Detecting Emotion Stimuli in Emotion-Bearing Sentences

Diman Ghazi, Diana Inkpen, and Stan Szpakowicz

School of Electrical Engineering and Computer Science
University of Ottawa, Ottawa, Ontario, Canada
diman.ghazi@gmail.com, Diana.Inkpen@uOttawa.ca,
szpak@eecs.uottawa.ca

Abstract. Emotion, a pervasive aspect of human experience, has long been of interest to social and behavioural sciences. It is now the subject of multi-disciplinary research also in computational linguistics. Emotion recognition, studied in the area of sentiment analysis, has focused on detecting the expressed emotion. A related challenging question, *why* the experiencer feels that emotion, has, to date, received very little attention. The task is difficult and there are no annotated English resources. FrameNet refers to the person, event or state of affairs which evokes the emotional response in the experiencer as emotion *stimulus*.[1] We automatically build a dataset annotated with both the emotion and the stimulus using FrameNet's *emotions-directed* frame. We address the problem as information extraction: we build a CRF learner, a sequential learning model to detect the emotion stimulus spans in emotion-bearing sentences. We show that our model significantly outperforms all the baselines.

1 Introduction

Causality is a semantic relation defined as "the relationship between cause and effect",[2] where the latter is understood as a consequence of the former. Causality detection used to depend on hand-coded domain-specific knowledge bases [17]. [16] defined semantic constraints to rank possible causality, and machine learning techniques now prevail.

[15] described automatic detection of lexico-syntactic patterns which express causation. There are two steps: discover lexico-syntactic patterns, and apply inductive learning to automatically detect syntactic and semantic constraints (rules) on the constituent components. [5] extracted possible causal relations between noun phrases, via a bootstrapping method of causality extraction using cue phrases and word-pair probabilities. A simple supervised method in [3] trained SVM models using features derived from WordNet and the Google N-gram corpus; providing temporal information to the causal relations classifier boosted the results significantly. All that work, however, addresses causality in general, by nature very different from detecting the cause of an emotion. The causal relation between two parts of the text are sought. We look for the emotion stimulus when the emotion can be conveyed explicitly in the text or be implicit.

[1] Most authors talk generically of *cause*.
[2] http://www.oxforddictionaries.com

A. Gelbukh (Ed.): CICLing 2015, Part II, LNCS 9042, pp. 152–165, 2015.
DOI: 10.1007/978-3-319-18117-2_12

Table 1. The elements of the emotion frame in FrameNet

Core	Event	The occasion or happening in which Experiencers in a certain emotional state participate.
	Experiencer	The person or sentient entity who experiences or feels the emotions.
	Expressor	It marks expressions that indicate a body part, gesture or other expression of the Experiencer that reflects his or her emotional state.
	State	The abstract noun that describes a more lasting experience by the Experiencer.
	Stimulus	The person, event, or state of affairs that evokes the emotional response in the Experiencer.
	Topic	The general area in which the emotion occurs. It indicates a range of possible Stimulus.
Non-Core	Circumstance	The condition(s) under which the Stimulus evokes its response.
	Degree	The extent to which the Experiencer's emotion deviates from the norm for the emotion.
	Empathy-target	The individual or individuals with which the Experiencer identifies emotionally and thus shares their emotional response.
	Reason/ Explanation	The explanation for why the Stimulus evokes a certain emotional response.
	Manner	It is any description of the way in which the Experiencer experiences the Stimulus which is not covered by more specific frame elements. Manner may also describe a state of the Experiencer that affects the details of the emotional experience.
	Parameter	A domain in which the Experiencer experiences the Stimulus.

A more specific task in causality analysis, most similar to our task, is to identify sources of opinions. [4] used machine-learning techniques to identify propositional opinions and their holders (sources). That pioneering work was limited in scope: only propositional opinions (which function as the sentential complement of a predicate), and only direct sources. [9], also among the pioneers in this field, viewed the problem as an information extraction task and tackled it using sequence tagging and pattern matching techniques simultaneously. They hypothesized that information extraction techniques would be well-suited to source identification because an opinion statement can be viewed as a kind of speech event with the source as the agent. [8] identify two types of opinion-related entities, expressions of opinions and sources of opinions, along with the linking relation between them. [19] analyzed judgment opinions, which they define as consisting of valence, a holder and a topic. All that work invokes interchangeably the terms *source of an opinion* and *opinion holder*. Although the source can be the reason that implies the opinion, it is mainly seen as the opinion holder – thus, it is a task very different from ours.

