3M3-OS-07d-6

Joint Estimation of Multiple Affects from Crowdsourced Annotations

Lei Duan, Satoshi Oyama, Haruhiko Sato, Masahito Kurihara [*1]

[*1] Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Japan

To build a high quality multi-affect annotated corpus for supervised affect prediction system, this paper explores some models used to estimate multiple affects for each sentence from multi-affect crowdsourced annotations given by fewer annotators, taking into consideration annotators' abilities and dependencies among affects.

1. Introduction

Expression of emotion, namely affect, is an important part of natural language. For a text-based affect prediction system using supervised learning algorithms, the quality of training data is critical to its performance. This paper explores some methods for estimating multiple affects from multi-affect annotations obtained by crowdsourcing.

2. Task Design

The annotated narrative text is a Japanese children's fairy story called "Little Masa and a red apple (政ちゃんと赤いりんご)". There are 78 lines in this story. 57 annotators are gathered from a crowdsourcing marketplace, Lancers. Each of them has to read some lines, then check the actor's emotions expressed by each line spontaneously. The response for each line can be multiple-choice. If an annotator has a feeling of that none of the emotions is reflected by the line, he (or she) should check "neutral". Each line is annotated by 30 of them. An affect set with 10 affects are chosen on the basis of "Emotive Expression Dictionary" (Nakamura, 1993), with their annotating frequency shown as Table 1. After annotating, a total of 3120 labels are collected, including 2768 affects and 352 neutrals.

Table 1 Frequency of affects and neutral

Affect	Frequency	Affect	Frequency
anger (An)	623	disgust (Di)	265
relief (Re)	362	surprise (Su)	243
neutral	*352*	fondness (Fo)	226
happiness (Ha)	306	fear (Fe)	107
sadness (Sa)	298	shame (Sh)	68
excitement (Ex)	270	**total**	**3120**

3. Estimation Algorithm

3.1 Naïve Voting

Naïve voting is the simplest method to estimate true affects for each sentence from crowdsourced annotations. With Naïve Voting, one affect is estimated to be true for a sentence if it is annotated equal or more than certain times for this sentence.

3.2 DS Model

Naïve Voting assumes all annotators have the same ability. (Dawid et al, 1977) proposed a model that considered annotator' predilection for certain affects. It is originally designed to estimate one true affect for each sentence from single-affect annotations:

$$\mathrm{argmax}_j P\left(T_{i,j} = 1 \middle| \text{annotations of } i\right) \quad (1)$$

In DS Model, there are J affects and I sentences. $T_{i,j} \in \{0,1\}$ is a set of indicator variables. If affect j is true for sentence i then $T_{i,j} = 1$ and $T_{i,q} = 0 (q \neq j)$. The true affect of sentence i should be the one that can maximize the posterior probability in (1).

We have extended DS Model to estimate multiple affects for each sentence from multi-affect annotations into two models.

(1) Affect Independent DS Model (I-DS)

I-DS estimates whether an affect is true for a sentence with the idea of DS. Every affect is estimated for each sentence independently. Parameters of I-DS (and the following D-DS) in (1), whose meanings are different from DS, are shown in Table 2.

(2) Affect Dependent DS Model (D-DS)

Being different from I-DS, D-DS conceives all 10 affects as an interrelated whole. The truth of every affect in an annotation can be portrayed as a "conjoint affect". With the 10 affects in our experiment, there are 1024 (2^{10}) different conjoint affects. D-DS estimates which affects are true for a sentence simultaneously.

Let X^i be the vector of 10 affects for sentence i. In order to estimate parameters of D-DS with EM algorithm, the missing data, the posterior probabilities in (1), are initialized as

$$\begin{aligned} & P\left(T_{i,j} = 1 \middle| \text{annotations of } i\right) \\ & = P({x_1}^i, {x_2}^i, \ldots, {x_{10}}^i) \\ & = \frac{\text{number of times sentence } i \text{ is annotated with conjoint affect } j}{\text{number of times sentence } i \text{ is annotated}} \end{aligned} \quad (2)$$

Because there are 2^{10} possible conjoint affects, we need to compare 2^{10} probabilities for each sentence. For the large quantity of parameters, a greater number of annotations are necessary. But our purpose is to obtain more accurate data from fewer annotators. So an approximating method is proposed in 3.3.

Table 2 Variables of DS and I-DS/D-DS Model

	DS Model	I-DS Model	D-DS Model
j	an affect	whether an affect is true	a conjoint affect
$T_{i,j} = 1$	affect j is true for sentence i	the affect is true for sentence i	conjoint affect j is true for sentence i
annotation of sentence i	single-affect annotations	Boolean-affect annotations	multiple-affect annotations

3.3 Approximating Affect Dependent DS Model (AD-DS)

To avoid the shortage of D-DS model, the joint probability in (2) can be approximated by using

$$P\left({x_1}^i, {x_2}^i, \ldots, {x_{10}}^i\right) = \prod_{a=1}^{10} P\left({x_a}^i | {x_b}^i\right), \quad 0 \le b < a \quad (3)$$

This algorithm is proposed by (Chow et al, 1968). It approximates a nth-order joint probability distribution with a product of

Contact information: Lei Duan, Graduate School of Information Science and Technology, Hokkaido University, Kita14 Nishi9, Kita, Sapporo, 090-2816-8002, duanleibnu@gmail.com

$n-1$ 2nd-order component distributions. The product can be graphically represented by a dependence tree. Figure 1 shows an example of a dependence tree in our experiment.

