We propose a double standpoint evaluation method (DSEM) for systems analyzing and recognizing affect in utterances. We use this method to evaluate our ML-Ask system and show the differences in results for two different standpoints – recognitive and commonsensical. Evaluation based on the former shows system’s accuracy in affect recognition needed for user-agent communication. The one based on the latter verifies system’s unanimity with the general commonsense interpretation of the affect conveyed in the utterance. Such evaluation is relevant to confirm the results of the recognition evaluation and can be further applied to sentiment analysis. Both standpoints are relevant and applying DSEM in research on affect analysis can verify performance of a system in much wider way than measures widely accepted.

Keywords: affect analysis, analysis of emotiveness, evaluation methods.

1. Introduction

Research on mechanical processing of emotions, including recognition and analysis of affect is rather young discipline of study. Since being initiated by Picard only in 1995 [1] the field has gathered popularity at an exponential rate. Number of scientists has been proposing their ideas on how to recognize emotions automatically, presenting higher or lower results. However, on what ground the results have been achieved, is a problem rather left unsaid till now. After over ten years of development of the field, there were none significant or reliable ideas on how to objectively evaluate the emotion recognition methods. Substitute and unreliable measures might put the fairness of the results in questioning or doubt. In this paper we propose a fair and thorough method of evaluation for systems analyzing the affect of utterances.

2. Problems with what to recognize and how

There are several urgent problems with evaluation methods we noticed in the studies on affect analysis/recognition.

2.1 Searching affect only in common words

Emotions are the domain of human-human communication. Therefore research on recognizing them should be focused on utterances and whole conversations rather than on isolated words. Although there are words which display more emotive coloring than the others, the emotiveness of such words becomes visible only in comparing to words similar semantically, but different pragmatically (see examples for Japanese in Table 1).

However, despite the self-evident nature of the above, there are still scientists developing methods for computing emotiveness of e.g. common nouns, like “pencil” or “laundry” [3], which drives us to nonsensical conclusions, like “laundry is joyful and dreadful”, and “pencil is favorable and enthusiastic”.

<table>
<thead>
<tr>
<th>non-emotive</th>
<th>emotive</th>
</tr>
</thead>
<tbody>
<tr>
<td>父</td>
<td>オヤジ</td>
</tr>
<tr>
<td>chichi</td>
<td>oyaji</td>
</tr>
<tr>
<td>father</td>
<td>old man</td>
</tr>
</tbody>
</table>

Table 1 Words with similar semantic-, but different emotive meaning [2].

2.2 Interfering in evaluation

In creating corpus for evaluation it is obligatory to keep authors’ interference as small as possible. However, it seems popular to interfere in the process of emotional tagging of the corpus [4], which might suggest that the work lacks objectivity.

2.3 Trap of commonsensical recognition

Recognition is from the definition a target-oriented process. Despite of that it is not rare to find works with evaluations of recognition results based on the approximated judgment of a third party of evaluators [5]. Often the evaluation is oversimplified to asking the evaluators whether system’s results were reasonable [6], although Rzepka states clearly that such way of evaluation is inappropriate since it depends highly on the evaluator’s imagination and experiences [7]. Employing the third party into the evaluation process contradicts the idea of recognition form the very beginning, however it is understandable that such evaluation method is far easier to perform than creating a fair and reliable corpus tagged by authors.
among which some were tagged only by a small number of
we used a set of sixty fully tagged items with additional thirty
conversations set in a specific context. For the first step however,
We distinguish at least three levels of emotiveness recognizable
emotiveness conveyed in a textual representation of an utterance.
also perform an evaluation of the same system using one of the
evaluation our former research - System ML-Ask [9, 10]. We
began gathering material for such corpus. We continue both:
appropriate corpus with multi-faceted tagging. Therefore we
perform the evaluation basing on the large corpus of utterances
tagged by authors of the utterances. The evaluation is performed
on a large number of sentences. To broaden the evaluation, we
apply the commonsense standpoint by taking into consideration
an opinion of the third party – large number of human evaluators.
However we do not ask the third party if the systems’ results
were reasonable, but make them perform the same actions as the
system. The conclusions for the evaluation are drawn on how
much the systems’ results coincide with the results of the general
commonsense of the third party.
All of the assumptions above make up DSEM – a broad
double standpoint evaluation method potentially capable to judge
the system fairly and without distortions.
3.1 Corpus for DSEM
For deep evaluation provided by DSEM there is a need for an
appropriate corpus with multi-faceted tagging. Therefore we
began gathering material for such corpus. We continue both:
gathering new utterances tagged by its authors and tagging by a
number of third-party human evaluators. Ultimately we plan to
gather both short and long utterances and add tagging for whole
conversations set in a specific context. For the first step however,
we used a set of sixty fully tagged items with additional thirty
among which some were tagged only by a small number of
evaluators.
4. DSEM for ML-Ask
We put the method into practice to perform a thorough
evaluation our former research - System ML-Ask [9, 10]. We
also perform an evaluation of the same system using one of the
shallow methods described before.
4.1 ML-Ask system – short description
ML-Ask is a system for multidimensional analysis of
emotiveness conveyed in a textual representation of an utterance.
We distinguish at least three levels of emotiveness recognizable
by the system. The utterance can: a) be either emotive or non-
emotive (or neutral), b) have a specified emotive value, and c)
convey specified feelings. This method of analyzing emotiveness,
on which ML-Ask was created, is based on Ptaszynski’s idea of
finding emotive elements in the text [11]. In an utterance made
by the user emotive elements are examined using the top-down
determined databases of emotive elements in speech. The
databases of each type of emotive elements appearing in
conversation in Japanese were gathered basing on different
researches. The databases are divided into interjections, emotive
mimetics (gitaigo), endearments, vulgar vocabulary, which
belong to lexical layer of speech, and symbols representing
emotive elements from non-lexical layer of speech, like
exclamation marks, syllable prolongation marks, etc. We also
added an algorithm recognizing emoticons, as symbols already
widespread and commonly used in everyday Internet
communication tools. A few simple examples of sentences
recognized this way as non-emotive value (A, B), and emotive
(A’, B’) are given below. The parts of each sentence that
constitute its emotiveness were written in bold letters.

