Emotion Classification Using Massive Examples Extracted from the Web

Ryoko TOKUHISA[†‡]
[†]Toyota Central R&D Labs., INC.
Nagakute Aichi JAPAN
tokuhisa@mosk.tytlabs.co.jp

Kentaro INUI[‡] **Yuji MATSUMOTO**[‡]
[‡]Nara Institute of Science and Technology
Ikoma Nara JAPAN
{ryoko-t,inui,matsu}@is.naist.jp

Abstract

In this paper, we propose a data-oriented method for inferring the emotion of a speaker conversing with a dialog system from the semantic content of an utterance. We first fully automatically obtain a huge collection of emotion-provoking event instances from the Web. With Japanese chosen as a target language, about 1.3 million emotion provoking event instances are extracted using an emotion lexicon and lexical patterns. We then decompose the emotion classification task into two sub-steps: sentiment polarity classification (coarse-grained emotion classification), and emotion classification (fine-grained emotion classification). For each subtask, the collection of emotion-proviking event instances is used as labelled examples to train a classifier. The results of our experiments indicate that our method significantly outperforms the baseline method. We also find that compared with the single-step model, which applies the emotion classifier directly to inputs, our two-step model significantly reduces sentiment polarity errors, which are considered fatal errors in real dialog applications.

1 Introduction

Previous research into human-computer interaction has mostly focused on task-oriented dialogs, where the goal is considered to be to achieve a given task as precisely and efficiently as possible by exchanging information required for the task through dialog (Allen et al., 1994, etc.). More recent research (Foster, 2007; Tokuhisa and Terashima, 2006, etc.), on the other hand, has been providing evidence for the importance of the affective or emotional aspect in a wider range of dialogic contexts, which has been largely neglected in the context of task-oriented dialogs.

A dialog system may be expected to serve, for example, as an *active listening* [1] partner of an elderly user living alone who sometimes wishes to have a chat. In such a context, the dialog system is expected to understand the user's emotions and sympathize with the user. For example, given an utterence *I traveled far to get to the shop, but it was closed* from the user, if the system could infer the user's emotion behind it, it would know that it would be appropriate to say *That's too bad* or *That's really disappointing*. It can be easily imagined that such affective behaviors of a dialog system would be beneficial not only for active listening but also for a wide variety of dialog purposes including even task-oriented dialogs.

To be capable of generating sympathetic responses, a dialog system needs a computational model that can infer the user's emotion behind his/her utterence. There have been a range of studies for building a model for classifying a user's emotions based on acoustic-prosodic features and facial expressions (Pantic and Rothkrantz, 2004, etc.). Such methods are, however, severely limited in that they tend to work well only when the user expresses his/her emotions by "exaggerated"

[1] Active listening is a specific communication skill, based on the work of psychologist Carl Rogers, which involves giving free and undivided attention to the speaker (Robertson, 2005).

Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 881–888
Manchester, August 2008