We focus on detecting the *emotion stimulus*. In FrameNet [13], an *experiencer* has a particular emotional *state*, which may be described in terms of a specific *stimulus* that invokes it, or a *topic* which categorizes the kind of stimulus. An explicitly named experiencer may be replaced by an *event* (with participants who are experiencers of the emotion) or an *expressor* (a body-part of gesture which would indicate the experiencer's state to an external observer). There can also be a *circumstance* in which the response occurs or a *reason* why the stimulus evokes the particular response in the experiencer.

Consider, for example, the sentence "In the Commons, Labour MPs unleashed their anger at the Liberal Democrats for promising to back the Government." "Labour MPs" is the *Experiencer*, "anger" is the expression of emotion, "the Liberal Democrats" is the *Emotion Stimulus*, and "for promising to back the Government" is the *Explanation*. We want our system to find the reason why Labour MPs were angry: to return the span *the Liberal Democrats* as the emotion stimulus.

[13] define six core frame elements for an emotion and six non-core elements (see Table 1). Of these, emotion stimulus seems to be the closest to saying *why* the experiencer feels that emotion. Therefore, here we focus on this aspect of emotion analysis.

We are particularly interested in the stimulus because determining *why* an emotion occurs has such intriguing applications as consumer behaviour analysis or mental-health care. It would be very useful for systems which answer question such as "how [x] feels about [y]" or "why [x] feels [y]". It also has practical importance to text summarization, because emotion expressions and emotion stimuli tend to be the most informative in an expressive sentence, so they can get higher weight in abstractive summarization.

We discuss emotion stimuli, a dataset we automatically built with emotion stimulus and emotion expression labels, and ways of detecting emotion stimuli. Section 2 covers the related work. Section 3 explains the process of collecting and annotating an emotion stimulus dataset using FrameNet data.[3] Section 4 discusses the features and the baselines for detecting emotion stimuli in emotion-bearing sentences as well as the experiments and the results. Section 5 concludes the paper and suggests future work.

2 Related Work

Researchers in the field of affective computing have investigated recognition, interpretation and representation of affect [30]. They consider a wide range of modalities such as affect in speech, facial display, posture and physiological activity. Due to the large volume of text data available on the Internet – blogs, email and chats – which are full of emotions, recently there has been a growing interest in automatic identification and extraction of sentiment, opinions and emotions in text. Besides, textual data on the Web take up little physical space and are easily transferred, so they have a high potential to be used for sharing ideas, opinion and emotions. It is also such an active area of research because its applications have spread to multiple domains, from consumer product reviews, health care and financial services to social events and political elections [14].

In order to recognize and analyze affect in written text – seldom explicitly marked for emotions – NLP researchers have come up with a variety of techniques, including the use of machine learning, rule-based methods and the lexical approach [28] [1] [2] [18] [37] [33] [6] [20]. Detecting emotion stimuli, however, is a very new concept in sentiment analysis. Emotion/stimulus interaction, an eventive relation, potentially yields crucial information in terms of information extraction. For example, we can predict future events or decide on the best reaction if we know the emotion cause [7].

[3] http://www.eecs.uottawa.ca/ diana/resources/emotion_stimulus_data/

Event-based emotion detection has been addressed in some previous research [38] [35] [24] but, to the best of our knowledge, only [7], [22] and [23] have worked on emotion cause detection. They explore emotions and their causes, focusing on five primary emotions – happiness, sadness, fear, anger and surprise – in Chinese texts. They have constructed a Chinese emotion corpus, but they focus on explicit emotions presented by emotion keywords, so each emotion keyword is annotated with its corresponding causes, if existing. In their dataset, they observe that most causes appear within the same clause of the representation of the emotion, so a clause might be the most appropriate unit to detect a cause. We find such granularity too large to be considered an emotion stimulus in English. Also, clauses were distinguished by punctuation: comma, period, question mark and exclamation mark. Just four punctuation marks are not enough to capture English clauses adequately.

Using linguistic cues, including causative verbs and perception verbs, the authors create patterns to extract general cause expressions or specific constructions for emotion causes. They formalize emotion cause detection as multi-label classification. Each instance may contain more than one label, such as "left-1, left-0", to represent the location of the clauses which are part of the cause. We have no evidence that their findings can be valid for English data. Their work is more of a solution for an over-simplified version of a complicated problem. Also, keyword spotting is used to detect the emotion word in the sentence and try to find the cause of that emotion, but not all emotions are expressed explicitly by emotion keywords.

In the end, what stands out is the fact that, as far as we know, there has been no significant previous work on emotion cause detection in *English* data. This may be due to the relative complexity of English in expressing emotions in text, or to the limitation in existing resources and datasets either for supervised machine learning or for evaluation purposes.[4] Therefore, we have collected data and built a new dataset for detecting emotion stimuli. Using the dataset, we also explore different baselines to establish the difficulty of the task. Finally, we train a supervised information extraction model which detects the emotion stimulus spans in emotion-bearing sentences.