A: 今日はいい天気です。
Kyō wa ii tenki desu.
It is a good weather today.
A': ああ、今日はええ天気だな ! (^o^)
Aa, kyō wa ee tenki dana ! (^o^)
Wow, now today is a fine weather! :D

B: 彼女は、大きいかさをもってきて、信之介を強く殴った。
Kanojo wa, ookii kasa wo mottekite, Shinnosuke wo tsuyoku
nagutta.
She brought a large umbrella and strongly hit Shinnosuke.
B': あいつぁ でっけーかさをもってきてやがって、シンちゃんをひ
でー ゴコロにじしまった！
Aitsaa dekkē kasa wo motekiyaagatte, Shin-chan wo hidē
bokoboko ni shichimatta !
That slut lugged a huge umbrella with her and beat the crap
out of Shin-chan.

After analyzing every utterance this way, the system returns a
verdict whether the utterance is emotive and what emotive
elements were found in the utterance. On this basis the system
proposes its emotive value of the sentence. The value is placed
on a scale of 0 to 5 and 1 point is counted for every piece of
emotive element found in the sentence (but with maximum value
of 5). In the next step, in the utterances determined as emotive,
the system, basing on a database of emotive expressions,
determines what specific feelings were conveyed. This database
as well as classification of emotion types is borrowed from
Nakamura’s collection [12].

4.2 ML-Ask system evaluation using DSEM
We performed an evaluation for ML-Ask using DSEM.
4.3.1 Accuracy evaluation
We evaluated three levels of analysis performed by the system.
(1) Emotive / non-emotive
The total accuracy of the system in determining whether an
utterance is emotive or not is the approximated balanced F-score $F_{E/NE}$
for recognizing both emotive and non-emotive utterances
represented as in figure (1) and amounted $F_{E/NE}=0.81$.

\[
F_{E/NE} = \frac{F \_E + F \_NE}{2}
\]  (1)
4.3.1 Tsuchiya’s evaluation method – short description

The emotion types result of recognizing any- and at least one feeling from the accuracy of specific emotion types the rest – the equivalent of commonsense. The unanimity between human evaluators gives us the level of unanimity of one evaluator with all (including the system treated equally with other evaluators). The consideration the tagging made by authors of the utterances. The third party human evaluators. In this stage we do not take into account how much the results coincide with the results of the speaker, we assented to a condition of almost-perfect match (a case when the emotive value differs between speaker and system by ±1 emotive point per utterance). The accuracy of setting the emotive value within this condition reached 67% for twelve items tagged this way.

(3) Specified emotions

By recognizing a specified type of emotion we understand a result of recognizing any- and at least one feeling from the utterance, including “non-emotive”. The emotive value recognition accuracy is counted as an $F_{ETR}$ value from an approximation of an accuracy to determine about “non-emotiveness” $F_{NE}$, and the accuracy to determine about the specific emotion types $F_{ET}$, as in figure (2). The system acquired accuracy of $F_{ETR} = 0.46$.

$$F_{ETR} = \frac{F_{ET} + F_{NE}}{2} \tag{2}$$

4.3.2 Commonsense evaluation

In evaluation based on the commonsensical standpoint, of the systems’ unanimity with the generally understood commonsense by checking how much the results coincide with the results of the third party human evaluators. In this stage we do not take into consideration the tagging made by authors of the utterances. The unanimity $U$ between one evaluator $Eva_A$ and the other $Eva_B$ is a simple relationship of the number of similarly tagged utterances $t_{W}$ to all of the utterances tagged by the human evaluators $t_i$ as showed in figure (3), and is counted for every pair of them (including the system treated equally with other evaluators). The average gives us the level of unanimity of one evaluator with all the rest – the equivalent of commonsense.

$$U^Eva_A^Eva_B = \frac{t_{sim}}{t_i} \tag{3}$$

(1) Emotive / non-emotive

The unanimity between human evaluators in determining whether the utterance is emotive or not was set at a very wide range between 47% and 95%. However, the average unanimity between all human evaluators was set at a very narrow bracket of 68% to 79% (with general average = 75%). In this point of evaluation the goal for the system was to fit in this bracket. However the systems’ unanimity with all evaluators reached 58%, which is a little above 77% of the general average unanimity between humans. The result is close to the result from the cognitive evaluation, which is very promising.

