Who Feels What and Why?
Annotation of a Literature Corpus with Semantic Roles of Emotions

Evgeny Kim and Roman Klinger
Institut für Maschinelle Sprachverarbeitung
University of Stuttgart, Pfaffenwaldring 5b, 70569 Stuttgart, Germany
evgeny.kim@ims.uni-stuttgart.de
romanklinger@ims.uni-stuttgart.de

Abstract

Most approaches to emotion analysis in fictional texts focus on detecting the emotion expressed in text. We argue that this is a simplification which leads to an overgeneralized interpretation of the results, as it does not take into account who experiences an emotion and why. Emotions play a crucial role in the interaction between characters and the events they are involved in. Until today, no specific corpora that capture such an interaction were available for literature. We aim at filling this gap and present a publicly available corpus based on Project Gutenberg, REMAN (Relational EMotion ANnotation), manually annotated for spans which correspond to emotion trigger phrases and entities/events in the roles of experiencers, targets, and causes of the emotion. We provide baseline results for the automatic prediction of these relational structures and show that emotion lexicons are not able to encompass the high variability of emotion expressions and demonstrate that statistical models benefit from joint modeling of emotions with its roles in all subtasks. The corpus that we provide enables future research on the recognition of emotions and associated entities in text. It supports qualitative literary studies and digital humanities. The corpus is available at http://www.ims.uni-stuttgart.de/data/reman.

Title and Abstract in German

Wer fühlt was und warum?
Annotation eines Literaturkorpus mit Semantischen Rollen von Emotionen


This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/
1 Introduction

Emotions are one of the crucial aspects of compelling narratives (Oatley, 2002; Ingermanson and Economy, 2009; Hogan, 2015). Not only do emotions help readers in literature comprehension (Barton, 1996; Robinson, 2005) but they also improve readers’ abilities of empathy and understanding of others’ lives (Mar et al., 2009; Kidd and Castano, 2013). This makes literature an interesting resource for the study of emotions, hence there is a growing interest in emotion-oriented text analysis among digital humanities scholars.

Emotion analysis and classification is a challenging task which has mostly been tackled with comparatively straight-forward approaches, at least in literary studies. For instance, Kim et al. (2017) and Reagan et al. (2016) show that emotions, recognized with dictionaries or bag-of-words models, serve as features for genre classification in fiction, however, only with limited performance. One reason is, presumably, that such approaches assume linearity of the story and ignore the semantic role structure of emotions: who feels the emotion and why, what caused the emotion, what is the target of it (Scarantino, 2016; Russell and Barrett, 1999). Consider the sentence “Jack is afraid of John because John has a knife”. Following structural approaches to defining emotional episodes, the sentence can be rephrased as “emotion of fear is experienced by Jack (experiencer) because John (target) has a knife (cause)”. Here, dictionary-based or bag-of-words approaches would probably capture that this sentence describes fear, however, would fail in attributing correct semantic roles to John and Jack and we would be forced to assume that their emotional experiences are equal, which is not the case.

To overcome these limitations of dictionary-based and bag-of-words approaches to emotion recognition from literary text, we contribute the corpus REMAN (Relational Emotion Annotation). To the best of our knowledge, this is the first dataset of literary excerpts which has annotations for emotions on a phrase level, for experiencers of each emotion, and for their targets and causes. Our work loosely follows the concept of directed emotions, as defined in FrameNet (Fillmore et al., 2003), and extends the work of Ghazi et al. (2015), who focus on detecting emotion stimulus in the FrameNet exemplary sentences annotated for emotions and causes. Our study is different in terms of the type of texts used for the annotation and the conceptualization of certain emotion components.

Our main contributions are therefore: (1) We present the first resource of fictional texts annotated for emotions, experiencers, causes, and targets. (2) We show that emotion annotation that takes into account not only strong emotion indicators (“afraid”), but also implicit emotions (“shaking fingers”) is valuable for the study of the language of emotions. (3) We provide results of baseline models to predict emotion words and roles separately and (4) show that the prediction performance of all subtasks benefits from joint prediction of experiencer, emotion words, and targets.