3 Data Collection and Annotation

[13] define 173 emotion-directed lexical units which correspond to different emotions. A lexical unit is a word/meaning pair (essentially a lexeme).[5] Typically, each sense of a polysemous word belongs to a different semantic frame. "Happy", "angry" and "furious" are examples of LUs in the Emotion-directed frame. For each of them, FrameNet annotators have labelled some sentences. We built a set of sentences marked with the emotion stimulus (cause), as well as the emotion itself. To collect a larger set of data, we used synonyms of emotion LUs to group the data into fewer basic emotions. In the manual synonym annotation task, we suggested Ekman's six emotions (*happiness*, *sadness*, *suprise*, *disgust*, *anger*, *fear*) [11], as well as *shame*, *guilt* and *hope*, posited in literature [29], to consider if the emotion LUs did not fit an Ekman emotion. We also

[4] [27] recently built a small dataset with 523 sentences for 22 emotions. We found it too scarce for machine learning methods of cause detection.

[5] The term will be henceforth abbreviated as LU.

allowed the annotators to propose a more appropriate emotion not on the list. In the end, we chose Ekman's emotions and *shame*.[6]

Some emotions fit a basic one; for example, *fury* clearly belongs to the *anger* category. On the other hand, *affront* is not so obvious. We will now discuss the dictionaries and thesauri we used to get emotions and their synonyms. Then we will describe the manual annotation. Next, we will apply the first round of tagging the emotions with one of the seven classes we chose. We separate the emotion LUs with a strong agreement between sources and annotations from those with weaker agreement. For the latter set of LUs, we relax one of the conditions. We use two more emotion lexicons to break the ties and classify more LUs. Finally we build two datasets, one with the group of emotions with strong agreement and the other with the result of the second phase of annotation added.

3.1 Annotating Emotion Lexical Units

We collected the synonyms from two trustworthy online sources.[7] Of the LUs collected from FrameNet, 14 are not covered at all. The two sources also do not always agree. To get a tie breaker, we resorted to manual annotation. We grouped the list of 173 emotion LUs from FrameNet into Ekman's six emotion classes by using the synonym list from the Oxford Dictionary and from thesaurus.com, joined the results and asked human annotators (fluent English speakers) to verify those annotations. We gave them an Excel table with each LU displayed in a row and each emotion in a column. For each word, the corresponding emotions were marked. A word could be classified into:

I one of Ekman's six emotions (110 words);
II two classes at the same time (14 words);
III none of the six emotion classes (49 words).

For group I, the annotators indicated if they disagreed with the existing classification by crossing the existing mark and indicating the emotion they think is more appropriate (if there is one). For group II, they chose only one of the two emotions they thought was closer to the LU and crossed out the other one. For group III, they chose one of the three suggested classes, *guilt*, *shame* and *hope*, and grouped the LU into one of them. Finally, there was a column for comments, where the annotators could write any other emotion they thought was more suitable as a synonym of the emotion LU.

The results of these annotations are used to build the dataset containing sentences with emotion stimulus tags. Section 3.2 presents the process of building the dataset.

3.2 Building the Dataset

Considering both thesauri and human annotators' tags, each LU was tagged with the emotion that had the highest number of votes. We combined the result of the Oxford

[6] *Guilt* and *hope* were not found useful, but *confusion* was suggested as another emotion with many synonym matches among the emotion LUs. This requires further study. *Confusion* is not considered in related research, and anyhow we need more evidence in the literature.

[7] `http://www.thesaurus.com`
`http://www.oxforddictionaries.com/thesaurus/`

Dictionary and thesaurus.com, so we had four annotations in total. As a result, 102 out of 173 LUs were labeled with high agreement (at least 3 out of 4 votes), which are shown as strong agreement in Table 2.

For the other LUs, we did not have enough information to group them into one of the seven emotions. We used two more sources – the NRC emotion lexicon [26] and the WordNet affect lexicon [34] – to break some of the ties. In this step, the LUs were labeled with an emotion for which the number of votes was more than half of the votes (three in our case) and the difference of the votes for the two top emotion classes was at least 2. This added 21 LUs to our set. While the NRC emotion lexicon was not very useful,[8] many ties were broken by WordNet Affect. The result of this extended list is shown as weaker agreement in Table 2.

Table 2. Distribution of LUs for each emotion class

Agreement	happiness	sadness	surprise	disgust
Strong	22	31	12	3
Weaker	22	32	12	8
	anger	fear	shame	total
Strong	19	13	2	102
Weaker	26	17	6	123

Using the list of grouped emotion synonyms, we collected FrameNet data manually labeled for each LU. Next, we selected the sentences which contain emotion stimuli and we assigned each sentence to its corresponding emotion class. The distribution of the instances in the dataset is shown in Table 3. Each instance is a complete sentence, 18 tokens on average, and contains one stimulus assigned to the emotion LU.

