2P2-7

Double Standpoint Evaluation Method for Affect Analysis Systems

Michal PTASZYNSKI [*1] Pawel DYBALA [*1] Rafal RZEPKA [*1] Kenji ARAKI [*1]

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

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".

non-emotive	emotive
父	オヤジ
chichi	*oyaji*
father	old man

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.

2.4 Tiny evaluation

The most popular problem, which often comes with the former one, is the number of the evaluators employed to the process. If researchers decide to check only how the third party evaluates the results, it is appropriate to ask as many evaluators as possible to get a wide view on the results. Unfortunately many scientists limit their evaluation to, e.g. five people [6], where others settle for even three [8]. Evaluation limited to such borders surely

Contact information: Michal Ptaszynski, Affiliation: Language Media Laboratory, Graduate School of Information Science and Technology, Hokkaido University, Address: Kita-ku, N-14 W-9, 060-0814 Sapporo, Japan, Tel: +81-11-706-7389, Fax: +81-11-709-6277, E-mail: ptaszynski@media.eng.hokudai.ac.jp

provides less cumbersome results, but it is highly questionable, whether it is sufficient at all.

3. Double Standpoint Evaluation Method

Basing on all problems stated above we could work out our own fair method of evaluation for affect analysis systems. First of all, as a premise, we do not take isolated words as an object of research. The minimal unit of our interest is an utterance. By the utterance we mean any act where a set of communicative signs is uttered by sender to receiver. It can be simple or consist of a number of sentences. To judge the true precision of a system 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 Shinnnosuke.
B': **あいつぁ でっけー**かさをもってき**やがって**、シン**ちゃん**を**ひでー ボコボコ**に**しちまった ！**
Aitsaa dekkē** kasa wo mottekiyagatte, 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 specified 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 calculated as in figure (1) and amounted $F_{E/NE}$=0.81.

$$F_{E/NE} = \frac{F_E + F_{NE}}{2} \qquad (1)$$

Such result is very promising, since the same value counted for human evaluators gave a wide range of results from 0.4 to 0.86. ML-Ask is placed in the top of this ranking, so we can say that the system recognizes emotiveness on a very high level.

(2) Emotive value

Since emotive value of an utterance is highly dependable on many constantly changing situational features and it is impossible to achieve a perfect match with analysis of only textual layer of an utterance, in the process of evaluating the system's unanimity with 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 emotion types 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.

$$F_{ETR} = \frac{F_{ET} + F_{NE}}{2} \qquad (2)$$

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_{sim} 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.

$$U_{Eva_B}^{Eva_A} = \frac{t_{sim}}{t_i} \qquad (3)$$

(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 recognitive 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-uncommonsensical" 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 – recognitive 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 commonsense 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

[1] Picard, R. W.: *Affective Computing*, MIT Technical Report #321, MIT Media Laboratory, 1995.

[2] Kamei, T., Kouno, R., Chino, E.: *The Sanseido Encyclopedia of Linguistics*, Vol. 6, Sanseido, 1996.

[3] Saito, T. et al.: *Meishi no kanjoshozokusei no chushutsu to sore ni motozuku meishi no ruijido no keisan* (Computing similarity of nouns on the basis of noun emotional affiliation), Proceedings of The Fourteenth Annual Meeting of The Association for Natural Language Processing, 2008.

[4] Wu, C. H., Chuang Z. J, Lin Y. C.: *Emotion Recognition from Text Using Semantic Labels and Separable Mixture Models*, ACM Transactions on Asian Language Information Processing, 2006.

[5] Alm, C. O., Roth, D., Sproat, R.: *Emotions from text: machine learning for text based emotion prediction*, HLT/EMNLP, Vancouver, 2005.

[6] Tsuchiya, S., Yoshimura, E., Watabe, H., Kawaoka, T.: *The Method of the Emotion Judgement Based on an Association Mechanism*, Journal of Natural Language Processing, Vol. 14, No. 3, The Association for Natural Language Processing, 2007.

[7] Rzepka, R., Araki, K.: *What About Tests In Smart Environments? On Possible Problems With Common Sense In Ambient Intelligence*, Proceedings of 2nd Workshop on Artificial Intelligence Techniques for Ambient Intelligence, IJCAI'07, 2007.

[8] Endo, D., Saito, S., Yamamoto, K.: *Kakariuke kankei wo riyo shita kanjoseikihyogen no chushutsu* (Extracting expressions evoking emotions using dependency structure), Proceedings of The Twelve Annual Meeting of The Association for Natural Language Processing, 2006.

[9] Ptaszynski, M., Dybala, P., Shi, W. H., Rzepka, R., Araki, K.: *Lexical Analysis of Emotiveness in Utterances for Automatic Joke Generation*, ITE Technical Report Vol.32, No.47, pp. 39-42, The Institute of Image Information and Television Engineers, 2007.

[10] Ptaszynski, M., Dybala, P., Rzepka, R., Araki, K.: *Effective Analysis of Emotiveness in Utterances Based on Features of Lexical and Non-Lexical Layers of Speech.*, Proceedings of The Fourteenth Annual Meeting of The Association for Natural Language Processing, 2008.

[11] Ptaszynski, M.: *Moeru gengo - Intānetto keijiban no ue no nihongo kaiwa ni okeru kanjōhyōgen no kōzō to kigōrontekikinō no bunseki - " 2channeru" denshikeijiban o rei toshite* - (Boisterous language. Analysis of structures and semiotic functions of emotive expressions in conversation on Japanese Internet bulletin board forum - 2channel -), UAM, Poznań, 2006.

[12] Nakamura, A.: *Kanjō hyōgen jiten* (Dictionary of Emotive Expressions), Tokyodo Publishing, 2004.

[13] Dybala, P., Rzepka, R., Araki, K.: *Dajare Generating Support Tool - Towards Applicable Linguistic Humor Processing*, Proceedings of The Fourteenth Annual Meeting of The Association for Natural Language Processing, 2008.