# A Construction of Implicit Emotion Cause Corpus

## A.1 Scheme for Data Collection and Annotation

Although there is a way to transform current explicit emotion cause dataset by masking emotion keywords, it still ignores some linguistic features of implicit emotion documents such as the lack of emotion context architecture.

Our implicit emotion cause corpus is built based on ISEAR, an implicit emotion corpus of self-reported affect, in which every sample describes events or situations that the participants had experienced of seven pre-defined emotions(*joy, fear, anger, sadness, disgust, shame, and guilt*) without mentioning the emotion explicitly. Firstly, in order to ensure the quality of the original English dataset, we correct the word and grammatical errors of the samples and remove duplicate sentences. We also get rid of samples that are too short to contain cause events. In order to eliminate the latent adverse effects of different languages when conducting comparative experiments with current explicit emotion cause benchmark dataset, several mature translation tools *Google Translate, Youdao Translate, Xunfei Translate* are used to translate all these English samples into Chinese in multiple rounds.

To evaluate the quality of translated Chinese samples, we ask two annotators to score each sample's translation accuracy ranging from level 1(*worst*) - level 5(*best*) according to the criteria in Table 1. For those samples below level 3, annotators will re-evaluate after proofreading by 2 professional native speakers. As shown in Table 2, 22% of samples are proofread to ensure their quality.

Finally a Chinese corpus contains 7423 implicit emotion cause samples is obtained based on original ISEAR samples. For each sample, annotators label the implicit emotion(*joy, fear, anger, sadness, disgust, shame, and guilt*) of original sentence and the set of events. For simplicity, a sample is given in the Fig 1 regardless of how they are represented in corpus. In order to better label and proofread the samples, a software has been developed to realize semi-automatic labeling of samples and have applied for software copyright.

| **Document** | 我显然受到了不公正的待遇，并且始终无法解释清楚这一点。(*I obviously suffer from the unfair treatment, and have never been able to explain this point clearly.*) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Emotion** | **Anger** | | | | | | | |
| **7-Tuple** | ***Att_Subj*** | ***Subj*** | ***Adv*** | ***P*** | ***Cpl*** | ***Att_Obj*** | ***Obj*** | ***Cause*** |
| **Events** | [NULL] | '我' (I) | '显然' (obviously) | '受到了' (suffer from) | [NULL] | '不公正的' (unfair) | '待遇' (treatment) | Y |
| | [NULL] | '我' (I) | '并且始终无法' (have never ) | '解释' (explain) | '清楚' (clearly) | [NULL] | '这一点' (this point) | N |

Figure 1: A Sample in the Corpus

## A.2 Analysis of the Corpus

In our implicit emotion cause corpus, there are totally 7423 samples. Sample distribution among different emotions and the number of events of each emotion is shown in Table 3. From Table 3, we find that the distribution of samples corresponding to each emotion is almost the same but there is a greater difference in the total number of events for each emotion that the biggest gap reaches 620.

In our annotation scheme, the definition of event is centered on verbs. However, in Chinese, cause events are categorized into two types: verbal events and nominal events. Verbal events refer to events that involve verbs, whereas nominal events are simply nouns.Therefore, we summarize the distribution of nominal events and verbal events for every emotion in the dataset. Further, we describe the distribution of samples containing different numbers of cause events for each emotion in the Table 4 and Table 5. Firstly, there are only 264 samples *(256 samples containing 1 nominal event + 8 samples containing more than 1 nominal event)* that contain nominal events and this part of samples just account for a very small proportion of all samples. This shows that if we ignore this part of the sample and focus our attention mainly on the extraction of verbal events, the effect of the model will not be reduced much. Secondly, 88.06% of the samples contain one or two verbal events and only 1.52% of sample have more than three verbal events. This may be caused by the length of the sample text.

