, film genre classification and art style recognition, which are thought to raise more variant opinions among people. Estimating labels not only from crowdsourced annotations but also from user-created content services is also an interesting direction for future work.

Appendix A

Proof of optimal label pairing pattern

Let $P(X)$ be the joint probability distribution over n labels x_1, x_2, ..., x_n, X denoting the n-dimensional vector $(x_1, x_2, ..., x_n)$. Under the condition that labels are pairwise correlated, the joint distribution over all labels takes the following form:

$$P'(X) = \prod_{\substack{i=1 \\ i<j(i)\leq n \\ i\neq j(i'),j(i)\neq j(i')\,(i'=1,\ldots,i-1)}}^{n-1} P\left(x_i, x_{j(i)}\right),$$

where $\left(x_i, x_{j(i)}\right)$ constitutes a label pair. The optimal pairing pattern is the one that minimizes the Kullback-Leibler divergence [90], which measures the difference between two probability distributions over the same event space, between $P(X)$ and $P'(X)$:

$$\begin{aligned} D(P \mid\mid P') &= \sum_X P(X) \log \frac{P(X)}{P'(X)} \\ &= -\sum_X P(X) \log \frac{1}{P(X)} - \sum_X P(X) \log P'(X) \\ &= -H(X) - \sum_X P(X) \sum_{\substack{i=1 \\ i<j(i)\leq n \\ i\neq j(i'),j(i)\neq j(i')\,(i'=1,\ldots,i-1)}}^{n-1} \log P\left(x_i, x_{j(i)}\right), \end{aligned} \tag{A.1}$$

where $H(X)$ on the right side is the joint entropy of all labels. Since $P\left(x_i, x_{j(i)}\right)$ is a component of $P(X)$,

$$-\sum_{X} P(X) \log P\left(x_i, x_{j(i)}\right) = -\sum_{x_i, x_{j(i)}} P\left(x_i, x_{j(i)}\right) \log P\left(x_i, x_{j(i)}\right),$$

where the right side of the equation is the joint entropy of label pair $\left(x_i, x_{j(i)}\right)$. Thus, Equation (A.1) becomes

$$D(P \mid\mid P') = -H(X) + \sum_{\substack{i=1 \\ i<j(i)\leq n \\ i\neq j(i'), j(i)\neq j(i') \, (i'=1,\ldots,i-1)}}^{n-1} H\left(x_i, x_{j(i)}\right).$$

Since $H(X)$ is independent of the pairing pattern, minimizing the Kullback-Leibler divergence $D(P \mid\mid P')$ is equivalent to minimizing the sum of the joint entropies of all label pairs:

$$\sum_{\substack{i=1 \\ i<j(i)\leq n \\ i\neq j(i'), j(i)\neq j(i') \, (i'=1,\ldots,i-1)}}^{n-1} H\left(x_i, x_{j(i)}\right).$$

If we depict the pairing pattern as an undirected graph, where labels are represented by vertices, and the weight of each edge is assigned the joint entropy of the two labels represented by the two vertices of the edge, the sum of the joint entropies of all label pairs can be minimized by finding the minimum-weight perfect matching of the graph. This solution can be achieved by using the Blossom algorithm [91].

Bibliography

[1] Seyda Ertekin, Cynthia Rudin, and Haym Hirsh. Approximating the crowd.

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

[3] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier R Movellan, and Paul 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.

[4] Peter Welinder, Steve Branson, Serge Belongie, and Pietro Perona. The multi-dimensional wisdom of crowds. In *Advances in Neural Information Processing Systems (NIPS)*, volume 10, pages 2424–2432, 2010.

[5] Satoshi Oyama, Yukino Baba, Yuko Sakurai, and Hisashi 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.

[6] Brian Massumi. Movement, affect, sensation, 2002.

[7] Eric Shouse. Feeling, emotion, affect. *M/c journal*, 8(6):26, 2005.

[8] Niko Besnier. Language and affect. *Annual review of anthropology*, pages 419–451, 1990.

[9] Kristie S Fleckenstein. Defining affect in relation to cognition: A response to susan mcleod. *Journal of Advanced Composition*, pages 447–453, 1991.

[10] Myriam Munezero, Calkin Suero Montero, Erkki Sutinen, and John Pajunen. Are they different? affect, feeling, emotion, sentiment, and opinion detection in text. *Affective Computing, IEEE Transactions on*, 5(2):101–111, 2014.

[11] Cecilia Ovesdotter Alm, Dan Roth, and Richard 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.

[12] 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.

[13] Jin Ha Lee and Xiao Hu. Generating ground truth for music mood classification using mechanical turk. In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*, pages 129–138. ACM, 2012.

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

[15] Robert Anderson, Bjorn Stenger, Vincent Wan, and Roberto Cipolla. Expressive visual text-to-speech using active appearance models. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3382–3389. IEEE, 2013.

[16] Mostafa Al Masum Shaikh, Antonio Rui Ferreira Rebordao, and Keikichi 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.

