Fine-grained Sentiment Analysis with 32 Dimensions

Xianchao Wu[1], Hang Tong[2], Momo Klyen[1]

[1]A.I.&Research, Microsoft Development Co., Ltd.
[2] Department of Mechano-Informatics, The University of Tokyo
{xiancwu, momokl}@microsoft.com, tonkou@mi.t.u-tokyo.ac.jp

***Abstract*—Understanding human's complicated and capricious emotions remains a fundamental challenge. In this paper, we propose a fine-grained sentiment analysis system which classify emotions into 32 categories. For one direction, we cover more detailed emotions and for the other direction, we further measure each emotion with strength, such as describing angry by annoyance, anger and range. Taking Japanese as a test language, we describe our methods of building the training data, of constructing deep neural network classifiers, and of evaluating the models.**

***Keywords*-sentiment analysis; deep learning**

I. Introduction

Representing million level users' psychological emotions through machine learning techniques remains as a fundamental challenge for numerous natural language processing (NLP) applications, such as building emotional chatbots [1] and classifying consumers' opinions [2]. It is problematic if one sentence is simply classified into categories that only include *positive*, *negative*, and *neutral*. Real-world applications prefer a fine-grained emotion category set in which most of our ordinary emotions are included, such as happy, angry, surprise, sad and so on. This taxonomy is for sure helpful for us to better understanding people's detailed emotions and consequently to construct better emotion-driven applications.

Moreover, the *strength* of a detected emotion should also be taken care of by an emotion understanding method. For example, the strength of *angry* is different in the following two sentences: (1) *What you did anger me, I do not accept your apology* and (2) *Jack could not control his angry to Mary and he started to beat her heavily*. With different strengths, the emotions, even in the same category of *angry*, are different and should be treated in different ways. For example, when generating a response in an emotion-care oriented chatbot, we hope to take advantage of the fine-grained emotional classification and consequently prepare (by ranking and/or generating) their respectively fine-grained emotional responses. Such as, *anger* sentence (1) can be responded by *shall I apologize to making you feeling better?*, and the *rage* sentence (2) can be responded by *Jack, stop!* or *Mary, run away!* or even *I am calling the police*. One interesting work introduced in [1] is that responses from five emotional categories (*like*, *happy*, *sad*, *disgust*, and *angry*) are generated under their emotional chatting machine. We argue that a combination of their emotional chatting machine and our proposed fine-grained sentiment analysis (SA) is supposed to achieve more in terms of detailed users' emotion-care.

The major obstacles that this paper tries to tackle include, (1) choose a fine-grained emotion taxonomy to better cover people's dominant sentiment feelings and prepare the related training data; and (2) utilize emotional words and characters in Japanese spoken sentences for constructing and training neural network classifiers.

Work done when Hang was an intern in Microsoft.

II. Related Work

SA has been well developed in terms of methodology, of from richer features to deeper neural networks. For example, recursive deep models were proposed in [3] for semantic compositionality over the Stanford Sentiment Treebank[1]. Shared model adaptation was proposed in [2] for personalized sentiment classification using Amazon product reviews [4] and Yelp restaurant reviews[2] which contain opinion ratings in discrete five-star range. [5] proposed a deep memory network to explicitly capture the importance of contextual words when inferring the sentiment polarity of an aspect.

However, the data aspect was developed rather slower than the algorithm aspect. By far, the dominant taxonomies used by most data sets still base on rough category tags such as *positive*, *negative* and *neutral*. For example, the Stanford Sentiment Treebank [3] attaches degrees (*strong* and *very strong*) to *positive* and *negative* categories. Another well-known benchmark set is the SemEval data[3] for aspect based *positive/negative/neutral* sentiment analysis which links opinions with aspects.

One Chinese dataset with rather detailed emotions is the NLPCC dataset which contains 23K sentences collected from Weibo[4] (a mobile/website application that is alike twitter in Chinese). This dataset contains 8 emotion categories *anger*, *disgust*, *fear*, *happiness*, *like*, *sadness*, *surprise*, and *other*. In the dataset, there are two infrequent classes, *Fear* (1.5%) and *Surprise* (4.4%). This dataset was used in [1] for building a bidirectional long-short term memory (LSTM) emotion classifier for enhancing their emotional chatting machine for generating content and emotion appropriate responses. The NRC word-emotion

[1]https://nlp.stanford.edu/sentiment/index.html
[2]https://www.yelp.com/dataset_challenge
[3]http://alt.qcri.org/semeval2014/task4/
[4]http://weibo.com

