Acquiring a Dictionary of Emotion-Provoking Events

Hoa Trong Vu[†,‡], Graham Neubig[†], Sakriani Sakti[†], Tomoki Toda[†], Satoshi Nakamura[†]

[†]Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma-shi, Nara, Japan

[‡]Vietnam National University, University of Engineering and Technology
E3 Building - 144 Xuan Thuy Street, Cau Giay, Hanoi, Vietnam

Abstract

This paper is concerned with the discovery and aggregation of events that provoke a particular emotion in the person who experiences them, or *emotion-provoking events*. We first describe the creation of a small manually-constructed dictionary of events through a survey of 30 subjects. Next, we describe first attempts at automatically acquiring and aggregating these events from web data, with a baseline from previous work and some simple extensions using seed expansion and clustering. Finally, we propose several evaluation measures for evaluating the automatically acquired events, and perform an evaluation of the effectiveness of automatic event extraction.

1 Introduction

"You look happy today, did something good happen?" This is a natural question in human dialogue, and most humans could think of a variety of answers, such as "I met my friends" or "I passed a test." In this work, we concern ourselves with creating resources that answer this very question, or more formally "given a particular emotion, what are the most prevalent events (or situations, contexts) that provoke it?"[1] Information about these *emotion-provoking events* is potentially useful for emotion recognition (recognizing emotion based on events mentioned in a dialogue), response generation (providing an answer to emotion-related questions), and answering social-science related questions (discovering events that affect the emotion of a particular segment of the population).

While there is very little previous research on this subject, one previous work of note by Tokuhisa et al. (2008) focused on emotion-provoking events purely from the viewpoint of emotion recognition. They used large corpus of examples collected from the Web using manual patterns to build a k-nearest-neighbors emotion classifier for dialog systems and found that the classifier significantly outperforms baseline methods. This method provides both an inspiration and a baseline for our work, but still lacks in that it makes no attempt to measure the quality of the extracted events, aggregate similar events, or rank events by prevalence, all essential factors when attempting to use extracted events for applications other than simple emotion recognition.

In this paper, we describe work on creating prevalence-ranked dictionaries of emotion-provoking events through both manual labor and automatic information extraction. To create a manual dictionary of events, we perform a survey asking 30 participants to describe events that caused them to feel a particular emotion, and manually cleaned and aggregated the results into a ranked list. Next, we propose several methods for extracting events automatically from large data from the Web, which will allow us to increase the coverage over the smaller manually created dictionary. We start with Tokuhisa et al. (2008)'s patterns as a baseline, and examine methods for improving precision and coverage through the use of seed expansion and clustering. Finally, we discuss evaluation measures for the proposed task, and perform an evaluation of the automatically extracted emotion-provoking events. The acquired events will be provided publicly upon acceptance of the paper.

2 Manual Creation of Events

In order to create a small but clean set of gold-standard data for each emotion, we first performed

[1] This is in contrast to existing sentiment lexicons (Riloff et al., 2003; Valitutti, 2004; Esuli and Sebastiani, 2006; Velikovich et al., 2010; Mohammad and Turney, 2013), which only record the sentiment orientation of particular words (such as "meet" or "friend"), which, while useful, are less directly connected to the emotions than the events themselves.

Emotions	Words
happiness	happy, glad
sadness	sad, upset
anger	angry, irritated
fear	afraid, scared
surprise	surprised, astonished
disgust	disgusted, terrible

Table 2: Seed words for each emotion.

a survey on emotion-provoking events. We did so by asking a total of 30 subjects (a mixture of male and female from 20-40 years of age) to write down five events that provoke each of five emotions: happiness, sadness, anger, fear, and surprise. As these events created according to this survey still have a large amount of lexical variation, we manually simplify them to their core and merge together events that have similar meanings.

Finally, for each emotion we extract all the events that are shared by more than one person. It should be noted that this will not come anywhere close to covering the entirety of human emotion, but as each event is shared by at least two people in a relatively small sample, any attempt to create a comprehensive dictionary of emotion-provoking events should at least be able to cover the pairs in this collection. We show the most common three events for each emotion in Table 1.

3 Automatic Extraction of Events

We also performed experiments attempting to automatically extract and aggregate events from Web data. As a starting point, we follow Tokuhisa et al. (2008) in defining a single reliable pattern as a starting point for event extraction:

I am EMOTION that EVENT

As this pattern is a relatively reliable indicator that the event is correct, most events extracted by this pattern will actually be emotion-provoking events. For instance, this pattern will be matched with the sentence "I am *happy* that *my mother is feeling better*", in which *my mother is feeling better* certainly causes happiness.

For the EMOTION placeholder, we take into account 6 emotions - happiness, sadness, anger, fear, disgust, and surprise - argued by Ekman (1992) to be the most basic. We manually create a short list of words that can be inserted into the above pattern appropriately, as shown in Table 2.