2 Related Work

Emotions have strong linguistic markers that define the tone of the text (Johnson-Laird and Oatley, 1989). This allows for different granularities of emotion annotation. The corpus which originates from the ISEAR project (Scherer and Wallbott, 1994) is an example of document-level annotation that includes descriptions of situations in which respondents had experienced various emotions. Examples of sentence-level annotations include the work by Alm et al. (2005), who annotate a corpus of children stories, and Strapparava and Mihalcea (2007), who label news headlines, but without specifying the textual markers of emotion. An early work which includes textual markers of emotions is Aman and Szpakowicz (2007), who annotate blogposts. Wiebe et al. (2005) annotate a corpus of news articles with emotions at a word and phrase level.

Recent works have mainly diverged from plain emotion annotation, following the idea of emotion theorists (Russell, 2003, i.a.) that causes of emotions are inseparable from emotions: Russo et al. (2011) build a corpus of Italian newspaper articles annotated with emotion key words and emotion cause phrases. Both Mei et al. (2012) and Gui et al. (2016) construct emotion-cause-annotated corpora for Chinese. Chen et al. (2010) adopt a rule-based approach based on linguistic patterns to detect emotion causes in the annotated Chinese corpus. Gui et al. (2017) present a question-answering approach to emotion cause extraction, also for Chinese.
When I mentioned the house, he seemed surprised.

All laughed at the mistake, and none louder than the fourth member of the parliament.

Figure 1: Example annotation from Hugo (1885), with one character, an emotion word, and event and cause and experiencer annotations.

Figure 2: Example annotation from Stimson (1943), with two characters who are experiencers of different emotions. Disgust and joy are annotated as a mixture of emotions. Both emotions have the same target.

Fewer works exist for English. Neviarouskaya and Aono (2013) annotate 500 sentences from an online forum with experiencer, emotion, and emotion cause and present a method for extracting linguistic relations between an emotion and its cause. Ghazi et al. (2015) collect exemplary sentences from FrameNet that have cause annotation and implement a model that extracts the causes of emotions. Following a similar approach, Mohammad et al. (2014) annotate Tweets for semantic roles.

Conceptually, our work partially overlaps with the FactBank corpus (Saurí and Pustejovsky, 2009), where “who thinks what” is taken into account as well. However, in contrast to FactBank, we do not predefine event-selecting predicates for emotion causes and targets, as those are defined by the annotators. In this sense, our work is also different from aspect-based sentiment analysis, where aspects of reviewed products are often predefined.

3 Annotation Task

The goal of the REMAN annotation project is to create a dataset of excerpts from fictional texts that are annotated for the phrases that lead to the association of the text with an emotion, the experiencer of the emotion (a character in the text, if mentioned), the target and the cause of the emotion, if mentioned (e.g., an entity, or event). An example of such an annotation is shown in Figures 1 and 2. As it can be seen from these depictions, each annotation includes textual span annotations such as emotions, characters, events, as well as relation annotations that establish relations between different text spans (cause, experiencer, target). In the following, we describe the conceptual background for each annotation layer in detail. The complete annotation guidelines are available online together with the corpus.

3.1 Phrase Annotation

3.1.1 Emotion

We conceptualize emotions as one’s experience that falls in the categories in Plutchik’s classification of emotions, namely anger, fear, trust, disgust, joy, sadness, surprise, and anticipation. In addition, we allow annotation with the class other emotion that covers cases when the emotion expressed in the text cannot be reliably categorized into one of the predefined eight classes. A list of the emotions along with example realizations can be found in Appendix A, Table 5.

Annotators are instructed to prefer span annotations of key words (e.g., “afraid”), except cases when emotions are only expressed with a phrase (e.g., “tense and frightened”) or indirectly (e.g., “the corners of her mouth went down”). Additionally, emotion spans are marked to be intensified (i.e., amplified), diminished (i.e., downtoned) and negated without marking the modifier or including the modifier. Each span is associated with one or more emotions (exemplified in Figure 2).
3.1.2 Entity
We conceptualize entities as mentions of something that has a clear identity of a person, object, concept, state, or event (see Table 6 in Appendix A). Entities are only annotated if they are experiencer, cause, or target of an emotion.

