Multiple Emotions Detection in Conversation Transcripts

Duc-Anh Phan, Hiroyuki Shindo, Yuji Matsumoto
Graduate School of Information and Science
Nara Institute of Science and Technology
8916-5, Takayama, Ikoma, Nara, 630-0192, Japan
{phan.duc_anh.oq3,shindo,matsu}@is.naist.jp

Abstract

In this paper, we present a method of predicting emotions from multi-label conversation transcripts. The transcripts are from a movie dialog corpus and annotated partly by 3 annotators. The method includes building an emotion lexicon bootstrapped from Wordnet following the notion of Plutchik's basic emotions and dyads. The lexicon is then adapted to the training data by using a simple Neural Network to fine-tune the weights toward each basic emotion. We then use the adapted lexicon to extract the features and use them for another Deep Network which does the detection of emotions in conversation transcripts. The experiments were conducted to confirm the effectiveness of the method, which turned out to be nearly as good as a human annotator.

1 Introduction

Along with the trend of "Affective Computing", the task of Emotion Detection in text has received much attention in the recent years. However, very little research has been working on the detection of multiple emotion simultaneously. Instead, most of them make simple assumption that emotions are mutually exclusive and focus on multi-class classification. In fact, the nature of human emotion is complicated: emotions have connections, some are opposite of each other, while some occur together at the same time, resonate and create another emotional state - dyads (Plutchik, 1980)

The survey by Dave and Diwanji (2015) predicted the need for Emotion Detection in *streaming data and the study of emotion flow during chatting*. In this paper, we tackle the simplified version of this task by detecting the emotions in *conversation*. The corpus we used is made of conversations among movie characters, who take turns in the conversation. Those turns are called utterances, which are then manually annotated in a multi-label manner.

Emotion detection in conversation is essentially different from identifying emotions in news headlines (Strapparava and Mihalcea, 2007) or Tweets (Bollen et al., 2011) where each instance is independent of each other. Generally, the expression of Emotion in general depends on the words being used. However, it also quite depends on the grammar structure and syntactic variables such as: negations, embedded sentence, and the type of sentence - question, exclamation, command or statement (Collier, 2014). Therefore, similar to the detection of emotions of sentences in a paragraph, the context information of the whole conversation and what is said in the previous utterance should be taken into consideration. The extraction of context features will be further explained in sub-section 4.3.1

Unlike other works (Li et al., 2015; Wang et al., 2015) where small sets of basic emotions are used, we annotated the dataset using the notion of Plutchik's basic emotion and dyads (1980). This eases the annotators' task since it offers annotators with wider range of emotion labels (8 basic and 23 combinations) to choose from.

Previous research often relied on a list of 6 basic emotions (Ekman et al., 1987) with some variants. However, this notion fails to show conflict side of some emotions. For example, people should not

Figure 1: Plutchik's basic emotion and dyads - image taken from `http://twinklet8.blogspot.jp`

feel happiness and sadness from the same incident altogether. Furthermore, Ekman's basic emotions are the result of observation made on human facial expressions so applying such notion in text classification task seems irrelevant. Newer works relies on dimensional representation using valence-arousal space (Calvo and Mac Kim, 2013; Yu et al., 2015)

Plutchik (1980) suggested 4 axes of bipolar **basic emotions**: Joy - Sadness, Fear - Anger, Trust - Disgust, Surprise - Anticipatation. These primary emotions may blend to form the full spectrum of human emotional experience. The new complex emotions formed by them are called **dyads** (Figure 1). Plutchik's notion reasonably explains the connection between emotions. Some emotions will not occur at the same time since they are on the opposite side of the axis. Complex emotions can also be viewed as combinations of primary ones. The idea enables us to approach emotion detection in a more comprehensive manner. In the future, we may address not only complicating mixture of emotions but also the intensity of each of them.

In this paper, we propose a three steps method for the detection of emotions in conversation: 1)Building Emotion Lexicon from Wordnet (Miller, 1995). 2) Using simple Neural Network to adapt the lexicon to the training data. 3) Using Deep Network with features extracted from adapted lexicon and classify the multi-label corpus.