For the EVENT placeholder, we allow any string of words, but it is necessary to choose the scope of the string that is referring to the emotion-provoking event. To this end, we use a syntactic parser and set a hard restriction that all events must be a subtree having root tag S and containing at least one noun phrase and one verb phrase.

Given these two restrictions, these patterns provide us with high quality event-emotion pairs, but the method is still lacking in two respects, lack of coverage and lack of ability to aggregate similar events. As both of these are essential to creating a high-quality and non-redundant dictionary of events, we make two simple extensions to the extraction process as follows.

3.1 Pattern Expansion

Pattern expansion, or bootstrapping algorithms are widely used in the information extraction field (Ravichandran and Hovy, 2002). In particular Espresso (Pantel and Pennacchiotti, 2006) is known as a state-of-the-art pattern expansion algorithm widely used in acquiring relationships between entities. We omit the details of the algorithm for space concerns, but note that applying the algorithm to our proposed task is relatively straightforward, and allows us to acquire additional patterns that may be matched to improve the coverage over the single seed pattern. We do, however, make two changes to the algorithm. The first is that, as we are interested in extracting events instead of entities, we impose the previously mentioned restriction of one verb phrase and one noun phrase over all events extracted by the patterns. The second is that we perform normalization of events to reduce their variability, namely removing all function words, replacing proper nouns with special symbol, and lemmatizing words.

3.2 Grouping events

The second improvement we perform is grouping the extracted events together. Grouping has a number of potential practical advantages, as noted frequently in previous work (Becker et al., 2011). The first is that by grouping similar events together, we can relieve sparsity issues to some extent by sharing statistics among the events in a single group. The second is that aggregating events together allows humans to browse the lists more efficiently by reducing the number of redundant entries. In preliminary experiments, we attempted several clustering methods and even-

Emotions	Events		
happiness	meeting friends	going on a date	getting something I want
sadness	someone dies/gets sick	someone insults me	people leave me alone
anger	someone insults me	someone breaks a promise	someone is too lazy
fear	thinking about the future	taking a test	walking/driving at night
surprise	seeing a friend unexpectedly	someone comes to visit	receiving a gift

Table 1: The top three events for each emotion.

tually settled on hierarchical agglomerative clustering and the single-linkage criterion using cosine similarity as a distance measure (Gower and Ross, 1969). Choosing the stopping criterion for agglomerative clustering is somewhat subjective, in many cases application dependent, but for the evaluation in this work, we heuristically choose the number of groups so the average number of events in each group is four, and leave a further investigation of the tuning to future work.

4 Evaluation Measures

Work on information extraction typically uses accuracy and recall of the extracted information as an evaluation measure. However, in this work, we found that it is difficult to assign a clear-cut distinction between whether an event provokes a particular emotion or not. In addition, recall is difficult to measure, as there are essentially infinitely many events. Thus, in this section, we propose two new evaluation measures to measure the precision and recall of the events that we recovered in this task.

To evaluate the precision of the events extracted by our method, we focus on the fact that an event might provoke multiple emotions, but usually these emotions can be ranked in prominence or appropriateness. This is, in a way, similar to the case of information retrieval, where there may be many search results, but some are more appropriate than others. Based on this observation, we follow the information retrieval literature (Voorhees, 1999) in adapting mean reciprocal rank (MRR) as an evaluation measure of the accuracy of our extraction. In our case, one event can have multiple emotions, so for each event that the system outputs, we ask an annotator to assign emotions in descending order of prominence or appropriateness, and assess MRR with respect to these ranked emotions. [2]

We also measure recall with respect to the manually created dictionary described in Section 2, which gives us an idea of what percent of common emotions we were able to recover. It should be noted that in order to measure recall, it is necessary to take a matching between the events output by the system and the events in the previously described list. While it would be ideal to do this automatically, this is difficult due to small lexical variations between the system output and the list. Thus, for the current work we perform manual matching between the system hypotheses and the references, and hope to examine other ways of matching in future work.

5 Experiments

In this section, we describe an experimental evaluation of the accuracy of automatic extraction of emotion-provoking events.

5.1 Experimental Setup

We use Twitter[3] as a source of data, as it is it provides a massive amount of information, and also because users tend to write about what they are doing as well as their thoughts, feelings and emotions. We use a data set that contains more than 30M English tweets posted during the course of six weeks in June and July of 2012. To remove noise, we perform a variety of preprocessing, removing emoticons and tags, normalizing using the scripts provided by Han and Baldwin (2011), and Han et al. (2012). CoreNLP[4] was used to get the information about part-of-speech, syntactic parses, and lemmas.

We prepared four systems for comparison. As a baseline, we use a method that only uses the original seed pattern mentioned in Section 3 to acquire emotion-provoking events. We also evaluate expansions to this method with clustering, with pattern expansion, and with both.

We set a 10 iteration limit on the Espresso algorithm and after each iteration, we add the 20

[2] In the current work we did not allow annotators to assign "ties" between the emotions, but this could be accommodated in the MRR framework.