Character An entity that acts as a character in the text. Character annotation should not omit important information (e.g., the annotation of “the man with two rings of the Royal Naval Reserve on his sleeve” is preferred over only annotating “the man”).

Event An event is an occasion or happening that plays a role in the text. Events can be expressed in many ways (see Table 6 in Appendix A for examples from the annotated dataset) and annotators are instructed to label the entire phrases including complementizers or determiners.

Other This is an umbrella concept for everything else that is neither a character nor an event, but fills as relation, described in the following.

3.1.3 Relation Annotation
Relations are semantic links between an emotion and other text spans and can be of type experiencer, cause, and target. In addition, we partially annotate coreferences to link personal pronouns to proper nouns. All relations, except Coreference, can only originate from the emotion annotations.

Experiencer The experiencer relation links an emotion span and entity of type character who experiences the emotion. If the text contains multiple emotions with multiple experiencers, they all are subject to relation annotation.

Target The target relation links an emotion span and entity of any type towards which the emotion experienced by the experiencer is directed. If there are multiple targets of the emotion, then all of them should also be included in the relation annotation. See Figure 2 for the example of a target annotation.

Cause The cause relation links an emotion span and entity of any type, which serves as a stimulus, something that evokes the emotional response in the experiencer. If there are multiple causes for the emotion, then all of them are included in separate relation annotations.

Coreference The annotators are instructed to annotate as an experiencer the character that is the closest to the emotion phrase in terms of token distance. If the closest mention of the character is a pronoun and the text provides a referent that has a higher level of specificity than the pronoun (i.e., a proper noun or a noun denoting a group or class of objects), then the annotators are asked to resolve the coreference.

4 Corpus Construction and Annotation
4.1 Selection
The corpus of 200 books is sampled from Project Gutenberg\(^1\). All books belong to the genre of fiction and were written by authors born after the year 1800. More detailed information on the distribution of authors and genres can be found in Appendix B.

We sample consecutive triples of sentences from this subsample of books. A triple is accepted for inclusion for annotation if the middle sentence includes a word from the NRC dictionary (Mohammad and Turney, 2013). We consider this middle sentence the target sentence and the annotators are instructed to label emotions in this second sentence only. Experiencers, causes and targets are annotated in the whole sentence triple if they refer to an emotion in the target sentence.

The sampling procedure is motivated by our observation that triples of sentences sampled with emotion dictionary show the best coverage in terms of the roles that are associated with the emotion. Ghazi et al. (2015) annotate only one sentence and speculate whether adding one sentence before and after will lead to better results. To check their hypothesis, we conduct a small pre-study experiment by extracting 100 random sentences from the Project Gutenberg with the NRC dictionary and analyze how often the roles of experiencer, cause, and target are found in the target sentence and in the window of up to five sentences before and after. The analysis shows that 98% of the texts include the experiencer in the target

\(^1\)http://www.gutenberg.org/
sentence, while cause and target is found in the target sentence in 67% of the texts. Another 29% of the texts include cause and target in the window of one sentence before and after the target sentence. The remaining texts include cause and target in the window of two (2%), three (1%), and four (1%) sentences around the target sentence. We therefore opt for three-sentence spans as they provide enough information regarding “who feels what and why” without creating unnecessary annotation overhead (cumulatively, 96% of cause and target are found in such sentence triples).

4.2 Annotation Procedure

The annotations were generated in a multistep process, visualized in Figure 3. The people involved in the annotation were either annotators or experts, whose roles did not overlap. The annotations (of spans and relations) were performed by three graduate students of computational linguistics (two native English speakers, one non-native speaker) within a three-month period. Arising questions were discussed in weekly meetings with the experts (the authors of the paper) and the results documented in the annotation guidelines. We used WebAnno\(^2\) (Yimam et al., 2013) as annotation framework. In the following, we discuss the four steps of generating the corpus.