Table 3. Distribution of labels in the emotion stimulus datasets. Dataset 1 contains synonym groups with strong agreement, Dataset 2 – also those with weaker agreement.

	happiness	sadness	surprise	disgust
Dataset 1	211	98	53	6
Dataset 2	211	107	53	38
	anger	fear	shame	total
Dataset 1	168	129	34	699
Dataset 2	199	144	68	820

As a complementary dataset, we also collected the sentences with no stimulus tag, yet containing expressions of one of the seven emotions. This dataset is much larger than the dataset with stimulus. This makes us wonder whether it is due to the existence of many emotion causes implicit in the context; or if it is because of other possible frame elements in the sentence we disregard, such as circumstances and explanation, which can indicate emotion stimuli; or if a sentence is not enough to always contain

[8] It puts many words into multiple negative classes such as *sadness*, *anger*, *disgust*. For example, *Despair* is labeled as anger, disgust, fear, and sadness.

the emotion stimulus and we should consider larger text portions (maybe the current sentence and the previous and next sentence). Nonetheless, we believe that building this dataset is useful in choosing one of these three reasons. As a future work, we also would like to extend the emotion-stimulus dataset by considering other frame elements such as circumstances and explanation, which can indicate emotion stimuli. The distribution of the emotion instances in this dataset is presented in Table 4.

Table 4. Distribution of labels in the emotion datasets with no stimulus

happiness	sadness	surprise	disgust
268	468	160	57
anger	fear	shame	total
284	279	77	1594

The next section explains how we use the emotion stimulus dataset to build a supervised model to learn emotion stimulus spans in emotion-bearing sentences.[9]

4 Automatically Detecting Emotion Stimulus

To assess the difficulty of detecting emotion stimuli, we develop baseline systems which work with intuitive features . We also explore various features and their effect on stimulus detection results. We have built labeled data annotated with emotion stimuli, so we also have the privilege to explore supervised learning methods for information extraction. Of the datasets explained in Table 3, we use the second one, with 820 instances. That gives us more data to learn from.

Statistical learning models help avoid the biases and insufficiency of coverage of manual rule and pattern detection methods. One of the most common learning paradigms for performing such labelling tasks are Hidden Markov Models or probabilistic finite-state automata to identify the most likely sequence of labels for the words in any given sentence [36]. Such models, however, do not support tractable inference, and they represent the data by assuming their independence. One way of satisfying both these criteria is to use a model which defines a conditional probability over label sequences given a particular observation sequence rather than using a joint distribution over both label and observation sequences.

Conditional random fields (CRFs) [21] are a probabilistic framework for labelling and segmenting sequential data, based on a conditional model which labels a novel observation sequence x by selecting the label sequence y maximizing the conditional probability p(y|x). We use CRF from MinorThird [10], because it allows error analysis: comparing the predicted labels with the actual labels by highlighting them on the actual text. It also ranks the features based on the weight, so we can see which features have contributed the most.

[9] The dataset we built indicates both the emotion expressed and the emotion stimulus in each sentence. In this work, however, we only detect emotion stimulus, and assume that the emotion expression is present (in our data, it is). Our future work will address both emotion expression and emotion stimulus detection at the same time.

4.1 Baselines

The baselines we explain here set the ground for comparing our results and evaluating the performance of different models. One of the main properties of emotions is that they are generally elicited by stimulus events: something happens to the organism to stimulate or trigger a response after having been evaluated for its significance [32]. That is why events seem to be the most obvious indicators of emotions; we build our first two baselines upon events. We *are* aware that they are not the only emotion stimuli, but we believe them to be important enough.

One problem with using events as emotion stimuli is that event detection itself is a challenging task. The literature suggests verbal and nominal events; the former are much more numerous [7]. A verb conveys an action, an occurrence or a state of being. We use verbs as a textual signal of events; as our first baseline, we mark verbs in a sentence as the emotion stimuli. We retrieve verbs with the OpenNLP POS tagger.[10] The second baseline is Evita [31], a tool which detects both nominal and verbal events; as an emotion stimulus, we select an event in a sentence at random.

We noted earlier that not only events can be stimuli. FrameNet defines a stimulus as the person, event or state of affairs that evokes the emotional response in the Experiencer. For the third baseline, we recognize an emotion stimulus in a larger portion of the sentence, a phrase or a syntactically motivated group of words. We use the OpenNLP chunker,[11] and randomly select as an emotion stimulus a chunk which contains a verb.