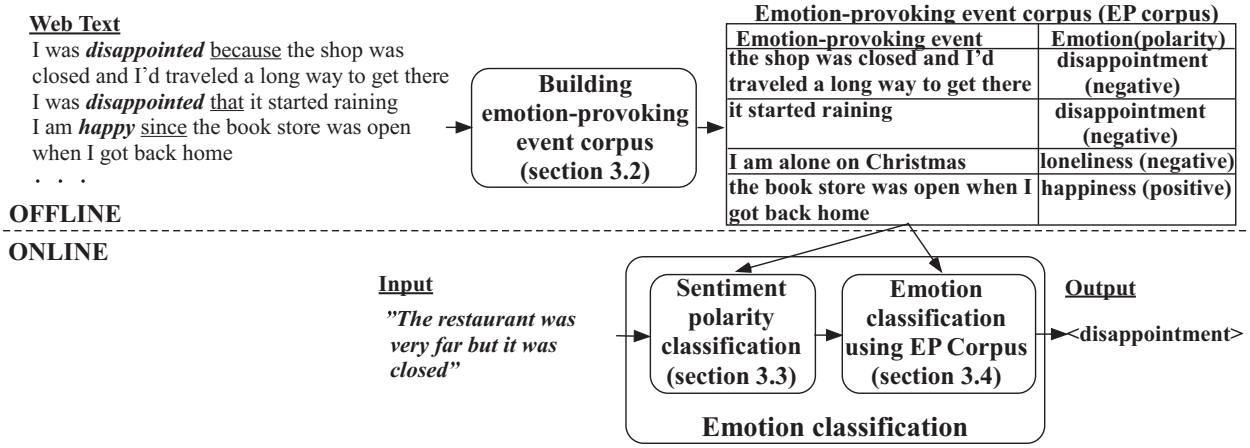

Figure 1: Overview of our approach to emotion classification

prosodic or facial expressions. Furthermore, what is required in generating sympathetic responses is the identification of the user's emotion in a finer grain-size. For example, in contrast to the above example of *disappointing*, one may expect the response to *My pet parrot died yesterday* should be *That's really sad*, wheras the response to *I may have forgotten to lock my house* should be *You're worried about that*.

In this paper, we address the above issue of emotion classification in the context of human-computer dialog, and demonstrate that massive examples of emotion-provoking events can be extracted from the Web with a reasonable accuracy and those examples can be used to build a semantic content-based model for fine-grained emotion classification.

2 Related Work

Recently, several studies have reported about dialog systems that are capable of classifying emotions in a human-computer dialog (Batliner et al., 2004; Ang et al., 2002; Litman and Forbes-Riley, 2004; Rotaru et al., 2005). ITSPOKE is a tutoring dialog system, that can recognize the user's emotion using acoustic-prosodic features and lexical features. However, the emotion classes are limited to *Uncertain* and *Non-Uncertain* because the purpose of ITSPOKE is to recognize the user's problem or discomfort in a tutoring dialog. Our goal, on the other hand, is to classify the user's emotions into more fine-grained emotion classes.

In a more general research context, while quite a few studies have been presented about opinion mining and sentiment analysis (Liu, 2006), research into fine-grained emotion classification has emerged only recently. There are two approaches commonly used in emotion classification: a rule-based approach and a statistical approach. Masum et al. (2007) and Chaumartin (2007) propose a rule-based approach to emotion classification. Chaumartin has developed a linguistic rule-based system, which classifies the emotions engendered by news headlines using the WordNet, SentiWordNet, and WordNet-Affect lexical resources. The system detects the sentiment polarity for each word in a news headline based on linguistic resources, and then attempts emotion classification by using rules based on its knowledge of sentence structures. The recall of this system is low, however, because of the limited coverage of the lexical resources. Regarding the statistical approach, Kozareva et al. (2007) apply the theory of (Hatzivassiloglou and McKeown, 1997) and (Turney, 2002) to emotion classification and propose a method based on the co-occurrence distribution over content words and six emotion words (e.g. joy, fear). For example, *birthday* appears more often with *joy*, while *war* appears more often with *fear*. However, the accuracy achieved by their method is not practical in applications assumed in this paper. As we demonstrate in Section 4, our method significantly outperforms Kozareva's method.

3 Emotion Classification

3.1 The basic idea

We consider the task of emotion classification as a classification problem where a given input sentence (a user's utterance) is to be classified either into such 10 emotion classes as given later in Table 1 or as ⟨neutral⟩ if no emotion is involved in the input. Since it is a classification problem, the task should be approached straightforwardly in a vari-

Table 1: Distribution of the emotion expressions and examples

Sentiment Polarity	10 Emotion Classes	Emotion lexicon (349 Japanese emotion words)	
		Total	Examples
Positive	happiness	90	嬉しい (happy), 狂喜 (joyful), 喜ぶ (glad), 歓ぶ (glad)
	pleasantness	7	楽しい (pleasant), 楽しむ (enjoy), 楽しめる (can enjoy)
	relief	5	安心 (relief), ほっと (relief)
Negative	fear	22	恐い (fear), 怖い (fear), 恐ろしい (frightening)
	sadness	21	悲しい (sad), 哀しい (sad), 悲しむ (feel sad)
	disappointment	15	がっかり (lose heart), がっくり (drop *one's* head)
	unpleasantness	109	嫌 (disgust), 嫌がる (dislike), 嫌い (dislike)
	loneliness	15	寂しい (lonely), 淋しい (lonely), わびしい (lonely)
	anxiety	17	不安 (anxiety), 心配 (anxiety), 気かかり (worry)
	anger	48	腹立たしい (angry), 腹立つ (get angry), 立腹 (angry)

ety of machine learning-based methods if a sufficient number of labelled examples were available. Our basic idea is to learn what emotion is typically provoked in what situation, from massive examples that can be collected from the Web. The development of this approach and its subsequent implementation has forced us to consider the following two issues.