**Step 1: Emotion phrase annotation** The annotators were asked to first decide whether the text expresses an emotion and which emotion that is. If any exists, they label the phrase, which led to their decision. The annotators were instructed to search for emotions that are expressed either as single words or phrases.

**Step 2: Emotion phrase aggregation** In the previous step, each annotator generates set of annotations. In this step, the expert heuristically aggregates all spans that overlap between annotators in a semi-automatic process: Concrete emotions are preferred over the “other-emotion”, annotations with modifier are preferred over annotations without, and shorter spans are preferred over longer spans. Overlapping annotations with different emotion labels are all accepted.

**Step 3: Relation annotation** Annotators are given the same texts they annotated for emotions in Step 1 with the aggregation from Step 2 from all annotators. Therefore, all annotators see the same texts and annotations as input in this step. For each emotion, the task is to annotate entities that are experiencers, targets, or causes of the emotion and establish relations between them. The annotators were instructed to tag only those entities that have a role of an experiencer, cause, and target. The decision on the entity and relation annotation is made simultaneously: For each emotion the annotators find who experiences the emotion (which character) and why (because of event, object, or other character).

**Step 4: Relation annotation aggregation and adjudication** This final step is a manual expert step: Aggregate the relation annotations provided by the annotators. Heuristically, we prefer shorter spans for entities, but guide ourselves with common sense. For instance, consider the phrase “[…] wishing rather to amuse and flatter himself by merely inspiring her with passion”. “Wishing” is labelled as emotion. One annotator tagged “to amuse and flatter himself by merely inspiring her with passion” as event, another tagged only “by merely inspiring her with passion”, which is incomplete, as the target of the emotion is the act of amusing and flattering oneself.

Note that we do not discard the rejected annotations but publish all annotations of all annotators.

---

\(^2\)https://webanno.github.io/webanno/
a ↔ b  b ↔ c  a ↔ c

<table>
<thead>
<tr>
<th>Type</th>
<th>(\kappa)</th>
<th>strict (F_1)</th>
<th>fuzzy (F_1)</th>
<th>(\kappa)</th>
<th>strict (F_1)</th>
<th>fuzzy (F_1)</th>
<th>(\kappa)</th>
<th>strict (F_1)</th>
<th>fuzzy (F_1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>anger</td>
<td>.25</td>
<td>25</td>
<td>39</td>
<td>.15</td>
<td>15</td>
<td>38</td>
<td>.18</td>
<td>18</td>
<td>33</td>
</tr>
<tr>
<td>anticipation</td>
<td>.09</td>
<td>9</td>
<td>23</td>
<td>.07</td>
<td>7</td>
<td>20</td>
<td>.18</td>
<td>18</td>
<td>39</td>
</tr>
<tr>
<td>sadness</td>
<td>.32</td>
<td>32</td>
<td>41</td>
<td>.22</td>
<td>23</td>
<td>41</td>
<td>.19</td>
<td>20</td>
<td>29</td>
</tr>
<tr>
<td>joy</td>
<td>.38</td>
<td>39</td>
<td>50</td>
<td>.40</td>
<td>40</td>
<td>55</td>
<td>.28</td>
<td>28</td>
<td>44</td>
</tr>
<tr>
<td>surprise</td>
<td>.26</td>
<td>26</td>
<td>43</td>
<td>.22</td>
<td>23</td>
<td>33</td>
<td>.27</td>
<td>27</td>
<td>37</td>
</tr>
<tr>
<td>trust</td>
<td>.17</td>
<td>17</td>
<td>26</td>
<td>.14</td>
<td>14</td>
<td>21</td>
<td>.12</td>
<td>13</td>
<td>32</td>
</tr>
<tr>
<td>disgust</td>
<td>.23</td>
<td>23</td>
<td>41</td>
<td>.10</td>
<td>10</td>
<td>26</td>
<td>.19</td>
<td>19</td>
<td>31</td>
</tr>
<tr>
<td>other</td>
<td>.07</td>
<td>7</td>
<td>7</td>
<td>.06</td>
<td>6</td>
<td>11</td>
<td>.08</td>
<td>8</td>
<td>22</td>
</tr>
<tr>
<td>fear</td>
<td>.35</td>
<td>35</td>
<td>48</td>
<td>.28</td>
<td>28</td>
<td>35</td>
<td>.28</td>
<td>28</td>
<td>41</td>
</tr>
<tr>
<td>character</td>
<td>.63</td>
<td>63</td>
<td>68</td>
<td>.48</td>
<td>49</td>
<td>51</td>
<td>.48</td>
<td>48</td>
<td>54</td>
</tr>
<tr>
<td>event</td>
<td>.29</td>
<td>31</td>
<td>60</td>
<td>.09</td>
<td>10</td>
<td>30</td>
<td>.32</td>
<td>34</td>
<td>44</td>
</tr>
<tr>
<td>other</td>
<td>.11</td>
<td>12</td>
<td>28</td>
<td>.11</td>
<td>11</td>
<td>18</td>
<td>.20</td>
<td>21</td>
<td>23</td>
</tr>
<tr>
<td>experiencer</td>
<td>.65</td>
<td>73</td>
<td>73</td>
<td>.48</td>
<td>57</td>
<td>57</td>
<td>.46</td>
<td>55</td>
<td>55</td>
</tr>
<tr>
<td>cause</td>
<td>.20</td>
<td>28</td>
<td>28</td>
<td>.34</td>
<td>39</td>
<td>39</td>
<td>.26</td>
<td>26</td>
<td>32</td>
</tr>
<tr>
<td>target</td>
<td>.27</td>
<td>28</td>
<td>36</td>
<td>.18</td>
<td>29</td>
<td>29</td>
<td>.14</td>
<td>14</td>
<td>28</td>
</tr>
</tbody>
</table>