The remainder of the paper is organized as follows. Section 2 summarizes related work on emotion detection. Section 3 discusses the nature of our dataset and explains the annotating scheme. Section 4 proposes our approach which includes the 3 steps mentioned above. Section 5 evaluates the lexicon, the effectiveness of the adapted lexicon and the proposed method in general. Section 6 gives the conclusion and discusses future work.

2 Related Work

Most of the work in the field tried to define a small set of emotions (D'Mello et al., 2006; Yang et al., 2007) which involved only 3 and 4 emotional states respectively. Another work by Hasegawa et al. (2013) performed a *multi-class* classification on dialog data from Twitter in Japanese. They automatically labeled the obtained dialogs by using emotional expression clues, which is similar to our collocation list explained in sub-section 3.3. We propose a more comprehensive approach by exploiting Plutchik's notion which covers the full spectrum of human emotions to work on challenging multi-label conversation corpus.

Having the same notion, Buitinck et al. (2015) proposed a simple Bag of Words approach and tuned RAKEL for multi-label classification for movie reviews. We go further and work on conversation data where the exchange between characters and the context of the whole dialog are of great importance. The closest to our work is Li et al. (2015) on paragraphs and documents which tried to improve the sentence-level prediction of some special emotions which, due to data sparseness and inherent multi-label classification, were very hard to predict. They incorporated label dependency among labels and context dependency into the graph model to achieve such goal. However, their work is for paragraphs in Chinese. In our case, we take advantages of Deep Neural Network to capture the abstract representation of context information.

Our system is different from previous methods in four main ways:

- Plutchik's notion of primary emotions and dyads is incorporated in our system and provides scalability to address more than just primary emotions if needed in the future.

- We bootstrapped the lexicon and then adapted it to the training data which improved the classification result
- The proposed method includes 2 neural networks, one for adapting the lexicon and the other for multi-label classification of emotions.
- We use a set of manually constructed features instead of word-embedding directly for the Neural Network. The reason for that is further discussed in sub-section 4.3.1

3 Corpus, Dataset & Annotation Scheme

3.1 Movie Dialog Corpus

The Cornell Movie Dialog dataset [1] was originally used for understanding the coordination of linguistic style in dialogs (Danescu-Niculescu-Mizil and Lee, 2011). It includes in total 304,713 utterances (turns in conversation) out of 220,579 conversational exchanges between 10,292 pairs of 9,035 movie characters from 617 movies. The annotating scheme is as follows:

- One utterance may hold zero, one or more emotions at the same time. The list of emotions to assign includes Plutchik's 8 basic emotions and 23 dyads. The sytem will treat the dyads as combination of basic emotions. In case an utterance holds no emotion, it should be annotated with "None"
- The annotators need to assign the whole utterance which may have two or more sentences inside with a set of all emotions expressed inside it. There may be cases where confict emotions according to Plutchik's notion appear simultaneously in the same utterance.

The followings are some statistics of the corpus: total of 11,610 utterances, 10,008 of which are in the training data , 1,602 others are in the testing data, the average number of label per utterance is 1.29. We separated the *training data* which was annotated by only one annotator and the *testing data* which was annotated by all three annotators.

[1] http://www.mpi-sws.org/~cristian/Cornell_Movie-Dialogs_Corpus.html

3.2 Inter-Annotator Agreement

One of the most common Inter-Annotator Agreement measurement is the Kappa statistics (Cohen, 1960). Bhowmick et al. (2008) suggested a Kappa-based measurement for multi-class classification. However, none of them are applicable to our multi-label corpus because their ways of computing causes hypothetical probability of chance agreement P_e to be greater than 1 since there are cases where two or more labels are annotated to a given instance. Therefore, we measure the Kappa statistics for each emotion class and then average them as shown in Table 1. The survey by Artstein and Poesio (2008) suggested that low kappa scores are often observed in multi-label annotating tasks even when the annotators do not make much use of the ability to assign multiple tags.