[3] `http://www.twitter.com`

[4] `http://nlp.stanford.edu/software/corenlp.shtml`

Methods	MRR	Recall
Seed	46.3 (±5.0)	4.6 (±0.5)
Seed + clust	57.2 (±7.9)	8.5 (±0.9)
Espresso	49.4 (±2.8)	8.0 (±0.5)
Espresso + clust	**71.7** (±2.9)	**15.4** (±0.8)

Table 3: MRR and recall of extracted data (with standard deviation for 3 annotators).

most reliable patterns to the pattern set, and increase the seed set by one third of its size. These values were set according to a manual inspection of the results for several settings, before any evaluation was performed.

We examine the utility of each method according to the evaluation measures proposed in Section 4 over five emotions, happiness, sadness, anger, fear, and surprise.[5] To measure MRR and recall, we used the 20 most frequent events or groups extracted by each method for these five emotions, and thus all measures can be interpreted as MRR@20 and recall@20. As manual annotation is required to calculate both measures, we acquired results for 3 annotators and report the average and standard deviation.

5.2 Experimental Results

The results are found in Table 3. From these results we can see that clustering the events causes a significant gain on both MRR and recall, regardless of whether we use Espresso or not. Looking at the results for Espresso, we see that it allows for small boost in recall when used on its own, due to the fact that the additional patterns help recover more instances of each event, making the estimate of frequency counts more robust. However, Espresso is more effective when used in combination with clustering, showing that both methods are capturing different varieties of information, both of which are useful for the task.

In the end, the combination of pattern expansion and clustering achieves an MRR of 71.7% and recall of 15.4%. While the MRR could be deemed satisfactory, the recall is still relatively low. One reason for this is that due to the labor-intensive manual evaluation, it is not realistic to check many more than the top 20 extracted events for each emotion, making automatic evaluation metrics the top on the agenda for future work.

Emotions	MRR	Recall
happiness	93.9	23.1
sadness	76.9	10.0
anger	76.5	14.0
fear	48.3	24.3
surprise	59.6	0.0

Table 4: Average MRR and recall by emotion for the Espresso + clustering method.

However, even without considering this, we found that the events extracted from Twitter were somewhat biased towards common, everyday events, or events regarding love and dating. On the other hand, our annotators produced a wide variety of events including both everyday events, and events that do not happen every day, but leave a particularly strong impression when encountered. This can be seen particularly in the accuracy and recall results by emotion for the best system shown in Table 4. We can see that for some emotions we achieved recall approaching 25%, but for surprise we didn't manage to extract any of the emotions created by the annotators at all, instead extracting more mundane events such as "surprised I'm not fat yet" or "surprised my mom hasn't called me yet." Covering the rare, but important events is an interesting challenge for expansions to this work.

6 Conclusion and Future Work

In this paper we described our work in creating a dictionary of emotion-provoking events, and demonstrated results for four varieties of automatic information extraction to expand this dictionary. As this is the first attempt at acquiring dictionaries of emotion-provoking events, there are still many future directions that deserve further investigation. As mentioned in the experimental discussion, automatic matching for the evaluation of event extraction, and ways to improve recall over rarer but more impressive events are necessary. There are also many improvements that could be made to the extraction algorithm itself, including more sophisticated clustering and pattern expansion algorithms. Finally, it would be quite interesting to use the proposed method as a tool for psychological inquiry, including into the differences between events that are extracted from Twitter and other media, or the differences between different demographics.

[5] We exclude disgust, as the seed only matched 26 times over entire corpus, not enough for a reasonable evaluation.

References

Hila Becker, Mor Naaman, and Luis Gravano. 2011. Beyond trending topics: Real-world event identification on Twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM11)*.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200.

Andrea Esuli and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation*, pages 417–422.

John C Gower and GJS Ross. 1969. Minimum spanning trees and single linkage cluster analysis. *Applied statistics*, pages 54–64.

Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages: makn sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 368–378.

Bo Han, Paul Cook, and Timothy Baldwin. 2012. Automatically constructing a normalisation dictionary for microblogs. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 421–432, Jeju Island, Korea, July. Association for Computational Linguistics.

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Patrick Pantel and Marco Pennacchiotti. 2006. Espresso: leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 113–120.

Deepak Ravichandran and Eduard Hovy. 2002. Learning surface text patterns for a question answering system. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 41–47.

Ellen Riloff, Janyce Wiebe, and Theresa Wilson. 2003. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 25–32. Association for Computational Linguistics.

Ryoko Tokuhisa, Kentaro Inui, and Yuji Matsumoto. 2008. Emotion classification using massive examples extracted from the web. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 881–888.

Ro Valitutti. 2004. Wordnet-affect: an affective extension of wordnet. In *In Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1083–1086.

Leonid Velikovich, Sasha Blair-Goldensohn, Kerry Hannan, and Ryan McDonald. 2010. The viability of web-derived polarity lexicons. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 777–785.

Ellen M Voorhees. 1999. The trec-8 question answering track report. In *Proceedings of TREC*, volume 99, pages 77–82.