Table 1: Pairwise inter-annotator agreement for the phrase annotation and relation annotation. \(F_1\) is in %. Regarding the relation scores, in strict \(F_1\), a TP holds if the relation annotation is the same and the entity it points to has the same label and span. In fuzzy \(F_1\), a TP holds if the relation annotation is the same and the entity it points to is the same, but the span boundary of the entity is not necessarily the same.

5 Results

In the following, we first discuss annotation statistics and then provide results of baseline models trained on our resource.

5.1 Inter-annotator Agreement and Consistency of the Annotations

We use pairwise Cohen’s Kappa coefficient (\(\kappa\)) on the token level and \(F_1\) on a phrase level with exact and fuzzy match to calculate the agreement of the phrase annotation and \(F_1\) to estimate the agreement of the relation annotation. For \(F_1\) calculation, we use two approaches: strict that requires labels and spans to be identical, and fuzzy that accepts an annotation to be a true positive if the annotations of two annotators overlap by at least one token. Table 1 reports the agreement scores for emotion, entity, and relation annotations between each pair of annotators.

Joy has the highest number of instances (336) and the highest agreement scores (average \(\kappa = 35\)), followed by fear (\(\kappa = 0.30\)) and sadness (\(\kappa = 0.24\)). Other emotion has the lowest agreement with average \(\kappa = 0.07\). For entity annotation, especially for character annotation, the agreement is higher, with the highest agreement between two annotators being \(\kappa = 0.63\). The agreement on the event and other entities is low (\(\kappa = 0.23 \text{ and } 0.14 \text{ and } F_1 = 25 \text{ and } 14\), respectively). This is presumably the case because event annotations are often comparably long. This also holds, to a lower extend, for character annotations. If we allow partial overlaps to count as a match, the average \(F_1\) increases to 57 for character (an increase of 4 percentage points (pp)), 44 for event (increase by 19 pp), and 23 for other category (increase by 9 pp).

For relation annotations, higher agreement scores are also observed with fuzzy evaluation (\(F_1\) increase for experiencer, cause and target by 10 pp, 7 pp, and 12 pp respectively). These results are in line with previous studies on emotion cause annotation (Russo et al., 2011), and show that disagreements mainly come from the different spans of the entities, though they overlap.