978-1-5386-1981-0/17/$31.00

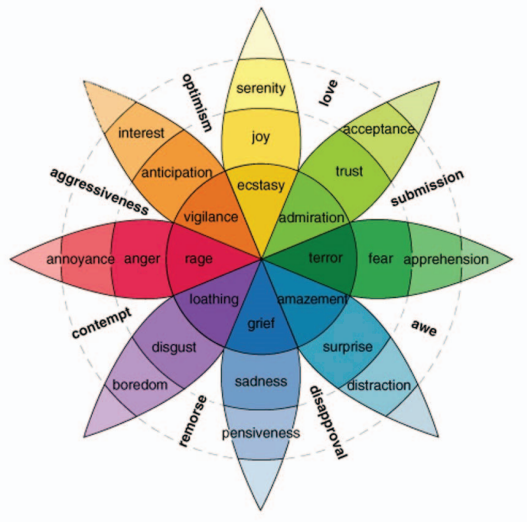

Figure 1. Plutchik's wheel of emotions.

association lexicon[5] (aka EmoLex) [6] includes around 15K unigrams of English which were manually annotated with eight basic emotions (*anger*, *fear*, *anticipation*, *trust*, *surprise*, *sadness*, *joy*, and *disgust*) and two sentiments (*negative* and *positive*). They also used Google Translator to obtain the Japanese version.

The major difference we made is that we build a fine-grained sentiment analysis dataset for Japanese language with not only detailed emotional categories but also the *strength* of the categories. As far as our knowledge, this is the first investigation about classification of textual messages using 32 emotions for one Asian language.

III. Fine-grained Emotion Labels

We follow one of the most influential classification theory proposed in [7], in which human emotions are classified into 32 categories: 8 primary emotions, 3 *strengths* for each primary emotion, and 8 combined emotions (such as *love* = *joy* + *trust*). Figure 1 illustrates Plutchik's wheel of emotions, in which, each *petal* stands for one primary emotion, and primary emotions intensify as they move from outside to the center of the wheel, with colors of from shallow to deep.

IV. Dataset

We collect emotional sentences from our in-house Japanese chatbot log. We follow the road-map described in [8], [9] to first build emotional seed words and then cover sentences.

Emotional Seed Words

First, we extend those 32 Japanese emotion labels in the wheel of emotions by word2vec [10]. This extension collects candidate seed words which are likely to have similar word meanings with the corresponding emotion labels. However, the defect of the extension is that contextually similar words can actually have opposite emotions. For example, *ureshii*/*joy* and *kanashii*/*sad* have a high word2vec similarity score since they are often followed by identical nouns (such as *movie*, *story*). To alleviate this shortage, we follow [9] and utilize a bilingual phrase translation table [11] to translate Japanese words into English and then translate back to Japanese to finally obtain another synonym set for each seed word. Then, we pick an intersection of the word2vec extension set and the word alignment synonym set. In addition, we manually collect the synonyms of each emotion labels from websites[6] of Japanese synonym dictionaries. As a result, we obtained 13,688 seed words for the entire 32 labels.

Emotional Sentence Collection Using Seed Words

The emotion of one sentence is likely to depend on its specific words. We leverage this tendency to guide our automatic labeling of sentences. For instance, we automatically label a sentence as *joy*, if it contains *joy*, or any other seed synonym words of *joy*. In total, we extract about 200K emotional sentences (except *neutral*) out of 500K sentences.

This automatically collected dataset has several limitations. First, there are few emotional sentences labeled as *admiration*, *loathing* or *distraction*. This is because such words or their synonyms rarely appear on conversational logs. This reflects that only focusing on seed words is inefficient to collect a balanced dataset. Second, we have so many *surprise* sentences because we admit consecutive exclamation marks, "!!", to be one seed word of *surprise* label. Thanks to that admittance, we can obtain not only large amount of *surprise* sentences, but also many "!!"-included emotional sentences, which should be labeled as tags other than *surprise*. Third, in Figure 1, the color intensity corresponds to the emotion intensity. However, these emotional labels are likely to share synonym seed words. For example, *hate* is considered to be a variant of not only *loathing*, but also *disgust*. This makes the automatic method to be more fragile.

Refining Dataset by Human Annotation

To tackle above problems and to still employ the benefits of the automatic method, we manually re-annotate part of the emotion labels. We extract 10K sentences out of the whole dataset, and instruct native speakers to (1) correct the wrong emotion labels and (2) determine the intensity of sentence. Specially, all the labels of one sentences will be kept from different annotators. By using this manually corrected dataset, We make word frequency vocabulary for each labels. From the vocabulary, we calculate the probability of the words contained by emotions. We utilize this vocabulary to re-calculate the sentence emotion labels. After this process, we obtain a refined emotion sentence dataset. The statistics of the final dataset are listed in Table I.