Some strong emotions: "Anger", "Fear", "Surprise" have better agreement scores as they have indicators such as question marks and excalmation forms. Nevertheless, they are easier for human to identify because they are the basic emotions that we - human inherits from animals. They are the emotions that trigger the "fight or flight" and "stop and examine" response. (Plutchik, 1980)

Due to the time constraint, we had neither the time to show annotators the movies footage nor an adequate amount of sessions to work together and seek a better degree of agreement. Because the annotators only worked with the text data, it was very difficult for them to visualize the situation and make correct judgment.

3.3 ISEAR dataset for Collocation features

We also use ISEAR dataset [2] for the process of producing collocation features. In the ISEAR datase, student respondents, both psychologists and non-psychologists, were asked to report situations in which they had experienced 7 major emotions. Five out of them are completely identical to the basic emotions of Plutchik's . In each case, the questions covered the way they had appraised the situation and how they reacted. Therefore, to our belief, this dataset would provide good collocation features for the 5 identical emotions of our corpus. We mine

[2] http://www.affective-sciences.org/researchmaterial

Emotion class	Kappa Stat
Anger*	0.300
Fear*	0.303
Disgust*	0.127
Trust	0.102
Joy*	0.101
Sadness*	0.131
Surprise	0.575
Anticipation	0.110
Average (by class)	0.219
No. of utterances	1,602

* indicates that these emotions are also in ISEAR dataset

Table 1: Kappa Agreement score.

this dataset for words which frequently appear together with one emotion. If a word also appears in other emotions situation, it loses its place as the indicator toward one specific emotion and we discard it from the collocation list. The use of this collocation list in our work is closely similar to emotional expression clues in (Hasegawa et al., 2013).

4 The Proposed Method

4.1 Bulding Lexicon

Using Lexicon is proven to provide significant improvement in identifying the emotion conveyed by a word (Mohammad, 2012). Therefore, in our case, we built a new lexicon, each lexical item of which displays not only its association with Plutchik's basic emotions but also how strong the association is.

We define the primary emotions and dyads in Plutchik's theories as the seeds of our lexicon. Throughout Wordnet, we search for *synonyms, hypernyms, hyponyms* of the seeds. A reverse lemmatisation is necessary to retrieve related verbs, adjectives and adverbs and their derived forms (verb forms and comparative, superlative adjectives) of the seeds. We keep tracks of the original nouns and the seeds where the new words were derived from (Table 2). Note that sometimes a word was derived from different nouns and seeds, which suggests mixed emotional states.

Each lexical item in the lexicon has a vector of values on each axis of the basic emotions: Joy - Sadness, Fear - Anger, Trust - Disgust, Surprise - Anticipatation . We manually assign the primary emotions with a value vector of 1, 0 or -1 and the dyads with 0.5, 0 or -0.5, depending on the axes they belong. For example, "joy" came from the axis of Joy-Sadness, thus, its vector is [1,0,0,0] while the vector for "sadness" is [-1,0,0,0] (Table 3). The dyad "love" came from primary emotions "joy" and "trust", hence its vector is [0.5,0.5,0,0]. It is to be noted that the minus sign only indicates that the emotion is on the other side of the axis. It is not a suggestion of negative emotion in any case.

Words	Original Nouns - **Seeds**
joy	(primary)- **joy**
sadness	(primary)- **sadness**
fear	(primary)- **fear**
love	(dyad)- **love**
benevolent	benevolence- **love**
worship	worship-**fear** , worship-**love**

Table 2: Wordnet expansion.