[23] used as a baseline the clause with the first verb to the left of the emotion keyword. In English, however, there are single-clause sentences such as "My grandfather's death made me very sad." or "I was surprised to hear the ISIS news." with both the emotion state and the stimulus. The whole sentence would be returned as an emotion stimulus. Even so, we believe that it is worth exploring and investigating how useful a clause will be in detecting emotion stimuli in English. As the next baseline, we select a random clause as the stimulus. In OpenNLP parse trees,[12] we take the S, and SBAR tags as indicators of independent and dependent clauses in a sentence. Next, we randomly choose one of the clauses as the emotion stimulus.

Finally, we use Bag-of-Words as a typical baseline for all NLP tasks and for the sake of comparison. The previous baselines were rule-based systems with simple heuristics. For this baseline, we apply CRF sequence learner from MinorThird to all the unigrams in the text. In the *sequence annotator learner* we select *CRF Learner* as the classifier, 100 (a MinorThird default) as the number of iterations over the training set, and *5-fold cross validation* as the evaluation option.[13]

[10] `http://opennlp.apache.org/documentation/manual/opennlp.html#tools.postagger`

[11] `http://opennlp.apache.org/documentation/1.5.2-incubating/manual/opennlp.html#tools.chunker`

[12] `http://opennlp.apache.org/documentation/manual/opennlp.html#tools.parser`

[13] The dataset is too small for 10-fold cross validation. We only wanted to use unigrams as features in the baseline, but the CRF learner in MinorThird adds the previous labels of each token as a feature. The results, then, are higher than when using only unigram features.

Table 5. Baselines, the results. For random chunks, a quick experiment shows that verb phrase and noun phrase chunks are only 30% of the text. For Bag-of-Words, the span-level evaluation scores are 0.3293, 0.2132 and 0.2588.

	Precision	Recall	F-measure
Verb	0.212	0.059	0.093
Evita events	0.265	0.044	0.076
Random Chunk	0.292	0.0692	0.112
Random Clause	0.419	0.532	0.469
Bag-of-Words	0.5904	0.5267	0.5568

The baseline results are presented in Table 5. In span detection problems, the evaluation measures can either be based on the number of matching tokens or be more strict and consider the exact spans and the number of exact matches. Consider the sentence "His doctors were astounded that he survived the surgery." The emotion stimulus span ought to be "that he survived the surgery." If we return "that he survived" instead, token-based measures find three tokens matches, but span-based measures treat this as no match. Naturally, the value of token-level measures is higher than span-level measures. That is why, to build the higher-bound baseline, we report the token level precision, recall and F-measure.

The results indicate very low coverage in the first three baselines while the clause and Bag-of-Words baselines are much higher. The reason can be that the data in FrameNet are well-formed and carefully collected. Having a quick look at the instances shows that the emotion stimulus tends to be longer than just a verb or a phrase. Stimuli are long enough to say why a particular emotion was experienced. Therefore as a baseline the random clause and Bag-of-Words experiments have higher coverage. We believe that, although the first baselines' results are really low when used as the only feature in a simple rule-based system, they still are interesting features to study as features in our machine learning methods. In the next section we will discuss adding these features, and compare the results with the baseline.

4.2 Features and Results

Corpus-Based. We use a set of corpus-based features built in MinorThird's text analysis package. Among the features there are the lower-case version of each single word, and analogous features for tokens in a small window to either side of the word. Here we set the window size to three as suggested by the literature [9]. Additional token-level features also include information whether the token is a special character such as a comma, and orthographic information. For example, there is a feature, the character pattern "X+", which indicates tokens with all capital letters. The features are grouped into positive and negative. *Positive* refers to the group of features built and weighted based on the tokens within the stimulus span. *Negative* are the features related to all the tokens outside of the targeted span. This feature extraction process results in 23,896 features used in our learning process.

We applied *CRF Learner* with the same settings as the Bag-of-Words baseline in the previous section. The result of these experiments are shown in Table 6.

Table 6. Results of detecting emotion stimulus using different features

	Token Precision	Token Recall	Token F-measure	Span Precision	Span Recall	Span F-measure
Corpus-Based	0.7460	0.7017	0.7232	0.5658	0.5402	0.5527
Corpus-Based + Event	0.766	0.756	0.761	0.567	0.561	0.5644
Corpus-Based + Chunker	0.776	0.761	0.7688	0.564	0.556	0.5603
Corpus-Based + Clause	0.809	0.731	0.768	0.623	0.564	0.592
Corpus-Based + Event + Chunker + Clause	0.811	0.746	0.777	0.666	0.593	0.6280

An analysis of our learnt model and the feature weights shows that, for the positive tokens, the left-side token features have a higher weight than the right-side tokens. It is the opposite for the negative tokens. Also, the highest-weighted token features include "at", "with", "about", "that" and emotion words such as "delight", "concerned", "ashamed", "anger" for the left-side tokens.