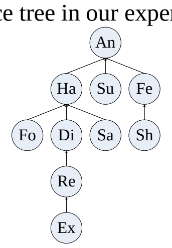


Figure.1 Example of a dependence tree

(The optimal approximation in Exp AD-DS with all sentences in 4.3.)

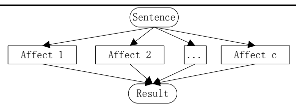
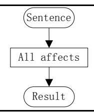
The product represented by Figure 1 is

$$P(X) = P(x_{An})P(x_{Ha}|x_{An})P(x_{Su}|x_{An})P(x_{Fe}|x_{An})P(x_{Fo}|x_{Ha}) P(x_{Di}|x_{Ha})P(x_{Sa}|x_{Ha})P(x_{Sh}|x_{Fe})P(x_{Re}|x_{Di})P(x_{Ex}|x_{Re})$$

To improve the performance of AD-DS, we intended to build one tree for each sentence (Sen AD-DS). But because of the limited annotations for each sentence, none of the 78 sentences can be denoted with a complete tree. So we also examined the performance of that all testing sentences share the same tree in the experiment (Exp AD-DS) to observe AD-DS intensively.

The relationships and flows of proposed multi-affect estimation models based on DS are summarized in Table 3.

Table 3 Relations and flows of proposed models

	Affect Independent	Affect Dependent
DS Model	I-DS	D-DS
		AD-DS
Flow		

4. Estimation Experiments

4.1 Accuracy Evaluation

For each sentence, we randomly divided the annotators who annotated the sentence into several groups, in order to exam the performance of each model on fewer annotators.

To evaluate model accuracy scientifically, each affect is estimated as a numeric value for a sentence on a scale from 1 to 0, where 1 means the sentence completely expresses this affect, and 0 means the sentence does not express this affect at all. So a model result of an annotator group of a sentence and the gold standard of this sentence can be treated as a 10-dimensional vector. The similarity between them is measured with cosine. The accuracy of a model is evaluated with the average of similarities between its results of annotator groups and gold standards of sentences.

4.2 Gold Standard

In our experiment, Naïve Voting in 3.1 constitutes the gold standard of each sentence, given its simplicity and objectivity. The rate of one affect for a sentence is estimated by the annotated probability: if a sentence is annotated by n annotators, m of whom choosed the affect, the rate of the affect is m/n.

4.3 Results and Discussion

Table 4 shows the accuracies of models, with the number of annotators in a group being 3 (10 groups), 5 (6 groups) and 10 (3 groups) respectively.

Table 4 Accuracies of models with different numbers of annotators in a group

	3 annotators	5 annotators	10 annotators
I-DS	0.738	0.772	0.832
D-DS	0.670	0.675	0.726
Sen AD-DS	0.673	0.694	0.721
Exp AD-DS	0.702	0.729	0.749

For all the 3 kinds of grouping, I-DS shows the best performance. Exp AD-DS performed a little better than D-DS and Sen AD-DS. All 4 models performs better along with the growth of the number of annotators in a group. In fact, one sentence is annotated merely 1.18 affects on average in the annotating task. It seems that the instruction to the annotators is ambiguous, which only stated that multiple-choice is possible. So almost all annotators annotated only 1 affect for a sentence. In this circumstance, I-DS performs best by its "independent" property. But 1.18 seems insufficient as "multiple" for our expectation: joint estimation of multiple affects. To make up this deficiency, we also examed models with the top 4 sentences whose numbers of annotated affects are largest. The average of annotated affects of these 4 sentences is 1.81. Table 5 shows the accuracies.

Table 5 Accuracies of models with top 4 sentences

	3 annotators	5 annotators	10 annotators
I-DS	0.824	0.776	0.762
D-DS	0.815	0.847	0.828
Sen AD-DS	0.805	0.838	0.752
Exp AD-DS	0.782	0.793	0.723

At this time although I-DS still performs best when there are 3 annotators in each group, the superiority with the other 3 models is not so obvious. When the number of annotators in a group increased to 5, D-DS performes best; Sen and Exp AD-DS come second and third. It demonstrates the superiority of affect dependent models for multi-affect estimation. But with the number of annotators in a group increasing to 10, accuracies of all 4 models decreased. Perhaps the number of annotators is too large for the relatively small number of sentences, so some confusion is arisen.

5. Conclusion

In this paper, we proposed some models to estimate multiple affects for each sentence from multi-affect crowdsourced annotations given by fewer annotators. Being restricted by the mild dependencies among annotated affects, the result of our research is not very prominent. We plan to conduct an experiment with instructions that encourage annotators to choose affects expressed by the sentence as many as possible, and examine the models proposed in this paper with these multi-affect annotations in our future work.

Acknowledgement

This work was supported in part by JSPS KAKENHI 24650061.

References

[1] A. Nakamura. 1993. Kanjo hyogen jiten [Dictionary of Emotive Expressions] (in Japanese), *Tokyodo Publishing*, Tokyo.

[2] A. P. Dawid and A. M. Skene. 1979. Maximum Likelihood Estimation of Observer Error-rates using the EM Algorithm. *Applied Statistics*, Vol. 28, No.1, pp.20-28.

[3] C. K. Chow and C. N. Liu. 1968. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, Vol. IT-14 No. 3, 462–467.