In addition, we calculate the wup similarity (Wu and Palmer, 1994) between a new word and the seed it came from, based on the depth of the two senses in the Wordnet taxonomy and that of their Least Common Subsumer.

$$wup(word, seed) = \frac{2 * dep(lcs)}{dep(word) + dep(seed)} \quad (1)$$

We assumed that the higher the similarity, the closer emotional state of the word to the seed. Thus, the value vector of a word is the sum of the products of each seed vector and the similarity between the word and such seed.

$$vector(word) = \sum_{k=1}^{n} vector(seed_k) \times wup(word, seed_k). \quad (2)$$

For example, in the case of the word "worship", we first calculate the wup scores between the word and its two seeds: fear and love (Table 2). Next, they are multiplied by the vectors of the seeds fear-[0,0,1,0] and love-[0.5,0.5,0,0], and then summed up to get the result (Table 3).

4.2 Adapting Lexicon to Training data

We understand that a lexicon bootstrapped from a general domain resource such as Wordnet has its effectiveness limited when it is applied on a specified

Words	J-S	T-D	F-A	S-An
joy	1	0	0	0
sadness	-1	0	0	0
fear	0	0	1	0
love	0.5	0.5	0	0
benevolent	0.47	0.47	0	0
worship	0.14	0.14	0.29	0

Table 3: Value vector of some words. (*J-S: Joy-Sadness, T-D: Trust-Disgust, F-A: Fear-Anger, S-A: Surprise-Anticipation*)

domain. In order to partly solve this problem, we built a simple neural network with one input layer and one output softmax layer.

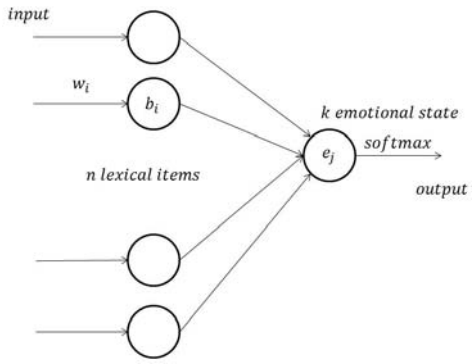

Figure 2: Adapting lexicon for emotional state e_j

The input to the network is the Bag-of-Words features of the training data. We then steps by steps, try to do binary classification on the basic emotion e_j. Let each node of the network be corresponding to each lexical item in the Lexicon. The biases a node b_i are initialized according to the value of each lexical items in the lexicon while the weights w_i are randomly initialized. After each step, we update the biases and weights and then repeat the process for the whole 8 basic emotions. Figure 2 shows the structure of the network. We use the log-likelihood as the cost function for the network input: $C = -\ln(a_y^L)$ where a^L is the output of the final layer and y is the desired output. In the end, we updated the lexicon with new values from the network. We will discuss about the improvements made by the network later in section 5.

4.3 Deep Network for Multi-label Classification

4.3.1 Features Extraction

The process of feature selection for the network is an heuristic one. We initially used a lot of features and then through logistic regression, unimportant features such as the genre of the movie or n-grams features were filtered out.

The core part of the extraction process is to take advantages of the lexicon to transform an utterance to a vector of values expressing the tendency towards each emotion state. This task is done in a rule-based manner (Algorithm 1 and Figure 3). Each word in the utterance is mapped to the lexicon to retrieve the value vector. The representation vector of an utterance is the sum vector of all the word inside it. The negation and word dependency are also taken into account when we calculate the sum with the help of NLTK (Bird et al., 2009) dependency parsing.

Data: Movie Dialogs
Result: Tendency Features
$utterance\_value \leftarrow 0$
foreach *word in utterance* **do**
  $value \leftarrow retrieve\_from\_lexicon(word)$
  $dependencies \leftarrow$
  $check\_dependency(word, utterance)$
  if *value & check_negation(dependencies)* **then**
    $value \leftarrow -value$
  end
  $utterance\_value+ = value$
end

Algorithm 1: Tendency Features extracting algorithm

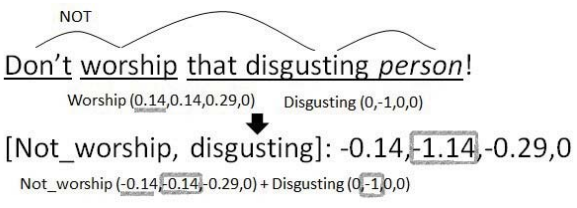

Figure 3: Extracting tendency features

Each utterance in the dataset is presented by the following compact set of 22 features:

1. The sum vector of the current utterance which suggest the *local tendency*.

2. The sum vector of all the utterances in the lexicon that appear in the conversation which provides the *context of the conversation*.