First, we have to consider the quantity and accuracy of emotion-provoking examples to be collected. The process we use to collect emotion-provoking examples is illustrated in the upper half of Figure 1. For example, from the sentence *I was disappointed because the shop was closed and I'd I traveled a long way to get there*, pulled from the Web, we learn that the clause *the shop was closed and I'd traveled a long way to get there* is an example of an event that provokes ⟨disappointment⟩. In this paper, we refer to such an example as an *emotion-provoking event* and a collection of event-provoking events as an *emotion-provoking event corpus (an EP corpus)*. Details are described in Section 3.2.

Second, assuming that an EP corpus can be obtained, the next issue is how to use it for our emotion classification task. We propose a method whereby an input utterance (sentence) is classified in two steps, sentiment polarity classification followed by fine-grained emotion classification as shown in the lower half of Figure 1. Details are given in Sections 3.3 and 3.4.

3.2 Building an EP corpus

We used ten emotions *happiness, pleasantness, relief, fear, sadness, disappointment, unpleasantness, loneliness, anxiety, anger* in our emotion classification experiment. First, we built a hand-crafted lexicon of emotion words classified into the ten emotions. From the Japanese Evaluation Expression Dictionary created by Kobayashi et al. (2005), we identified 349 emotion words based

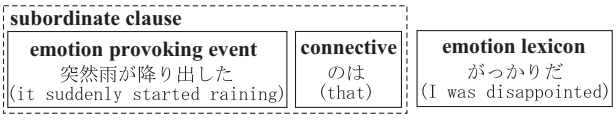

Figure 2: An example of a lexico-syntactic pattern

Table 2: Number of emotion-provoking events

10 Emotions	EP event	10 Emotions	EP event
happiness	387,275	disappoint-ment	106,284
pleasantness	209,682	unpleasantness	396,002
relief	46,228	loneliness	26,493
fear	49,516	anxiety	45,018
sadness	31,369	anger	8,478

on the definition of emotion words proposed by Teramura (1982). The distribution is shown in Table 1 with major examples.

We then went on to find sentences in the Web corpus that possibly contain emotion-provoking events. A subordinate clause was extracted as an emotion-provoking event instance if (a) it was subordinated to a matrix clause headed by an emotion word and (b) the relation between the subordinate and matrix clauses is marked by one of the following eight connectives: ので, から, ため, て, のは, のが, ことは, ことが. An example is given in Figure 2. In the sentence "突然雨が降り出したのは がっかりだ *(I was disappointed that it suddenly started raining)*", the subordinate clause "突然雨が降り出した *(it suddenly started raining)*" modifies "がっかりだ *(I was disappointed)*" with the connective "のは *(that)*". In this case, therefore, the event mention "突然雨が降り出した *(it suddenly started raining)*" is learned as an event instance that provokes⟨disappointment⟩.

Applying the emotion lexicon and the lexical patterns to the Japanese Web corpus (Kawahara and Kurohashi, 2006), which contains 500 million sentences, we were able to collect about 1.3 million events as causes of emotion. The distribution is shown in Table 2.

