1J3-05

Multi-affect Estimation Considering Consistencies among Crowdsourced Annotations

Lei Duan Satoshi Oyama Haruhiko Sato Masahito Kurihara

Graduate School of Information Science and Technology, Hokkaido University

We propose taking consistencies of story emotionality and character personality among crowdsourced annotations into account to estimate multiple affect labels for narrative sentences. Experimental results show that our approach enable the general consensus among large crowds to be effectively estimated using the opinions of a handful of crowdsourcing annotators. This will reduce the cost of preparing training data for use with narrative-oriented affect prediction techniques with minimal degradation in the quality of the result.

1. Introduction

Several machine learning techniques in the field of artificial intelligence are aimed at simulating emotion comprehension. Given the complexity of human thinking, these techniques have one thing in common: they more naturally fit the paradigm of multiple-label prediction than that of single-label prediction from a finite outcome space of affect labels. Moreover, the quality of training data directly influences the success or failure of these techniques. Take narrative-oriented affect prediction as an example. People have different tendencies in detecting subjective affect feelings, so they may experience the same narrative sentence differently. This means that the high quality training data for use with supervised affect prediction techniques should be in accord with the general consensus among large crowds. However, collecting data from large crowds is almost impossible due to the extremely high cost in time and expense.

Crowdsourcing is an economical and efficient approach to performing tasks that are difficult for computers but easy for humans, and labeling is one of its main applications. Obtaining crowdsourced annotations is a promising way of collecting training data for comprehension-simulation techniques. The *majority vote* most objectively reflects the general consensus if the number of voters is large enough. It is based on the implicit assumption that all voters have the same probability of making an error. On the other hand, crowdsourcing annotators are rarely trained and generally do not have the abilities needed to accurately perform the task. Moreover, some of them may simply give random responses as a mean to earn easy money. Therefore, if the number of collected annotations is less than a certain unknown number, the detrimental effect of the noisy responses will be significant, and treating the responses given by different annotators equally will produce poor results.

Dawid and Skene [Dawid 79] have proposed a model that considered annotators' predilections for certain labels. Furthermore, narrative, as a genre of literature characterized by its descriptive force, almost always subject to some affect tendencies. In particular, considering the elementary level of children's psychological development, children's stories and fairy tales often have vibrant affection tint and distinct character personalities so that children's attention can be certainly attracted. For this reason, we focus on affect consistencies of story emotionality and character personality among crowdsourced annotations. We attempt to incorporate the consistencies to our affect-inference process. The aim is to best estimate multiple true affect labels, which reflect the general consensus among large crowds, from multi-labeled annotations of narrative sentences provided by a handful of crowdsourcing annotators. This would reduce the cost of preparing multi-labeled training data for use with narrative-oriented affect prediction techniques with minimal degradation in the quality of the results.

2. Statistical Models

To incorporate dependency relationships among affect labels, we use the concept of conjoint-affect. Let J be the set of optional affect labels. A conjoint-affect represents a subset of J. The problem setting is similar to that of [Dawid 79]. Let I denote the set of annotated sentences. We first depict the story emotionality as the distribution of conjoint-affects $p_{\hat{J}}\left(\hat{J} \subseteq J\right)$, which is the ratio of the sentences that expresses conjoint-affect $\hat{J}$ among I. Let $c_i\,(i \in I)$ denote the character of sentence i, and let $I_{c_i}\,(i \in I)$ denote the sentences of character c_i. Similar to the story emotionality, the personality of character c_i is also represented by the distribution of conjoint-affects $m_{c_i\hat{J}}\left(i \in I, \hat{J} \subseteq J\right)$, which is the ratio of the sentence that expresses conjoint-affect $\hat{J}$ among I_{c_i}. $T_i \subseteq J\,(i \in I)$ represents the true conjoint-affect, namely the multiple true affects, for sentence i. K is the set of annotators, and $n^k_{i\hat{L}} \in \mathbb{N}\left(k \in K, i \in I, \hat{L} \subseteq J\right)$ denotes the number of times that annotator k annotated sentence i with the conjoint-affect $\hat{L}$. Let $E[T_i = \hat{J}]$ represent the expectation of true conjoint-affect of sentence i:

$$E[T_i = \hat{J}] = \Pr\left(T_i = \hat{J} \mid \left\{n^k_{i\hat{L}}\right\}_{k \in K, \hat{L} \subseteq J}, p_{\hat{J}}, m_{c_i\hat{J}}\right)$$

Contact: Lei Duan, Graduate School of Information Science and Technology, Hokkaido University, Kita 14, Nishi 9, Kita-ku, Sapporo, Hokkaido 060-0814, Japan, Tel: 090-2816-8002, Fax: 011-706-7831, duan@ec.hokudai.ac.jp

The true conjoint-label for sentence i is the one achieves the maximum expectation. This means that, the true conjoint-label T has to keep consistent with the affect tendencies of story emotionality p and character personality m.