5.2 Difficulties with Measuring IAA

As we showed in Section 5.1, the agreement across all annotation layers is comparably low. There are several reasons for that. Indeed, emotion annotation is highly subjective, but it is not the only subjective category. The cause and target of the emotion are not always clearly recognizable in the text and are also
Table 2: Corpus statistics for emotions annotations. Columns indicate the number of times each emotion was annotated. “in NRC1” shows how many of 1 token annotations are in the NRC dictionary (percentage is given relative to 1 token annotations). “in NRC2” shows how many multi-word annotations include at least one word from NRC.

Table 3: Corpus statistics for relation annotation. Columns indicate the number of times each role was assigned to an entity and how often the respective emotions are in relation to the entity.

subjective categories (two annotators may find two different causes for the same emotion), hence the low agreement scores across all categories. The only exception are experiencer annotations, which are the most reliable among all annotations and match the substantial agreement scores of character annotation (the only type of entities that can be involved in an experiencer relation).

We illustrate the difficulties the annotators face when annotating emotions with roles with the following example: “they had never seen ... what was really hateful in his face; ... they could only express it by saying that the arched brows and the long emphatic chin gave it always a look of being lit from below ...” All annotators agree on the character (“they”) and the emotion (“hateful” expressing disgust). Similarly, both annotators agree that the disgust is related to properties of the face which is described, however, one annotator marks “his face” as target, the other marks the more specific but longer “the arched brows and the long emphatic chin gave it always a look of being lit from below” as cause.

If we abstract away from the text spans, both annotators agree that the emotion of disgust has something to do with “his face”, however they disagree on the target annotation and the cause annotation. So, though conceptually, the annotations by two people are similar, this is not captured by our calculation of inter-annotator agreement.
5.3 Corpus Details

Tables 2 and 3 show the total number of annotations for each category. The REMAN corpus consists of 1720 sentence triples, 1115 of which include an emotion. For the emotion category, joy has the highest number of annotations, while anger has the lowest number of annotations. In most cases, emotion phrases are single tokens (e.g., “monster”, “irksome”), out of which 47% on average are found in the NRC dictionary. Other emotion has the largest proportion of annotations that span more than one token (36% out of all annotations in this category), which is in line with our expectation that lower levels of specificity for emotion annotation make it more difficult to find a single token that indicates an emotion.

For entities, character has the highest number of annotations. As one can see, the experiencer relation dominates the dataset (48%), followed by target (28%) and cause (24%) relations. Note that each character can experience more than one emotion, hence the difference between the number of characters and the experiencers. Table 3 also shows how many times each emotion triggered certain relation. In this sense, joy has triggered the most experiencer and cause relations, which is still related to the prevalence of the annotations for this emotion in the dataset.

6 Baseline Model

We provide a baseline for automatically predicting the structures we annotated. For this first model, we map the relations to span prediction tasks. This is feasible because characters, entities, and other were only annotated if they fill one of the roles, experiencer, target, or cause. Therefore, the prediction task boils down to a sequence prediction task of emotion phrases (for the different emotions) and the potential mentions of experiencers, targets, and causes. Note that we lose the actual relation information in this simplification.

Consider the example depicted in Figure 1: The phrase “I mentioned the house” is labelled as an event and is assigned a role of a cause for the emotion of surprise, and the word “he” is labelled as a character and is assigned a role of an experiencer of the same emotion. We represent these relationships by tagging “I mentioned the house” as cause and “he” as experiencer using inside-outside-beginning (IOB) encoding capturing the text spans that are linked by relations with an emotion.

We use two sequence labelling models, conditional random fields (CRF) (Lafferty et al., 2001) and bidirectional long short-term memory networks with a CRF layer (biLSTM-CRF), which both provide a good performance in sequence prediction tasks (Benikova et al., 2014; Huang et al., 2015). In addition, we analyze the difficulty of predicting the emotion for a full sentence triple, independent of segments. In the following, we further specify the experimental setting in detail.

6.1 Experimental setting

Experiment 1: Coarse-grained emotion classification In this experiment, the task is to classify the emotions which occur in the sentence triple which forms the instance under consideration. This is therefore a coarse abstraction of the structured prediction tasks presented in this paper. However, this constitutes the most straight-forward task in emotion analysis. We use a dictionary-based approach and a bag-of-words-based classifier.