3. The sum vector of the previous utterance in the conversation which also provides the *context of previous exchange* (of what triggered the current emotion).

4. The *polarity* (negative/ positive) score of the sentence.

5. *Features* such as: length, is_it_a_question, is_it_an_exclamatory_sentence.

6. *Collocation features* which indicate the number of appearances of words inside the ISEAR collocations list.

The reason for us to use extracted features is that it is very hard to capture the context of both the conversation and previous exchange using direct word-embedding. While using a recurrent neural network can solve the latter, it is a challenge to address the first. Each conversation has different number of utterances, it may hurt the performance of the system and result in network architecture complexity if we use a non-fixed size window to monitor all the utterances in a same conversation.

4.3.2 Building the Deep Network

The structure of the network is built as shown in Figure 4. The raw input is generalized to produce a small set of features. These features are fed to the network as input layer. We have 2 fully connected hidden layers and an output layer. Since the task is a multi-label classification problem where softmax cannot be used,the output layer is change into sigmoid we add a set of threshold values (one for each basic emotions). Only the labels, whose output values greater than the threshold are considered valid. The thresholds are randomly initialized and then updated after each epochs the same way we updated the biases and weights. In our implementation of the network, Theano (Bastien et al., 2012) was used to take advantages of GPU computing power.

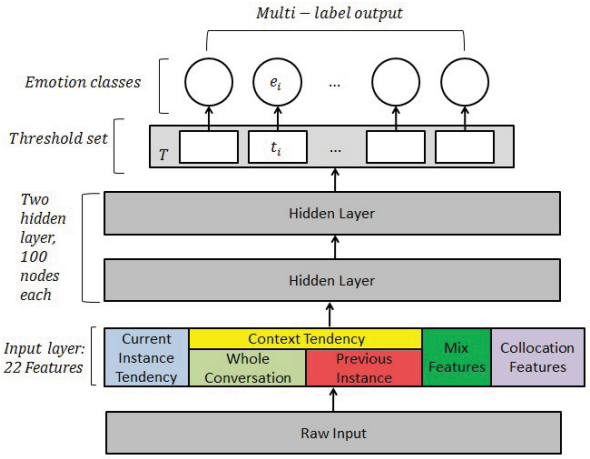

Figure 4: Structure of the Deep Network

The global cost function, similar to Zhang and Zhou (2006), is defined to reward the system for right predictions and severely punish for wrong ones in equation 3.

$$E = \sum_{i}^{m} = \frac{1}{|Y_i||\bar{Y}_i|} \sum_{(k,l)\in Y_i \times \bar{Y}_i} \exp\left(-(c_k^i - c_l^i)\right) \tag{3}$$

Let X be the set of all m instances. Let $Y = \{1, 2, .., Q\}$ be the set of all possible labels, Y_i is the set of true labels for ith instance x_i and $\bar{Y}_i$ is the set of the labels not belong to x_i. Obviously, $Y_i \cup \bar{Y}_i = Y$. We define E as the global cost function of the network. c^i is the set of actual outputs of the model for input x_i, each label has its own output. c_k^i is the output of label k belongs to the set of true labels, $k \in Y_i$. Meanwhile, c_l^i is the output of label l for $l \in \bar{Y}_i$. The difference $c_k^i - c_l^i$ measure the output of the system between the labels, which an instance belong to and which it doesn't. Naturally, we want this difference to be as big as possible.

5 Experiments

5.1 Experiment Setting

Corpus

As mention above, we used the annotated movie dialog corpus for testing our method. For the **gold standard** of the test data, we applied the majority rules on the annotation. If one emotion is annotated by two or more annotators, we accept it as a true label for the utterance.