[5]http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

[6]http://www.weblio.jp/

Table I
NUMBER OF SENTENCES (#) PER LABEL.

label	#
ecstasy	191
joy	882
serenity	559
vigilance	91
anticipation	23
interest	668
rage	2,165
anger	160
annoyance	1,003
loathing	1
disgust	58
boredom	218
grief	517
sadness	1,630
pensiveness	31
amazement	220
surprise	222

label	#
distraction	197
terror	1,892
fear	13,750
apprehension	1,211
admiration	5
trust	516
acceptance	167
optimism	60
aggressiveness	397
contempt	338
remorse	172
disapproval	31
awe	64
submission	77
love	416
neutral	503,758
Total	535,396

V. EXPERIMENTS AND NEURAL NETWORKS

A. Setup

In order to focus on fine-grained emotion classification and balance the training data, we remove the large-amount *neutral* category during model training. From the dataset of Table I, we randomly select validation datasets and test datasets at the ratio of train:validate:test = 8:1:1.

To represent input sentences with word and character level features, we compare 4 types of models: unigram bag-of-words (BoW), word-based convolution neural network (CNN) [12], character recurrent CNN (RCNN) [13] and its variant of word-RCNN model. Generally, after the representations by embedding layers and then CNN and/or RNN layers, all the features are sent to two fully-connected layers and finally a softmax output layer.

For all models, we use Adam [14] as the optimizer, setting $\alpha = 1e-5$ and weight decay to be 0.9. We use unigrams in our word-based models. The embedding dimensions for words and characters are all 256. For the BoW model, we use a 256-dimension vector for each unigram word and two full-connection layers of first 256-256 projection and then 256-to-32 projection. All the elements of bias vectors were initialized to be zero. Any other weight matrices were initialized by sampling from the Gaussian distribution of mean 0 and variance 0.01.

We use four lengths (3, 5, 7, and 9) of convolution functions in the word-CNN model, the word-RCNN model, and the char-RCNN model. Each convolution function projects an input 256-dimension vector into another 256-dimension vector. Specially, for the Word-RCNN and char-RCNN model's recurrent layer, we adapt LSTM which projects a 256*4-dimension vector into a 256-dimension vector. Then, we build a full-connection layer of 256-to-256 projection with $Relu$ active function and finally another full-connection layer which projects the 256-dimension vector to a vector that is with the same dimension of the target emotion categories. We initialized the recurrent weight matrices in LSTMs as random orthogonal matrices. Note that, the networks of word-RCNN and char-RCNN are identical except that the embedding layer takes unigram words or unigram characters as inputs. The difference between word-RCNN and Word-CNN is to whether use a LSTM layer after CNN or not. We experimentally choose $tanh$ function for word-based CNN, and select $Relu$ for other models. We train all models for 15 epochs, the point when validation accuracies get stable.

Table II
CLASSIFICATION ACCURACIES.

Model	Direct	Ensemble
word-CNN	**0.815**	**0.790**
char-RCNN	0.759	0.757
word-RCNN	0.781	0.774
BoW	0.722	0.699

We focus on the intensity of emotions to confirm the effect of direct and ensemble learning. Besides directly predicting 32 classes (we depict it as 'Direct' in Table II), we also propose to first (1) classify 16 emotions, composed of 8 primary emotions and 8 combined emotions mentioned in Section III and then (2) a intensity classifier to further predict the intensity ('Ensemble' in Table II) of these 8 primary emotions. This 'Ensemble' method is more alike a cascaded decision making network that first attempts to separate briefly an input sentence into primary emotions and then predict a fine-grained strength of that emotion. One motivation of building this method is by the observation of shared adverb or adjective modifier words for these emotions, such as *quite/extremely angry/happy*.

B. Results and Discussion

We describe our results on Table II, in which, word-CNN achieves the highest accuracy over other models. We explain it for two reasons: first, the convolution layer in word-based CNN succeeded to focus on characteristic emotional words and adjacent supplemental words in the sentence, that is why word-CNN outperformed the BoW model. Second, compared with char-RCNN and word-RCNN, word-CNN has a relatively simple architecture. Without RNN layers, the model worked effectively on this task because it only needs to keep smaller amount of parameters than other RNN models. The ensemble learning does not show any improvement in all models. We suppose this is because both of the classification errors from two separate network are accumulated. As a result, the overall accuracy is dropped by the two classification errors.

C. Simulate Plutchik's Wheel of Emotions

For eight primary emotions with 24 intensity variants, we applied principal component analysis (PCA) to the sentence features, each extracted by word-CNN. To decrease visual overlap between classes, we visualize in three dimensions. The sentence was all sampled from the training dataset. We depict the results on figure 2. To easily compare our work to Plutchik's wheel of emotions, we make our plot color correspond to Table 1. We can argue that there are smooth flow among *rage*, *anger* and *annoyance* as the value in T2 axis increases larger.