Although the result of these experiment significantly outperform all the baselines, we notice that the span precision and recall are much lower than at the token level. The reason is that the syntactic structure of a sentence is not considered in this set of features. According to the ranked features, many function words are among the highest-weighted features. This means that this task is very structure-dependent.

A few examples showcase some of the shortcomings of this model by comparing what is learnt (blue) versus what is the actual stimulus (green).

- "Colette works at marshalling our feelings of revulsion {{at this} voracious creature who has almost killed the poor box thorn.}" This example shows that, although these features might be useful to detect the beginning of the emotion stimulus, detecting the end of the span seems more challenging for them.
- "He was petrified {of the clippers} {at first}." In this case the model has learned that many emotion stimuli start with the word "at", so it chooses "at first" regardless of its semantic and syntactic role in the sentence.
- "At a news conference {at the Royal Geographical Society in London} , they described the mental and physical anguish {of their 95-day trek}." Lacking semantic features, the model does not recognize that a location cannot be an emotion stimulus alone.

Looking at the predicted labels and comparing them with the actual labels shows that we need deeper semantic and syntactic features (explained in the next sections).

Events. FrameNet's definition of emotion stimulus treats events as one of the main factors in detecting stimuli. That is why we use a tool to automatically detect events and add them to the features. The following examples show how events can be the main part of emotion stimuli.

- "I am desolate that Anthony has died."
- "His last illness was the most violent , and his doctors were astounded that he survived it ."
- "I join the Gentleman in expressing our sorrow at that tragic loss."

Evita [31] is a tool which develops algorithms to tag mentions of events in text, tag time expressions, and temporally anchor and order the events. The EVENT tag is used to annotate those elements in a text which mark the semantic events described by it. Syntactically, events are typically verb phrases, although some nominals, such as "crash" in "killed by the crash", will also be annotated as events. Evita's event classes are *aspectual*, *I-action*, *I-state*, *occurrence*, *perception*, *reporting* and *state*. The result of adding the event tags to the previous features in presented in Table 6.

Chunker. Text chunking divides a text into syntactically coherent segments like noun groups or verb groups, but does not specify their internal structure, nor their role in the main sentence. We use the OpenNLP chunker to tag the data with the chunks, because we believe that the chance of an emotion stimulus starting or ending in the middle of a chunk is very low. A chunker should help improve the span precision and recall.

Here are examples with the actual and predicted emotion stimulus label, using the previous model. We believe that considering chunks should help reduce the kind of errors which these examples illustrate.

- "Their cheerfulness and delight {{at still being} alive} only made Charlie feel more guilty." "Being alive" should be placed in one chunk, therefore the span will not end in the middle of a chunk.
- "Feeling a little frightened {{of the dead body behind} him in the cart} , he stopped for some beer at a pub , where he met Jan Coggan and Laban Tall." Again, "behind him in the cart" should be in one chunk, so by using a chunker we would know the predicted span was incorrect.

Clause. In English grammar, a clause is the smallest grammatical unit which can express a complete proposition. There are two different types of clauses, independent and dependent. An independent clause can stand alone as a complete sentence. Dependent clauses can be nominal, adverbial or adjectival. Noun clauses answer questions like "who(m)?" or "what?" and adverb clauses answer questions like "when?", "where?", "why?".[14] Although there might not be many cases when the whole clause is the emotion stimulus, there are some cases, as mentioned below, which make it worthwhile to look into clauses and considering them among the features.

To mark the clauses in a sentence, we use the OpenNLP parser. As suggested in the literature [12], we use the *SBAR* tag, which represents subordinate clauses in the parse trees, to identify dependent clauses in a sentence. We use the *S* tag inside the sentence to indicate independent clauses. The output of the parser is shown in the following example which shows how the emotion stimulus tag exactly aligns with the SBAR tag which indicates the subordinate clause.

- "I am pleased that they have responded very positively."[15]
- "I was so pleased she lived until just after Sam was born."[16]

[14] `http://www.learnenglish.de/grammar/clausetext.html`

[15] The parse is "I am pleased [SBAR that [S they have responded very positively.]]"

[16] The parse is "I was so pleased [SBAR [S she lived until [SBAR just after [S Sam was born.]]]]"

The result of adding the clause tags to the previously discussed features in presented in Table 6. At the end, we show the result of combining all the discussed features.

These results show that each set of features improves our span-learning model, while clause-based features are most effective among events, chunks and clause feature sets. Also, the combination of all features significantly outperforms every baseline. More improvement could come from adding more fine-grained features to each feature group. For example, we can add the type, tense and aspect of an event – provided by the Evita tool. We can also improve our chunk-based features by postprocessing the chunker's result: combining relevant chunks into longer chunks. For example, two noun phrases with a preposition between them can give a longer noun phrase; this could be more useful in our task. Finally, although the CRF results are promising, we ought to explore other sequential learning methods such as maximum-entropy Markov models (MEMM), or conditional Markov models (CMM).