Tables 3 and 4 show the results of our evalua-

Table 4: Examples from in the EP corpus

EP-Corpus			Result of evaluation	
Emotion-provoking Event	Emotion word	10 Emotions (P/N)	Polarity	Emotion
花持ちが悪い (A flower died quickly)	残念だ (diappointed)	⟨disappointment(N)⟩	Correct	Correct
敵が多い (There are a lot of enemies)	飽きる (lose interest)	⟨unpleasantness(N)⟩	Correct	Context-dep.
ちんげん菜が多い (There is a lot of Chinese cabbage)	嬉しい (happy)	⟨happiness(P)⟩	Context-dep.	Context-dep.
ジュースが飲みたい (I would like to drink orange juice)	大変だ (terrible)	⟨unpleasantness(N)⟩	Error	Error

Table 3: Correctness of samples from the EP corpus

	Polarity	Emotion
Correct	1140 (57.0%)	988 (49.4%)
Context-dep.	678 (33.9%)	489 (24.5%)
Error	182 (9.1%)	523 (26.2%)

tion for the resultant EP corpus. One annotator, who was not the developer of the EP corpus, evaluated 2000 randomly chosen events. The "Polarity" column in Table 3 shows the results of evaluating whether the sentiment polarity of each event is correctly labelled, whereas the"Emotion" column shows the correctness at the level of the 10 emotion classes. The correctness of each example was evaluated as exemplified in Table 4. *Correct* indicates a correct example, *Contex-dep.* indicates a context-dependent example, and *Error* is an error example. For example, in the case of *There are a lot of enemies* in Table 4, the "Polarity" is *Correct* because it represents a negative emotion. However, its emotion class⟨unpleasantness⟩ is judged *Context-dep*.

As Table 3 shows, the Sentiment Polarity is correct in 57.0% of cases and partially correct (Correct + Context-dep.) in 90.9% of cases. On the other hand, the Emotion is correct in only 49.4% of cases and partially correct in 73.9% of cases. These figures may not seem very impressive. As far as its impact on the emotion classification accuracy is concerned, however, the use of our EP corpus, which requires no supervision, makes remarkable improvements upon Kozareva et al. (2007)'s unsupervised method as we show later.

3.3 Sentiment Polarity Classification

Given the large collection of emotion-labelled examples, it may seem straightforward to develop a trainable model for emotion classification. Before moving on to emotion classification, however, it should be noted that a user's input utterance may not involve any emotion. For example, if the user gives an utterance *I have a lunch at the school cafeteria every day*, it is not appropriate for the system to make any sympathetic response. In such a case, the user's input should be classified as ⟨neutral⟩.

The classification between *emotion-involved* and *neutral* is not necessarily a simple problem, however, because we have not found yet any practical method for collecting training examples of the class⟨neutral⟩. We cannot rely on the analogy to the pattern-based method we have adopted to collect emotion-provoking events — there seems no reliable lexico-syntactic pattern for extracting neutral examples. Alternatively, if the majority of the sentences on the Web were neutral, one would simply use a set of randomly sampled sentences as labelled data for ⟨neutral⟩. This strategy, however, does not work because neutral sentences are not the majority in real Web texts. As an attempt, we collected 1000 sentences randomly from the Web and investigated their distribution of sentiment polarity. The results, shown in Table 5, revealed that the ratio of neutral events was unexpectedly low. These results indicate the difficulty of collecting neutral events from Web documents.

Taking this problem into account, we adopt a two-step approach, where we first classify a given input into three sentiment polarity classes, either positive, negative or neutral, and then classify only those judged positive or negative into our 10 fine-grained emotion classes. In the first step, i.e. sentiment polarity classification, we use only the positive and negative examples stored in the EP corpus and assume sentence to be neutral if the output of the classification model is near the decision boundary. There are additional advantages in this approach. First, it is generally known that performing fine-grained classification after coarse classification often provides good results particularly when the number of the classes is large. Second, in the context of dialog, a misunderstanding the user's emotion at the sentiment polarity level would be a disaster. Imagine that the system says *You must be happy* when the user in fact feels sad. As we show in Section 4.2, such fatal errors can be reduced by taking the two-step approach.