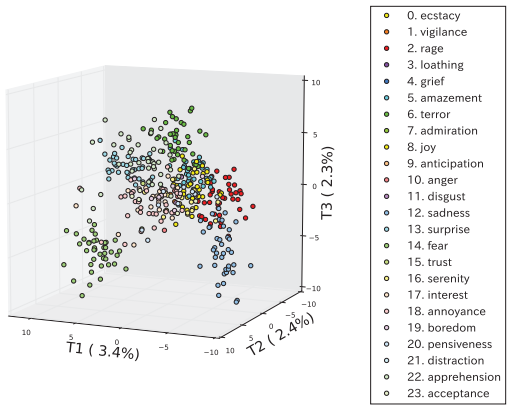

Figure 2. Simulation on wheel of 24 emotions (excluded 8 combined emotions, such as *love*).

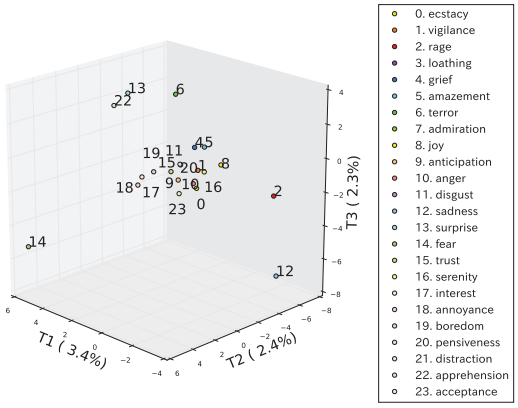

Figure 3. Visualization of mean points of 24 emotions.

Plutchik [7] assumed that distant emotion pairs have opposite color (such as *rage* in red and *fear* in green). Figure 2, however, shows that *rage* (blight red, middle-right side, "2") and *terror* (blight green, top-middle side, "6") are adjacent emotions in PCA's feature space (a *surprise* frequently causes a strong emotion of *angry*), so are *joy* (yellow, middle side, "8") and *sadness* (water blue below, right-bottom side, "12") (the extreme of *joy* is actually *sadness* as exposed by an ancient Chinese idiom *LeJiShengBei*, i.e., *extremely happy leads to sadness*), even though they are not adjacent in the wheel of emotions. Figure 3 shows that most strong primary emotions or emotions with rather larger training sentences (e.g., *rage*, *terror*, *sadness*) are in the corners or relatively far from the central of the coordinate.

VI. Conclusion and Future Work

Taking Japanese as a test language, we have constructed fine-grained SA models with 32 emotions. We focused on building emotional lexicons and sentence-level training data, training word/character-based neural networks, and evaluating the networks. We believe that this is only one step toward fine-grained SA and there are still challenges for building high-quality and large-scale training data and consequently applying the models to real-world applications, such as emotion-driven text/image generating. We take these as our future work.

References

[1] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu, "Emotional chatting machine: Emotional conversation generation with internal and external memory," *CoRR*, vol. abs/1704.01074, 2017.

[2] L. Gong, M. Al Boni, and H. Wang, "Modeling social norms evolution for personalized sentiment classification," in *Proceedings of Association for Computational Linguistics*, 2016, pp. 855–865.

[3] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of Empirical Methods in Natural Language Processing*, 2013, pp. 1631–1642.

[4] J. McAuley, R. Pandey, and J. Leskovec, "Inferring networks of substitutable and complementary products," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 785–794.

[5] D. Tang, B. Qin, and T. Liu, "Aspect level sentiment classification with deep memory network," in *Proceedings of Empirical Methods in Natural Language Processing*, 2016, pp. 214–224.

[6] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word-emotion association lexicon," vol. 29, no. 3, pp. 436–465, 2013.

[7] R. Plutchik, *Emotion: A psychoevolutionary synthesis*. Harpercollins College Division, 1980.

[8] S. Aman and S. Szpakowicz, "Using roget's thesaurus for fine-grained emotion recognition," in *Proceedings of International Joint Conference on Natural Language Processing*, 2008, pp. 312–318.

[9] X. Wu, Y. Kikura, M. Klyen, and Z. Chen, "Sentiment analysis with eight dimensions for emotional chatbots," *Natural Language Process Conference (Japan)*, 2017.

[10] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of Neural Information Processing Systems*, 2013, pp. 3111–3119.

[11] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Computational linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

[12] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of Empirical Methods in Natural Language Processing*, 2014, pp. 1746–1751.

[13] Y. Kim, Y. Jernite, D. Sontag, and A. M. Rush, "Character-aware neural language models." in *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, 2016, pp. 2741–2749.

[14] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980