5 Conclusion and Future Directions

We have framed the detection of emotion causes as finding a stimulus element as defined for the emotion frame in FrameNet. We have created the first ever dataset annotated with both emotion stimulus and emotion statement;[17] it can be used for evaluation or training purposes. We used FrameNet's annotated data for 173 emotion LUs, grouped the LUs into seven basic emotions using their synonyms and built a dataset annotated with both the emotion stimulus and the emotion. We applied sequential learning methods to the dataset. We also explored syntactic and semantic features in addition to corpus-based features. We built a model which outperforms all our carefully built baselines.

The set we built in this work is small and well-formed, and contains carefully built data annotated by humans. To show the robustness of our model and to study the problem thoroughly,we would like in the future to extend our dataset in two ways: first to study *Circumstances* and *Explanation* frame elements to investigate whether they can also indicate emotion stimuli to be added to our dataset. Secondly, we would like to use semi-supervised bootstrapping methods to add instances of other existing emotion datasets which do not have emotion cause labels.

Also as a preliminary step to emotion stimulus detection, we would like first to define whether the sentence contains an emotion stimulus and then detect the emotion stimulus span. In this work, we built a dataset with emotion statements with no stimulus tag which could be used for this purpose.

Last but not least, we believe that an emotion stimulus and the emotion itself are not mutually independent. Although in this work we did not take the emotion of the sentences into account, in the future we would like to detect both the emotion and the emotion stimulus at the same time and to investigate whether indicating emotion causes can improve emotion detection and vice versa.

[17] [25] see stimulus narrowly as one towards whom the emotion is directed.

References

1. Alm, C.O., Roth, D., Sproat, R.: Emotions from Text: Machine Learning for Text-based Emotion Prediction. In: HLT/EMNLP, pp. 347–354 (2005)
2. Aman, S., Szpakowicz, S.: Identifying expressions of emotion in text. In: Matoušek, V., Mautner, P. (eds.) TSD 2007. LNCS (LNAI), vol. 4629, pp. 196–205. Springer, Heidelberg (2007)
3. Bethard, S., Martin, J.H.: Learning Semantic Links from a Corpus of Parallel Temporal and Causal Relations. In: Proc. ACL 2008 HLT Short Papers, pp. 177–180 (2008)
4. Bethard, S., Yu, H., Thornton, A., Hatzivassiloglou, V., Jurafsky, D.: Automatic Extraction of Opinion Propositions and their Holders. In: 2004 AAAI Spring Symposium on Exploring Attitude and Effect in Text, pp. 22–24 (2004)
5. Chang, D.S., Choi, K.S.: Incremental cue phrase learning and bootstrapping method for causality extraction using cue phrase and word pair probabilities. Information Processing and Management 42(3), 662–678 (2006)
6. Chaumartin, F.R.: UPAR7: A knowledge-based system for headline sentiment tagging. In: Proc. 4th International Workshop on Semantic Evaluations, SemEval 2007, pp. 422–425 (2007)
7. Chen, Y., Lee, S.Y.M., Li, S., Huang, C.R.: Emotion cause detection with linguistic constructions. In: Proc. 23rd International Conference on Computational Linguistics, COLING 2010, pp. 179–187 (2010)
8. Choi, Y., Breck, E., Cardie, C.: Joint extraction of entities and relations for opinion recognition. In: Proc. 2006 Conference on Empirical Methods in Natural Language Processing, EMNLP 2006, pp. 431–439 (2006)
9. Choi, Y., Cardie, C., Riloff, E., Patwardhan, S.: Identifying sources of opinions with conditional random fields and extraction patterns. In: Proc. Human Language Technology and Empirical Methods in Natural Language Processing, HLT 2005, pp. 355–362 (2005)
10. Cohen, W.W.: Minorthird: Methods for Identifying Names and Ontological Relations in Text using Heuristics for Inducing Regularities from Data (2004), `http://minorthird.sourceforge.net`
11. Ekman, P.: An argument for basic emotions. Cognition & Emotion 6(3), 169–200 (1992)
12. Feng, S., Banerjee, R., Choi, Y.: Characterizing Stylistic Elements in Syntactic Structure. In: Proc. the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012, pp. 1522–1533 (2012)
13. Fillmore, C.J., Petruck, M.R., Ruppenhofer, J., Wright, A.: FrameNet in Action: The Case of Attaching. IJL 16(3), 297–332 (2003)
14. Ghazi, D., Inkpen, D., Szpakowicz, S.: Prior versus contextual emotion of a word in a sentence. In: Proc. 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis, WASSA 2012, pp. 70–78 (2012)
15. Girju, R.: Automatic detection of causal relations for Question Answering. In: Proc. ACL 2003 Workshop on Multilingual Summarization and Question Answering, MultiSumQA 2003, vol. 12, pp. 76–83 (2003)
16. Girju, R., Moldovan, D.: Mining Answers for Causation Questions. In: AAAI Symposium on Mining Answers from Texts and Knowledge Bases (2002)
17. Kaplan, R.M., Berry-Rogghe, G.: Knowledge-based acquisition of causal relationships in text. Knowledge Acquisition 3(3), 317–337 (1991)
18. Katz, P., Singleton, M., Wicentowski, R.: SWAT-MP: the SemEval-2007 systems for task 5 and task 14. In: Proc. 4th International Workshop on Semantic Evaluations, SemEval 2007, pp. 308–313 (2007)
19. Kim, S.M., Hovy, E.: Identifying and Analyzing Judgment Opinions. In: Proc. HLT/NAACL 2006, pp. 200–207 (2006)