Table 5: Distribution of the Sentiment polarity of sentences randomly sampled from the Web

Sentiment Polarity	Number	Ratio
positive	650	65.0%
negative	153	15.3%
neutral	117	11.7%
Context-dep.	80	8.0%

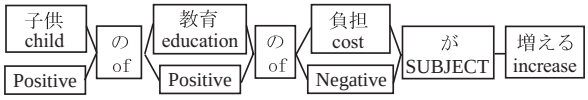

Figure 3: An example of a word-polarity lattice

Various methods have already been proposed for sentiment polarity classification, ranging from the use of co-occurrence with typical positive and negative words (Turney, 2002) to bag of words (Pang et al., 2002) and dependency structure (Kudo and Matsumoto, 2004). Our sentiment polarity classification model is trained with SVMs (Vapnik, 1995), and the features are {1-gram, 2-gram, 3-gram} of words and the sentiment polarity of the words themselves. Figure 3 illustrates how the sentence "子供の教育の負担が増える *(The cost of educating my child increases)*" is encoded to a feature vector. Here we assume the sentiment polarity of the "子供 *(child)*" and "教育 *(education)*" are positive, while the "負担 *(cost)*" is negative. These polarity values are represented in parallel with the corresponding words, as shown in Figure 3. By expanding {1-gram, 2-gram, 3-gram} in this lattice representation, the following list of features are extracted: 子供 *(child)*, *Positive*, 子供 *(child)*-の *(of)*, *Positive*-の *(of)*, 子供 *(child)*-の *(of)*-教育 *(education)*, *etc.*. The polarity value of each word is defined in our sentiment polarity dictionary, which includes 1880 positive words and 2490 negative words. To create this dictionary, one annotator identified positive and negative words from the 50 thousand most frequent words sampled from the Web. Table 6 shows some examples.

3.4 Emotion Classification

For fine-grained emotion classification, we propose a k-nearest-neighbor approach (kNN) using the EP corpus.

Given an input utterance, the kNN model retrieves k-most similar labelled examples from the EP corpus. Given the input *The restaurant was very far but it was closed* as Figure 1, for example, the kNN model finds similar labelled examples, say, labelled example {the shop was closed and I'd traveled far to get there} in the EP corpus.

Table 6: Examples of positive and negative words

P	子供 (child), 夏休み (summer vacation), 役立つ (useful), 成功する (succeed)
N	負担 (cost), 難しい (difficult), 難しい (difficult), 失敗する (failure)

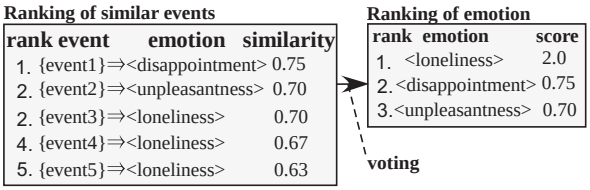

Figure 4: Emotion Classification by kNN (k=5)

For the similarity measure, we use cosine similarity between bag-of-words vectors; $sim(I, EP) = \frac{I \cdot EP}{|I||EP|}$ for input sentence I and an emotion-provoking event EPin the EP corpus. The score of each class is given by the sum of its similarity scores.

An example is presented in Figure 4. The emotion of the most similar event is ⟨disappointment⟩, that of the second-most similar event is ⟨unpleasantness⟩ tied with ⟨loneliness⟩. After calculating the sum for each emotion, the system outputs ⟨loneliness⟩ as the emotion for the input I because the score for ⟨loneliness⟩ is the highest.

4 Experiments

4.1 Sentiment polarity classification

We conducted experiments on sentiment polarity classification using the following two test sets:

TestSet1: The first test set was a set of utterances which 6 subject speakers produced interacting with our prototype dialog system. This data include 31 positive utterances, 34 negative utterances, and 25 neutral utterances.

TestSet2: For the second test set, we used the 1140 samples that were judged *Correct* with respect to sentiment polarity in Table 3. 491 samples (43.1%) were positive and 649 (56.9%) were negative. We then added 501 neutral sentences newly sampled from the Web. These samples are disjoint from the EP corpus used for training classifiers.

For each test set, we tested our sentiment polarity classifier in both the two-class (positive/negative) setting, where only positive or negative test samples were used, and the three-class (positive/negative/neutral) setting. The performance was evaluated in F-measure.