Evaluation Metrics In our study, 4 common evaluation metrics which have been popularly used in multi-label classification problems (Godbole and Sarawagi, 2004; Li et al., 2015) are employed to measure the performance of our system to the baselines. Let Y_i be set of true labels for a given sample, then Y_i' is the set the labels predicted by a system. Let m be the total number of samples.

1. *Hamming score* or accuracy in multi-label classification, gives the degree of similarity between the ground truth set of labels and the predicted set of labels.

$$Hammingscore = \frac{1}{N}\sum_{i}^{N}\frac{|Y_i \cap Y_i'|}{|Y_i \cup Y_i'|} \quad (4)$$

2. *Precision:* the fraction of correctly predicted labels over all the predicted labels in the set.

$$Precision = \frac{1}{N}\sum_{i}^{N}\frac{|Y_i \cap Y_i'|}{|Y_i'|} \quad (5)$$

3. *Recall:* the fraction of correctly predicted labels over all the true labels in the set.

$$Recall = \frac{1}{N}\sum_{i}^{N}\frac{|Y_i \cap Y_i'|}{|Y_i|} \quad (6)$$

4. *F1-measure:* the harmonic mean of Precision and Recall. In our study, we gave equal importance to Precision and Recall.

$$F1 = \frac{2 * precision * recall}{precision + recall} \quad (7)$$

5.2 Experimental Results

To evaluate the system, we tried to replicate other works and applied them on our new corpus. A similar work is Buitinck et al. (2015) which use the same Plutchik's basic emotions and work on multi-label data. We used similar Meka's [3] RAkEL method and Bag-of-Words approach as the first baseline. We understand that Buitinck et al. (2015)'s system is fine-tuned for their corpus, therefore, it is a little unfair to apply it to our corpus and make comparison. Therefore, the second baseline is Meka's DBPNN which is reported as generally having better accuracy than RAkEL (Fernandez-Gonzalez et al., 2015).

We decided that the most important baseline is the human annotation. We calculated the evaluation metrics based on the annotation made by each annotator against the gold standard and averaged the result by the total number of annotators. Another baseline is our own system using the lexicon before adaptation. Figure 5 compares the performance of our system to the baselines.

vs. Bag-of-Words Approaches: Our system, with and without lexicon adaptation performed remarkably better than the simple approaches using Meka's DBPNN and RAkEL. It exceeded the better DBPNN significantly in *Hamming Score* by 7.28 and 7.19, *Recall* by 12.85 and 5.95, *F1-measure* by 7.33 and 4.33 respectively. We argue that the context features played as an important factor here.

Lexicon Adaptation vs. No Adaptation: We can clearly see the improvements made by the adaptation on our system in *Recall and F1-measure*,which are increased by 6.9 and 3.0. This confirmed the necessity of the adaptation step.

vs. Human Annotator: This is the most important baseline, which explains how well our system performs in comparison with a Human Annotator. Please note that these values are averaged by the total number of annotators after the judgment made by each annotator are compared to the gold standard. Our system is slightly worse than a Human Annotator in all 4 metrics by 0.43 in *Hamming Score*, 0.79, 1.67, 1.69 in *Precision, Recall and F1-measure* respectively.

These results confirmed the performance of our method which is slightly worse than such of an human annotator. On the other hand, our method is more efficient than simple Bag-of-Words approaches. We also confirmed the improvement made by the Lexicon Adaptation step to our system.