[17] Maria do Rosário Dias Dias, Sofia de Sacadura Botte Lynce de Faria, Soraia Cristina Momade Ibrahim, et al. Im like a river: a health education instrument for stuttering. *Revista de Psicologia da IMED*, 5(2), 2013.

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

[19] 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.

[20] Cecilia Ovesdotter 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.

[21] Robert E Schapire and Yoram Singer. Boostexter: A boosting-based system for text categorization. *Machine learning*, 39(2):135–168, 2000.

[22] Naonori Ueda and Kazumi Saito. Parametric mixture models for multi-labeled text. In *Advances in neural information processing systems*, pages 721–728, 2002.

[23] Shenghuo Zhu, Xiang Ji, Wei Xu, and Yihong Gong. Multi-labelled classification using maximum entropy method. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 274–281. ACM, 2005.

[24] Seungji Yang, Sang-Kyun Kim, and Yong Man Ro. Semantic home photo categorization. *Circuits and Systems for Video Technology, IEEE Transactions on*, 17(3):324–335, 2007.

[25] Guo-Jun Qi, Xian-Sheng Hua, Yong Rui, Jinhui Tang, Tao Mei, and Hong-Jiang Zhang. Correlative multi-label video annotation. In *Proceedings of the 15th international conference on Multimedia*, pages 17–26. ACM, 2007.

[26] 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.

[27] 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.

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

[29] Shantanu Godbole and Sunita Sarawagi. Discriminative methods for multi-labeled classification. In *Advances in Knowledge Discovery and Data Mining*, pages 22–30. Springer, 2004.

[30] Juho Rousu, Craig Saunders, Sandor Szedmak, and John Shawe-Taylor. Kernel-based learning of hierarchical multilabel classification models. *The Journal of Machine Learning Research*, 7:1601–1626, 2006.

[31] Eneldo Loza Mencia and Johannes Fürnkranz. Efficient pairwise multilabel classification for large-scale problems in the legal domain. In *Machine Learning and Knowledge Discovery in Databases*, pages 50–65. Springer, 2008.

[32] Robert Moskovitch, Shiva Cohen-Kashi, Uzi Dror, Iftah Levy, Amit Maimon, and Yuval Shahar. Multiple hierarchical classification of free-text clinical guidelines. *Artificial Intelligence in Medicine*, 37(3):177–190, 2006.

[33] John P Pestian, Christopher Brew, Paweł Matykiewicz, DJ Hovermale, Neil Johnson, K Bretonnel Cohen, and Włodzisław Duch. A shared task involving multi-label classification of clinical free text. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 97–104. Association for Computational Linguistics, 2007.

[34] Nadia Ghamrawi and Andrew McCallum. Collective multi-label classification. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 195–200. ACM, 2005.

[35] Adriano Veloso, Wagner Meira Jr, Marcos André Gonçalves, and Mohammed Javeed Zaki. Multi-label lazy associative classification. In *PKDD*, volume 4702, pages 605–612. Springer, 2007.

[36] Ioannis Katakis, Grigorios Tsoumakas, and Ioannis Vlahavas. Multilabel text classification for automated tag suggestion. *ECML PKDD discovery challenge*, 75, 2008.

[37] Gemma Boleda, Sabine Schulte im Walde, and Toni Badia. Modelling polysemy in adjective classes by multi-label classification. In *EMNLP-CoNLL*, pages 171–180, 2007.

[38] Matthew R Boutell, Jiebo Luo, Xipeng Shen, and Christopher M Brown. Learning multi-label scene classification. *Pattern recognition*, 37(9):1757–1771, 2004.

[39] Min-Ling Zhang and Zhi-Hua Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition*, 40(7):2038–2048, 2007.

[40] Andreas P Streich and Joachim M Buhmann. Classification of multi-labeled data: A generative approach. In *Machine learning and knowledge discovery in databases*, pages 390–405. Springer, 2008.

[41] Tao Li and Mitsunori Ogihara. Detecting emotion in music. In *ISMIR*, volume 3, pages 239–240, 2003.

[42] Amanda Clare and Ross D King. Knowledge discovery in multi-label phenotype data. In *Principles of data mining and knowledge discovery*, pages 42–53. Springer, 2001.

[43] André Elisseeff and Jason Weston. A kernel method for multi-labelled classification. In *Advances in neural information processing systems*, pages 681–687, 2001.

[44] Hendrik Blockeel, Leander Schietgat, Jan Struyf, Sašo Džeroski, and Amanda Clare. *Decision trees for hierarchical multilabel classification: A case study in functional genomics*. Springer, 2006.

[45] Yi Zhang, Samuel Burer, and W Nick Street. Ensemble pruning via semi-definite programming. *The Journal of Machine Learning Research*, 7:1315–1338, 2006.

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

[47] Luis Von Ahn. Human computation. In *Design Automation Conference, 2009. DAC'09. 46th ACM/IEEE*, pages 418–419. IEEE, 2009.

[48] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew 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.

[49] Alexander Sorokin and David Forsyth. Utility data annotation with amazon mechanical turk. *Urbana*, 51(61):820, 2008.

[50] Christopher H Lin, Mausam, and Daniel 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.