Another important feature of samples in the corpus is that

| Translation Level | Criteria |
|---|---|
| Level 1 | The translation is obscure and difficult to understand, and the meaning of the source text cannot be delivered. |
| Level 2 | The translation does not reflect the semantics of the source text, and there exists: a) Main components of the source text can be translated in target text, but it cannot constitute a fluent translation due to sequence problems, logical errors, serious grammatical errors, etc. b) The translation is basically fluent, but serious problems (such as translation errors in negatives and double negatives, serious omissions, mistranslation of keywords, and excessive translation of the original text) exist in target texts. |
| Level 3 | The translation can reflect the semantics of the source text and is basically fluent (the order of grammatical components such as subject, predicate and object are correct), but there are improper translations of semantic keywords, omissions or mistranslations of non-keywords, etc. |
| Level 4 | The translation can reflect the semantics of the source text and is basically fluent (the order of grammatical components such as subject, predicate and object are correct), but there are a small number of improper words or improper collocates, etc. |
| Level 5 | The translation completely and explicitly reflects semantics and the translation is fluent. |

Table 1: Human evaluation of translated documents

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Before-PR | 4% | 12% | 6% | 34% | 44% |
| After-PR | 0% | 0% | 0% | 34% | 66% |

Table 2: Human evaluation of translated documents

| Emotion | sample | | event | |
|---|---|---|---|---|
| | number | P(%) | number | P(%) |
| *Joy* | 1098 | 14.79% | 1967 | 12.35% |
| *Fear* | 1066 | 14.36% | 2530 | 15.88% |
| *Anger* | 1110 | 14.95% | 2587 | 16.24% |
| *Sadness* | 1120 | 15.09% | 2195 | 13.78% |
| *Disgust* | 978 | 13.18% | 2021 | 12.68% |
| *Shame* | 1014 | 13.66% | 2196 | 13.78% |
| *Guilt* | 1037 | 13.97% | 2437 | 15.29% |

Table 3: Sample and Event Distribution of Different Emotions

| Emotion | nominal event number | | |
|---|---|---|---|
| | #0 | #1 | # >1 |
| *Joy* | 1061 | 35 | 2 |
| *Fear* | 1016 | 47 | 3 |
| *Anger* | 1072 | 37 | 1 |
| *Sadness* | 1080 | 40 | 0 |
| *Disgust* | 935 | 42 | 1 |
| *Shame* | 986 | 28 | 0 |
| *Guilt* | 1009 | 27 | 1 |
| *Total* | 7159 | 256 | 8 |
| *P(%)* | 96.44% | 3.45% | 0.11% |

Table 4: Distribution of Samples containing Nominal Events

| Emotion | verbal event number | | | | |
|---|---|---|---|---|---|
| | #0 | #1 | #2 | #3 | # >3 |
| *Joy* | 49 | 819 | 177 | 43 | 10 |
| *Fear* | 61 | 641 | 270 | 69 | 25 |
| *Anger* | 49 | 711 | 246 | 85 | 19 |
| *Sadness* | 52 | 815 | 195 | 46 | 12 |
| *Disgust* | 64 | 680 | 167 | 50 | 17 |
| *Shame* | 49 | 673 | 229 | 49 | 14 |
| *Guilt* | 35 | 654 | 260 | 72 | 16 |
| *Total* | 359 | 4993 | 1544 | 414 | 113 |
| *P(%)* | 4.84% | 67.26% | 20.80% | 5.58% | 1.52% |

Table 5: Distribution of Samples containing Verbal Events

subjects of most samples are pronouns(such as *I*, *he* or *she*) without obvious emotional information. From this point of view, we should pay more attention to the verbs of events that contains more emotional information when building models. This is consistent with our definition of events with verbs at the core. The lack of emotional information inherent in the task and context information in the corpus we constructed makes it necessary to introduce external knowledge into our model.

# B Experiments

## B.1 Metrics

We use the commonly accepted measure proposed by Lee for emotion cause extraction. In this measure, if a proposed emotion cause covers the annotated answer, the sequence is considered correct. The performance measures are Precision(P), Recall(R), and F-measure(F1), which are defined by

$$Precision = \frac{\sum correct\_cause\_events}{\sum proposed\_cause\_events},$$
$$Recall = \frac{\sum correct\_cause\_events}{\sum annotated\_cause\_events},$$
$$F - Measure = \frac{2 * Precision * Recall}{Precision + Recall}.$$

## B.2 Experimental Settings

For all our experiments, pre-trained word vectors on Chinese Wikipedia using Word2Vec toolkit are leveraged to initialize the word embeddings. The dimension of word embedding is set to be 300. And the dimensions of query, key and value of attention module are 64, 64 and 64 respectively.

The maximum numbers of words in each original texts and cause event texts are set to be 169 and 44 respectively. The network is trained based on a RMSprop optimizer with a mini-batch size 32 and a learning rate 0.0001. The epoch number is set to be 8. Based on a pilot study, we find the best value for the parameter $\lambda$ balancing the impacts of valence and arousal on computing emotion affectiveness is 0.25.