Classification result for each emotion class: Table 4 shows the distribution of emotion classes and reports the classification result of each emotion class in the corpus. Imbalance can be seen among classes in the corpus. We observed the expected "All-No-Recurrence" problem for minority classes of Joy and Sadness (high accuracy and near zero F1) as the corpus is unbalanced. "Surprise" is the class with the highest Agreement score (Table 1), it also

[3] http://meka.sourceforge.net/#about

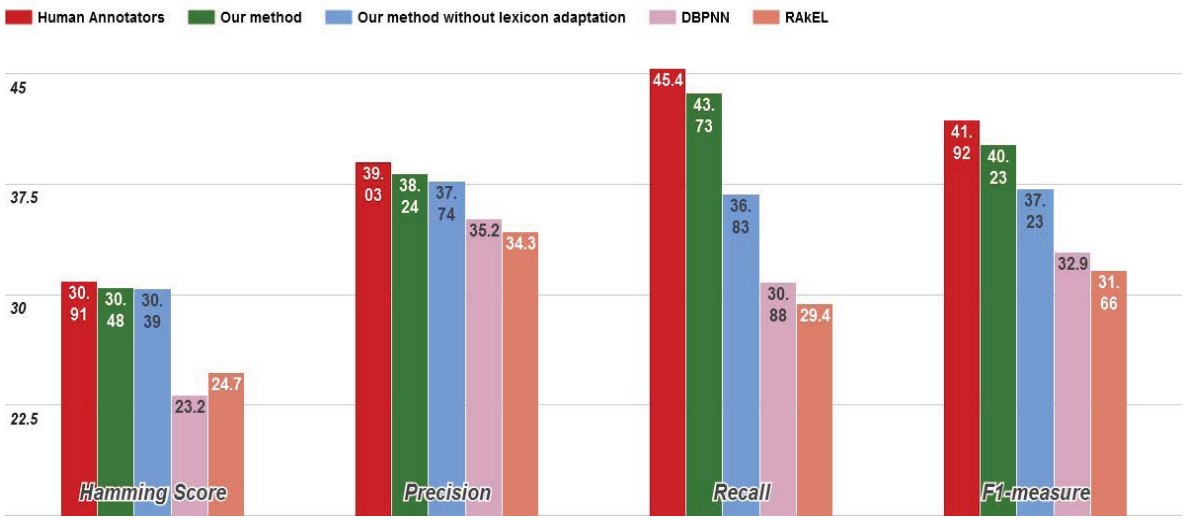

Figure 5: Evaluation of the system vs four baselines: 1) Human Annotators, 2) The system without lexicon adaptation, 3) DBPNN 4) RAkEL

Emotion	Percentage	Accuracy	F1
Anger	18.48%	0.615	0.265
Fear	16.52%	0.70	0.285
Disgust	16.52%	0.65	0.275
Trust	13.35%	0.69	0.313
Joy	5.56%	0.92	0.01
Sadness	5.18%	0.92	0.01
Surprise	17.01%	0.605	0.34
Anticipation	38.72%	0.395	0.27

Table 4: Accuracy and F1 for each emotion class.

achieves the highest F1 among other classes. While "Anticipation" class is a dominant class in the corpus, it suffers from low Agreement score. As a consequence, the classification result for this class is also not high. From this result, we can hope that in the future when movie footage are included, not only the agreement score but the system performance will also go up as well.

6 Conclusion & Future Work

In this paper, we propose our method of detecting and classifying emotions from a conversation corpus. The corpus is a set of movie dialogs annotated with multi-label emotions following Plutchik's notion of basic emotions and dyads. Our method involves building a lexicon from Wordnet using some seed emotion words, adapting the lexicon to the corpus, extracting a feature set from the input and classifying the emotions accordingly with the help of a deep neural network. The experiments show that our method's power to detecting emotion is comparable to that of a human annotator. However, one may argue that the disagreement among annotators may have affected the result. As discussed above, we hope to solve this problem by including the movies' footage in our annotating scheme.

At the time of the submission, we are adding the footages as well as improving annotating scheme to have higher Kappa statistics and evaluate again our method. Once finished, the corpus will be published for other researchers to use. In the future, we also want to further exploit the method by incorporating emotion detection on voices and images and monitoring complex emotions other than the basic ones and their intensity.

Acknowledgments

This research was supported by CREST project of Japan Science and Technology Agency. We are grateful to our colleagues from Computational Linguistics Lab, NAIST, Japan who provided insights and expertise that greatly assisted the research. We also thank the reviewers for their valuable comments that further improved our work.