[51] Hyun Joon Jung and Matthew Lease. Improving quality of crowdsourced labels via probabilistic matrix factorization. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[52] Yukino Baba, Hisashi Kashima, Kei Kinoshita, Goushi Yamaguchi, and Yosuke Akiyoshi. Leveraging non-expert crowdsourcing workers for improper task detection in crowdsourcing marketplaces. *Expert Systems with Applications*, 41(6):2678–2687, 2014.

[53] Andrew Mao, Ariel D Procaccia, and Yiling Chen. Better human computation through principled voting. In *AAAI*, 2013.

[54] Aashish Sheshadri and Matthew Lease. Square: A benchmark for research on computing crowd consensus. In *First AAAI Conference on Human Computation and Crowdsourcing*, 2013.

[55] Stefanie Nowak and Stefan 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.

[56] Jonathan Bragg, Daniel S Weld, et al. Crowdsourcing multi-label classification for taxonomy creation. In *First AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 2013.

[57] Victor S Sheng, Foster Provost, and Panagiotis 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.

[58] Pinar Donmez, Jaime G Carbonell, and Jeff 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.

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

[60] Akira Nakamura. Kanjo hyogen jiten [Dictionary of Emotive Expressions] (in Japanese). *Tokyodo*, 1993.

[61] Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. Affect analysis model: novel rule-based approach to affect sensing from text. *Natural Language Engineering*, 17(01):95–135, 2011.

[62] Saif Mohammad. From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 105–114. Association for Computational Linguistics, 2011.

[63] Weiyuan Li and Hua Xu. Text-based emotion classification using emotion cause extraction. *Expert Systems with Applications*, 41(4):1742–1749, 2014.

[64] Kerstin Bischoff, Claudiu S Firan, Wolfgang Nejdl, and Raluca 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 (JCDL)*, pages 285–294. ACM, 2009.

[65] Taner Danisman and Adil Alpkocak. Feeler: Emotion classification of text using vector space model. In *AISB 2008 Convention Communication, Interaction and Social Intelligence*, volume 1, page 53, 2008.

[66] Zhilei Liu, Shangfei Wang, Zhaoyu Wang, and Qiang 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.

[67] 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.

[68] Catherine Pelachaud. *Emotion-oriented systems*. John Wiley & Sons, 2013.

[69] Ryong Lee and Kazutoshi Sumiya. Measuring geographical regularities of crowd behaviors for twitter-based geo-social event detection. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*, pages 1–10. ACM, 2010.

[70] Yang Liao, Masud Moshtaghi, Bo Han, Shanika Karunasekera, Ramamohanarao Kotagiri, Timothy Baldwin, Aaron Harwood, and Philippa Pattison. Mining micro-blogs: opportunities and challenges. In *Computational Social Networks*, pages 129–159. Springer, 2012.

[71] Kaiquan Xu, Jiexun Li, and Yuxia Song. Identifying valuable customers on social networking sites for profit maximization. *Expert Systems with Applications*, 39(17):13009–13018, 2012.

[72] Fausto Giunchiglia, Pavel Shvaiko, and Mikalai Yatskevich. Semantic matching. In *Encyclopedia of Database Systems*, pages 2561–2566. Springer, 2009.

[73] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[74] Fausto Giunchiglia, Pavel Shvaiko, and Mikalai Yatskevich. S-match: an algorithm and an implementation of semantic matching. In *ESWS*, volume 3053, pages 61–75. Springer, 2004.

[75] Fausto Giunchiglia, Maurizio Marchese, and Ilya Zaihrayeu. Encoding classifications as lightweight ontologies. 2006.

[76] Souleiman Hasan, Sean O'Riain, and Edward Curry. Approximate semantic matching of heterogeneous events. In *Proceedings of the 6th ACM International Conference on Distributed Event-Based Systems*, pages 252–263. ACM, 2012.

[77] Pavel Shvaiko and Jérôme Euzenat. A survey of schema-based matching approaches. In *Journal on Data Semantics IV*, pages 146–171. Springer, 2005.

[78] Charles K Ogden and Ivor A Richards. The meaning of meaning. *New York*, 1923.

[79] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[80] George James Lidstone. Note on the general case of the bayes-laplace formula for inductive or a posteriori probabilities. *Transactions of the Faculty of Actuaries*, 8(182-192):13, 1920.

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

[82] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*, volume 81. The MIT Press, 2000.

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

[84] Lisa Feldman Barrett, Batja Mesquita, Kevin N Ochsner, and James J Gross. The experience of emotion. *Annual Review of Psychology*, 58:373, 2007.

[85] Hillel Aviezer, Ran R Hassin, Jennifer Ryan, Cheryl Grady, Josh Susskind, Adam Anderson, Morris Moscovitch, and Shlomo Bentin. Angry, disgusted, or afraid? studies on the malleability of emotion perception. *Psychological Science*, 19(7):724–732, 2008.

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

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

[88] Marc 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.

[89] Charles Darwin, Paul Ekman, and Phillip Prodger. *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[90] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[91] Jack Edmonds. Paths, trees, and flowers. *Canadian Journal of Mathematics*, 17(3):449–467, 1965.