Table 7: F-values of sentiment polarity classification (positive/negative)

	TestSet1		TestSet2	
	Pos	Neg	Pos	Neg
Word	0.839	0.853	0.794	0.842
Word + Polarity	0.833	0.857	0.793	0.841

Table 8: F-values of sentiment polarity classification (positive/negative/neutral)

	TestSet1		TestSet2	
	Pos	Neg	Pos	Neg
Word	0.743	0.758	0.610	0.742
Word + Polarity	0.758	0.769	0.610	0.742

Table 7 shows the results for the two-class setting, whereas Table 8 shows the results for the three-class. "Word" denotes the model trained with only word n-gram features, whereas "Word+Polarity" denotes the model trained with n-gram features extracted from a word-polarity lattice (see Figure 3).

The results shown in Table 7 indicate that both the "Word" and "Word+Polarity" models are capable of separating positive samples from negative ones at a high level of accuracy. This is an important finding, given the degree of the correctness of our EP corpus. As we have shown in Table 3, only 57% of samples in our EP corpus are "exactly" correct in terms of sentiment polarity. The figures in Table 7 indicate that context-dependent samples are also useful for training a classifier. Table 7 also indicates that no significant difference is found between the "Word" and "Word+Polarity" models. In fact, we also examined another model which used dependency-structure information as well; however, no significant gain was achieved. From these results, we speculate that, as far as the two-class sentiment polarity problem is concerned, word n-gram features might be sufficient if a very large set of labelled data are available.

On the other hand, Table 8 indicates that the three-class problem is much harder than the two-class problem. Specifically, positive sentences tend to be classified as neutral. This method has to be improved in future models.

4.2 Emotion classification

For fine-grained emotion classification, we used the following three test sets:

TestSet1 (2p, best): Two annotators were asked to annotate each positive or negative sentence in TestSet1 with one of the 10 emotion classess. The annotators chose only one emotion class even if they thought several emotions would fit a sentence. Some examples are shown in Table 9. The inter-annotator agreement is κ=0.76 in the kappa statistic (Cohen, 1960). For sentences annotated with two different labels (i.e. in the cases where the two annotators disagreed with), both labels were considered correct in the experiments — a model's answer was considered correct if it was identical with either of the two labels.

TestSet1 (1p, acceptable): One of the above two annotators was asked to annotate each positive or negative sentence in TestSet1 with all the emotions involved in it. The number of emotions for a positive sentence was 1.48 on average, and 2.47 for negative sentences. Table 10 lists some examples. In the experiments, a model's answer was considered correct if it was identical with one of the labelled classes.

TestSet2: For TestSet2, we used the results of our judgments on the correctness for estimating the quality of the EP corpus described in Section 3.2.

In the experiments, the following two models were compared:

Baseline: The baseline model simulates the method proposed by (Kozareva et al., 2007). Given an input sentence, their model first estimates the pointwise mutual information (PMI) between each content word cw_j included in the sentence and emotion expression $e \in$ {*anger, disgust, fear, joy, sudness, surprise*} by $PMI(e, cw) = \log \frac{hits(e,cw)}{hits(e)hits(cw)}$, where $hits(x)$ is a hit count of word(s) *x* on a Web search engine. The model then calculates the score of each emotion class E_i by summing the PMI scores between each content word cw_j in the input and emotion expression e_i corresponding to that emotion class: $score(E_i) = \sum_j PMI(e_i, cw_j)$. Finally, the model chooses the best scored emotion class as an output. For our experiments, we selected the following 10 emotion expressions:

嬉しい (happy), 楽しい (pleased), 安心 (relieved), 恐い (affraid), 悲しい (sad), 残念 (disappointed), 嫌 (hate), 寂しい (lonely), 不安 (anxious), 腹立たしい (angry)

For hit counts, we used the Google search engine.