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian J. Goodfellow, Arnaud Bergeron, Nicolas Bouchard, and Yoshua Bengio. 2012. Theano: new features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop.

Plaban Kr Bhowmick, Pabitra Mitra, and Anupam Basu. 2008. An agreement measure for determining inter-annotator reliability of human judgements on affective text. In *Proceedings of the Workshop on Human Judgements in Computational Linguistics*, pages 58–65. Association for Computational Linguistics.

S Bird, E Klein, and E Loper. 2009. Nltk book.

Johan Bollen, Huina Mao, and Alberto Pepe. 2011. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena.

Lars Buitinck, Jesse Van Amerongen, Ed Tan, and Maarten de Rijke. 2015. Multi-emotion detection in user-generated reviews. In *Advances in Information Retrieval*, pages 43–48. Springer.

Rafael A Calvo and Sunghwan Mac Kim. 2013. Emotions in text: dimensional and categorical models. *Computational Intelligence*, 29(3):527–543.

J Cohen. 1960. Kappa: Coefficient of concordance. *Educ. Psych. Measurement*, 20:37.

Gary Collier. 2014. *Emotional expression*. Psychology Press.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*.

Saurin Dave and Hiteishi Diwanji. 2015. Trend analysis in social networking using opinion mining a survey.

Sidney K D'Mello, Scotty D Craig, Jeremiah Sullins, and Arthur C Graesser. 2006. Predicting affective states expressed through an emote-aloud procedure from autotutor's mixed-initiative dialogue. *International Journal of Artificial Intelligence in Education*, 16(1):3–28.

Paul Ekman, Wallace V Friesen, Maureen O'Sullivan, Anthony Chan, Irene Diacoyanni-Tarlatzis, Karl Heider, Rainer Krause, William Ayhan LeCompte, Tom Pitcairn, Pio E Ricci-Bitti, et al. 1987. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology*, 53(4):712.

Pablo Fernandez-Gonzalez, Concha Bielza, and Pedro Larranaga. 2015. Multidimensional classifiers for neuroanatomical data. In *ICML Workshop on Statistics, Machine Learning and Neuroscience (Stamlins 2015)*.

Shantanu Godbole and Sunita Sarawagi. 2004. Discriminative methods for multi-labeled classification. In *Advances in Knowledge Discovery and Data Mining*, pages 22–30. Springer.

Takayuki Hasegawa, Nobuhiro Kaji, Naoki Yoshinaga, and Masashi Toyoda. 2013. Predicting and eliciting addressee's emotion in online dialogue. In *ACL (1)*, pages 964–972.

Shoushan Li, Lei Huang, Rong Wang, and Guodong Zhou. 2015. Sentence-level emotion classification with label and context dependence. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1045–1053, Beijing, China, July. Association for Computational Linguistics.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Saif Mohammad. 2012. Portable features for classifying emotional text. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 587–591. Association for Computational Linguistics.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. *Theories of emotion*, 1:3–31.

Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 70–74. Association for Computational Linguistics.

Zhongqing Wang, Sophia Lee, Shoushan Li, and Guodong Zhou. 2015. Emotion detection in code-switching texts via bilingual and sentimental information. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 763–768, Beijing, China, July. Association for Computational Linguistics.

Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics.

Changhua Yang, Kevin Hsin-Yih Lin, and Hsin-Hsi Chen. 2007. Emotion classification using web blog corpora. In *Web Intelligence, IEEE/WIC/ACM International Conference on*, pages 275–278. IEEE.

Liang-Chih Yu, Jin Wang, K Robert Lai, and Xue-jie Zhang. 2015. Predicting valence-arousal ratings of words using a weighted graph method. *Volume 2: Short Papers*, page 788.

Min-Ling Zhang and Zhi-Hua Zhou. 2006. Multilabel neural networks with applications to functional genomics and text categorization. *Knowledge and Data Engineering, IEEE Transactions on*, 18(10):1338–1351.