20. Kozareva, Z., Navarro, B., Vázquez, S., Montoyo, A.: UA-ZBSA: A headline emotion classification through web information. In: Proc. 4th International Workshop on Semantic Evaluations, SemEval 2007, pp. 334–337 (2007)
21. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: Proc. Eighteenth International Conference on Machine Learning, ICML 2001, pp. 282–289. Morgan Kaufmann Publishers Inc., San Francisco (2001)
22. Lee, S.Y.M., Chen, Y., Huang, C.R.: A text-driven rule-based system for emotion cause detection. In: Proc. NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, CAAGET 2010, pp. 45–53 (2010)
23. Lee, S.Y.M., Chen, Y., Li, S., Huang, C.R.: Emotion Cause Events: Corpus Construction and Analysis. In: Proc. Seventh International Conference on Language Resources and Evaluation (LREC 2010). European Language Resources Association (ELRA), Valletta (2010)
24. Lu, C.Y., Lin, S.H., Liu, J.C., Cruz-Lara, S., Hong, J.S.: Automatic event-level textual emotion sensing using mutual action histogram between entities. Expert Systems With Applications 37(2), 1643–1653 (2010)
25. Mohammad, S., Zhu, X., Martin, J.: Semantic Role Labeling of Emotions in Tweets. In: Proc. 5th, ACL Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 32–41 (2014)
26. Mohammad, S.M., Turney, P.D.: Emotions evoked by common words and phrases: using mechanical turk to create an emotion lexicon. In: Proc. NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, CAAGET 2010, pp. 26–34 (2010)
27. Neviarouskaya, A., Aono, M.: Extracting Causes of Emotions from Text. In: International Joint Conference on Natural Language Processing, pp. 932–936 (2013)
28. Neviarouskaya, A., Prendinger, H., Ishizuka, M.: Affect Analysis Model: novel rule-based approach to affect sensing from text. Natural Language Engineering 17(1), 95–135 (2011)
29. Ortony, A., Collins, A., Clore, G.L.: The cognitive structure of emotions. Cambridge University Press (1988)
30. Picard, R.W.: Affective Computing. The MIT Press (1997)
31. Pustejovsky, J., Lee, K., Bunt, H., Romary, L.: ISO-TimeML: An International Standard for Semantic Annotation. In: Proc. the Seventh International Conference n Language Resources and Evaluation (LREC 2010) (2010)
32. Scherer, K.R.: What are emotions? And how can they be measured? Social Science Information 44, 695–729 (2005)
33. Strapparava, C., Mihalcea, R.: Learning to identify emotions in text. In: Proc. 2008 ACM Symposium on Applied Computing, SAC 2008, pp. 1556–1560 (2008)
34. Strapparava, C., Valitutti, A.: WordNet-Affect: an Affective Extension of WordNet. In: Proc. 4th International Conference on Language Resources and Evaluation, pp. 1083–1086 (2004)
35. Tokuhisa, R., Inui, K., Matsumoto, Y.: Emotion classification using massive examples extracted from the web. In: Proc. 22nd International Conference on Computational Linguistics, COLING 2008, vol. 1, pp. 881–888 (2008)
36. Wallach, H.M.: Conditional random fields: An introduction. Tech. rep., University of Pennsylvania (2004)
37. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing Contextual Polarity: An Exploration of Features for Phrase-Level Sentiment Analysis. Computational Linguistics 35(3), 399–433 (2009)
38. Wu, C.H., Chuang, Z.J., Lin, Y.C.: Emotion recognition from text using semantic labels and separable mixture models. ACM Transactions on Asian Language Information Processing (TALIP) 5(2), 165–183 (2006)