As an interesting fact we might add that for sentences with a perfect match, where both authors and all evaluators were unanimous about the emotiveness (18 of 60 sentences), systems’ results reached 100%.

(2) Emotive value

With the condition of almost-perfect match (see 4.3.1(2)) assented, the unanimity of ML-Ask with eight human evaluators (2 females and 6 males) was set at a range of 50% to 88%. Although the bracket is wide, we consider this result as satisfactory, since emotional intensity setting is highly subjective among people. For the comparison, the approximate of unanimity among the evaluators themselves about the emotive value of the sentences was set at a level of 34% - 74%.

(3) Specified emotions

For evaluation of specified emotion recognition by the system we performed another survey. We asked twelve different people (2 females, 10 males) about emotions conveyed in emotive utterances (with a possibility of specifying more than one feeling). In many cases the results differed significantly and there were sentences with emotions unidentifiable by some evaluators. For such conditions the following assumptions were made. If ML-Ask guessed at least one of the emotion types classified by all evaluators per sentence, or the systems’ classification coincided with the majority, the result was positive. In final results ML-Ask achieved an accuracy of 45% of the human level in recognizing the specific types of emotions, which confirms the result acquired in the recognition evaluation.

The result is satisfactory and encouraging, although is not perfect, which arises from lacks in appropriate databases created on Nakamura’s collection [12]. However, we have already started to retrieve new entries from the Internet by using keywords from his collection to update the database, what clearly prognosticates the improvement.

4.3 Comparing DSEM to other evaluation method

We also compared DSEM to other evaluation method to show how evaluation method influences the view on a system. For comparison we took Tsuchiya’s et al. method, which, in their opinion decided that, the system they proposed was highly effective (accuracy 88%) [6].

4.3.1 Tsuchiya’s evaluation method – short description

In their evaluation, Tsuchiya et al. asked five people to verify how commonsensical were the results given by their system. The evaluators had three options: A) commonsensical, B) “non-commonsensical” and C) “uncommonsensical”. The result was counted positive for the evaluation if either A) or B) option were chosen. Furthermore, if at least two of the five evaluators gave a positive verdict, systems’ result was positive. In this rather lenient way Tsuchiya et al. showed their system achieved 88% of accuracy.
4.3.2 ML-Ask in the perspective of “non-uncommonsensicalness”

We performed an evaluation of results given by ML-Ask using Tsuchiya’s method. The system achieved the unbelievable accuracy of 97%, which would mean that it is almost perfect and simply outperforms all of the present ones. Although it is obviously our goal to achieve that, we would feel insecure to know that our evaluation method was chosen to increase the real results.

5 Conclusions

In this paper we presented DSEM - an objective method suitable for systems analyzing and recognizing emotions. The method is based on two standpoints of evaluation – recognition and commonsensical. We put the method into practice to evaluate system ML-Ask. The evaluation gave very promising results, but also helped us to realize what should be a question of concern in the future research on the project.

ML-Ask achieved a high accuracy result of 0.81 of balanced F-score in recognizing general emotiveness of an utterance. This level was confirmed in commonsensical evaluation with achieving a close result, 77% of general average unanimity between human evaluators.

The emotive value of an utterance is recognized by the system with an accuracy of 67%. Although the method was objectively confirmed as commonsensical, it is desirable to upgrade the method of setting emotive value to make it closer to the speakers’ intention.

The system recognizes specific types of emotions conveyed in utterance on a fair, but upgradeable level of 0.45 of balanced F-score. This level was also confirmed by the commonsensical evaluation.

However, what was the key thought of this paper, DSEM, in comparison to other methods clearly revealed distortions of the latter. We showed the advantage of DSEM in comparison to one of the most popular evaluation method in the field today. Although the method proposed by us requires more effort to perform it, it shows the results more accurately, without distortions and bending. It is desirable for this method to be accepted widely in field.

6 Future Work

We began gathering of a large evaluation corpus for DSEM to be used in the future research. At present the method is based on tagged utterances, although in the future we plan to prepare a broader corpus with tagging on whole conversations. This would help greatly in the future research on contextual recognition of emotions.

Since DSEM is a good mean of solving the problem of ambiguities, it also seems to be suitable for other fields of computer science contending with it, like sentiment analysis or humor processing [13].

References