Table 9: Examples of TestSet1 (2p, best)

	Annotator A	Annotator B
クリスマスにプレゼントをもらった (I got a Christmas present)	⟨happiness⟩	⟨happiness⟩
友達の家に遊びに行く (I'm going to go to my friend's house)	⟨pleasantness⟩	⟨pleasantness⟩
花見に行ったら突然雨が降り出した (It rained suddenly when I went to see the cherry blossoms)	⟨sadness⟩	⟨sadness⟩
渋滞でほとんど動かない (My car can't move because of the traffic jam)	⟨unpleasantness⟩	⟨anger⟩

Table 10: Examples of TestSet1 (1p, acceptable)

	Annotator A
クリスマスにプレゼントをもらった (I got a Christmas present)	⟨happiness⟩
友達の家に遊びに行く (I'm going to go to my friend's house)	⟨pleasantness⟩, ⟨happiness⟩
花見に行ったら突然雨が降り出した (It rained suddenly when I went to see the cherry blossoms)	⟨anger⟩, ⟨sad⟩, ⟨unpleasantness⟩, ⟨disappointment⟩
渋滞でほとんど動かない (My car can't move because of the traffic jam)	⟨unpleasantness⟩, ⟨anger⟩

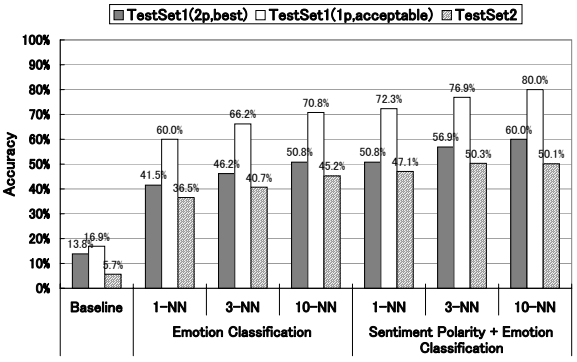

Figure 5: Results of emotion classification

k-NN: We tested the 1-NN, 3-NN and 10-NN models. In each model, we examined a single-step emotion classification and two-step emotion classification. In the former method, the kNN model retrieves k-most similar examples from the all of the EP corpus. In the latter method, when the sentiment polarity of the input utterance has obtained by the sentiment polarity classifier, the kNN model retrieves similar examples from only the examples whose sentiment polarity are the same as the input utterance in the EP corpus.

The results are shown in Figure 5. "Emotion Classification" denotes the single-step models, whereas "Sentiment Polarity + Emotion Classification" denotes the two-step models.

An important observation from Figure 5 is that our models remarkably outperformed the baseline. Apparently, an important motivation behind Kozareva et al. (2007)'s method is that it does not require any manual supervion. However, our models, which rely on emotion-provoking event instances, are also totally unsupervised — no supervision is required to collect emotion-provoking event instances. Given this commonality between the two methods, the superiority of our method in accuracy can be considered as a crucial advantage.

Regarding the issue of single-step vs. two-step, Figure 5 indicates that the two-step models tended to outperform the single-step models for all the test set. A paired t-test for TestSet2, however, did not reach significance [2]. So we next examined this issue in further detail.

As argued in Section 3.3, in the context of human-computer dialog, a misunderstanding of the user's emotion at the level of sentiment polarity would lead to a serious problem, which we call a *fatal error*. On the other hand, misclassifying a case of⟨happiness⟩ as, for example,⟨pleasantness⟩ may well be tolerable. Table 11 shows the ratio of fatal errors for each model. For TestSet2, the single-step 10-NN model made fatal errors in 30% of cases, while the two-step 10-NN model in only 17%. This improvement is statistically significant ($p<0.01$).

5 Conclusion

In this paper, we have addressed the issue of emotion classification assuming its potential applications to be human-computer dialog system including active-listening dialog. We first automatically collected a huge collection, as many as 1.3M, of emotion-provoking event instances from the Web. We then decomposed the emotion classification task into two sub-steps: sentiment polarity classification and emotion classification. In sentiment polarity classification, we used the EP-corpus as training data. The results of the polarity classification experiment showed that word n-gram features alone are more or less sufficient to classify positive and negative sentences when a very large amount of training data is available. In the emotion classification experiments, on the other hand,

[2]The data size of TestSet1 was not sufficient for statistical significance test

Table 11: Fatal error rate in emotion classification experiments

	Baseline	Emotion Classification			Sentiment Polarity + Emotion Classification
		1-NN	3-NN	10-NN	
TestSet1	49.2%	29.2%	26.2%	24.6%	15.4%
TestSet2	41.5%	37.6%	32.8%	30.0%	17.0%

we adopted the k-nearest-neighbor (kNN) method. The results of the experiments showed that our method significantly outperformed the baseline method. The results also showed that our two-step emotion classification was effective for fine-grained emotion classification. Specifically, fatal errors were significantly reduced with sentiment polarity classification before fine-grained emotion classification.