3. Empirical study

To evaluate the effectiveness of proposed model for multi-affect estimation, we conducted experiments using the Lancers crowdsourcing service[*1]. We choose two Japanese children's stories, "Although we are in love" [*2] and "Little Masa and a red apple"[*3], as the annotated texts, because we believed that children's stories will have relatively vibrant affection tint and distinct character personalities, which are the focus points of our research.

Annotators were asked to read the sentences and spontaneously check the character's affects generated by each sentence. While *the Big Six* [Cornelius 00] (i.e., happiness, fear, anger, surprise, disgust, and sadness) are typically used in affective computing research, We used ten affect classes in order to provide more choices to the annotators and thereby enable us to perform an in-depth study on multiple-label estimation. They were taken from the "Emotive Expression Dictionary" [Nakamura 93].

We used the general consensus as the gold standards. They were obtained by having each sentences annotated 30 times and then taking the majority vote. That is, the most often annotated conjoint-label for a sentence was used as the gold standard for that sentence. For the "Love" story, we asked every one of 30 annotators to annotate each sentence one time, which means that the 30 annotations for every sentence were provided by the same annotators. For the "Apple" story, annotators were not designated, so the 30 annotations for every sentence were provided by arbitrary annotators, and few if any of them annotated all of the sentences. This is a more realistic situation since it is not a good idea to submit a very big task to a crowdsourcing service because a big task tends to diminish annotator enthusiasm or even cause annotators to avoid the task. We conducted the "Apple" task in this way simply to examine the effects of "arbitrary annotator interference" on the model. Moreover, although our proposed models can handle a situation in which a sentence is annotated more than once by an annotator, it is still best to avoid this situation even though an annotator may interpret a linguistic unit differently at different times. Therefore, in our experiments, all the annotations for a sentence were obtained from different annotators, which means that $n^k_{i\hat{L}} \in \{0,1\} \left(k \in K, i \in I, \hat{L} \subseteq J\right)$.

To see the effect of the number of annotations per sentence on accuracy, we randomly split the annotations for a particular sentence into various numbers of groups of equal size, and estimated the multiple true affect labels for each sentence, given the annotations within each group. We conducted with five different group sizes: 3, 5, 10, 15, and 30.

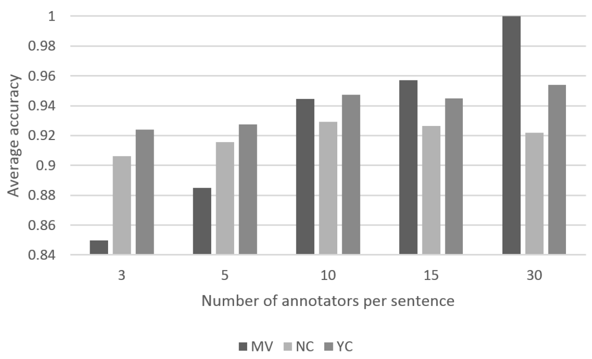

Figure 1: Average accuracies for affect prediction task for "Little Masa and a red apple" story

Since both the estimation result and the gold standard for a sentence can be regarded as a binary vector, the accuracy of a model was measured in terms of the average *Simple Matching Coefficient*, i.e., the average proportion of correct affect labels between the estimation results obtained from a group and the gold standards for all sentences. The average accuracies by groups for each group size were obtained with three models:

- MV: majority vote
- NC: model not considering consistencies
- YC: model considering consistencies

As shown in Figures 1 for the "apple" task, when the group size was 3, 5 or 10, almost all the statistical models achieved better accuracy than the MV. In other words, ten annotations at most for each sentence would be a reasonable number. Moreover, the YC model consistently outperformed the NC model, whose average accuracies remained basically unchanged as the group size increased. We obtained similar results for the "love" task. The experimental results show that considering consistencies of story emotionality and character personality among crowdsourced annotations is effective for the multi-affect estimation problem.

Acknowledgements

This work was supported in part by JSPS KAKENHI 24650061.

References

[Cornelius 00] Cornelius, R. R.: Theoretical approaches to emotion, in *ISCA Tutorial and Research Workshop (ITR-W) on Speech and Emotion* (2000)

[Dawid 79] Dawid, A. P. and Skene, A. M.: Maximum likelihood estimation of observer error-rates using the EM algorithm, *Applied statistics*, pp. 20–28 (1979)

[Nakamura 93] Nakamura, A.: Kanjo hyogen jiten [Dictionary of Emotive Expressions], *Tokyodo* (1993)

[*1] http://www.lancers.jp

[*2] 「僕たちは愛するけれど」
http://www.aozora.gr.jp/cards/001475/files/52111_47798.html

[*3] 「政ちゃんと赤いりんご」
http://www.aozora.gr.jp/cards/001475/files/52113_46622.html