For future work, we first need to examine other machine learning methods to see their advantages and disadvantages in our task. We also need an extensive improvement in identifying neutral sentences. Finally, we are planning to apply our model to the active-listening dialog system that our group has been developing and investigate its effects on the user's behavior.

References

Allen, James F., Lenhart K. Schubert, George Ferguson, Peter Heeman, Chung Hee Hwang, Tsuneaki Kato, Marc Light, Nathaniel G. Martin, Bradford W. Miller, Massimo Poesio, and David R. Traum. 1994. The TRAINS Project: A case study in building a conversational planning agent. *Journal of Experimental and Theoretical AI (JETAI)*.

Ang, Jeremy, Rajdip Dhillon, Ashley Krupski, Elizabeth Shriberg, and Andreas Stolcke. 2002. Prosody-Based Automatic Detection Of Annoyance And Frustration In Human-Computer Dialog. *Spoken Language Processing*, pages 2037–2040.

Batliner, A., K. Fischer, R. Huber, J. Spilker, and E. Noth. 2004. How to find trouble in communication. *Speech Communication*, 40(1-2):117–143.

Chaumartin, Francois-Regis. 2007. A knowledge-based system for headline sentiment tagging. *In Proceedings of the 4th International Workshop on Semantic Evaluations*.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement 20*, pages 37–46.

Foster, Mary Ellen. 2007. Enhancing Human-Computer Interaction with Embodied Conversational Agents. *Lecture Notes in Computer Science*, 4555:828–837.

Hatzivassiloglou, Vasileios and Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*.

Kawahara, Daisuke and Sadao Kurohashi. 2006. Case Frame Compilation from the Web using High-Performance Computing. *In Proceedings of the 5th International Conference on Language Resources and Evaluation*.

Kobayashi, Nozomi, Ryu Iida, Kentaro Inui, and Yuji Matsumoto. 2005. Collecting Evaluative Expressions for Opinion Extraction. *Lecture Notes in Artificial Intelligence*, 3248.

Kozareva, Zornitsa, Borja Navarro, Sonia Vazquez, and Andres Nibtoyo. 2007. UA-ZBSA: A Headline Emotion Classification through Web Information. *In Proceedings of the 4th International Workshop on Semantic Evaluations*.

Kudo, Taku and Yuji Matsumoto. 2004. A Boosting Algorithm for Classification of Semi-Structured Text. *In Proceedings of the EMNLP*.

Litman, Diane J. and Kate Forbes-Riley. 2004. Predicting Student Emotions in Computer-Human Tutoring Dialogues. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*.

Liu, Bing. 2006. Web Data Mining. *Springer*, pages 411–440.

Masum, Shaikh Mostafa AI, Helmut Prendinger, and Mitsuru Ishizuka. 2007. Emotion Sensitive News Agent: An Approach Towards User Centric Emotion Sensing from the News. *In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 76–86.

Pantic, Maja and Leon J. M. Rothkrantz. 2004. Facial Action Recognition for Facial Expression Analysis From Static Face Images. *IEEE Transactions on SMC-B*, 34(3):1449–1461.

Robertson, Kathryn. 2005. Active listening: more than just paying attention. *Aust Fam Physician*, 34(12):1053–1055.

Rotaru, Mihai, Diane J. Litman, and Katherine Forbes-Riley. 2005. Interactions between Speech Recognition Problems and User Emotions. *Proceedings 9th European Conference on Speech Communication and Technology*.

Teramura, Hideo. 1982. Japanese Syntax and Meaning. *Kurosio Publishers (in Japanese)*.

Tokuhisa, Ryoko and Ryuta Terashima. 2006. Relationship between Utterance and "Enthusiasm" in Non-Task-Orieinted Conversational Dialogue. *In Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*.

Turney, P.D. 2002. Thumbs up? thumbs down? semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424.

Vapnik, Vladimir N. 1995. The Nature of Statistical Learning Theory. *Springer*.