For the dictionary-based classification, we take the intersection between the words in the triple and NRC dictionary and assign the triple with the corresponding emotion labels. The F1 score is calculated by comparing the set of labels predicted by dictionaries against the set of gold labels for each triple. The gold labels come from the annotation of words and phrases within each triple. For the BOW approach, we convert each triple into a sparse matrix using all words in the corpus as features. We then classify the triples with a multi-layer perceptron with three hidden layers, 128 neurons each, with an initial learning rate of 0.01 that is divided by 5 if the validation score does not increase after two consecutive epochs by at least 0.001.

Experiment 2: Fine-grained emotion and role detection In this experiment, we evaluate the performance of fine-grained emotion and role (experiencer, target, and cause) prediction in a sequence labelling fashion, as described above. We instantiate separate CRF and biLSTM-CRF models for each relation, as
The CRF uses part-of-speech tags (detected with spaCy\(^3\) (Honnibal, 2013)), the head of the dependency, if it is capitalized, and offset conjunction with the features of previous and succeeding words as features. For the emotion category, we use the presence in the NRC dictionary in addition and, for experiencer, the presence in a list of English pronouns. We train for 500 iterations with L-BFGS (Liu and Nocedal, 1989) and L1 regularization.

The biLSTM-CRF model uses a concatenated output of two biLSTM models (one trained on word embeddings with dimension 300, and one trained on character embeddings from the corpus with dimension 100) as an input to a CRF layer. The word embeddings that we use as input are pre-trained on Wikipedia\(^4\) using fastText. We use Adam as activation function, a dropout value of 0.5, and train the model for 100 epochs with early stopping if no improvement is observed after ten consecutive epochs.

**Experiment 3: Potential for joint modelling of emotion and role prediction**  
The goal of this experiment is to understand if joint modelling of relations has the chance to contribute over learning each relation separately. To that end, we analyze the potential interactions between predictions with gold labels of all other predictions. Specifically, when training our models, we provide the classifier with the information which sequence of tokens is an experiencer (in the case of emotion phrase prediction) and which sequence of tokens is an emotion (in case of experiencer, cause, and target detection).

### 6.2 Results and Discussion

The results of all the experiments are summarized in Table 4. We evaluate our models in the same way we use $F_1$ for inter-annotator agreement: Firstly, by accepting a TP if it is exactly found (exact) and secondly, if at least one token is overlapping with the annotation (fuzzy).

**Experiment 1** Emotion classification with dictionaries and bag of words show mediocre performance. The recall with the dictionary classification is comparably high ($F_1 = 83$), which is due to the fact that texts were sampled using these dictionaries. However, as we said earlier, annotators are free to label any words and phrases as emotion-bearing, hence low precision and $F_1$ score. The MLP with BOW features does not perform better but shows increased precision at the cost of lower recall. A possible reason is that each triple may contain only one word that expresses the emotion with the rest of the words being neutral.

**Experiment 2** As results of this experiment show, the recall is low for all categories. A presumable reason is, as discussed in Section 5, that substantial number of emotion annotations are words or phrases that are not found in the NRC dictionary. On average, only 46\% of emotion annotations are single tokens

---

\(^3\)https://spacy.io/

\(^4\)As available at https://github.com/facebookresearch/fastText (Bojanowski et al., 2017)
that can be found in the NRC dictionary, but for some emotions this number is much lower (only 14% of *anticipation* annotation). The same applies to cause and target categories, as in most cases these are long spans of text (e.g., 94% of events are multiword expressions). This explains zero $F_1$ score for cause prediction with CRF and biLSTM-CRF and a better performance for target prediction with CRF, taking into account that most target relations is triggered by characters, 75% of which are single tokens (see Table 3).

The highest precision and $F_1$ across all categories is observed for the emotion category with biLSTM-CRF ($F_1 = 43$ and fuzzy $F_1 = 48$). The strict $F_1$ is by 12 pp higher than predicted with dictionaries and with BOW in text classification experiment.

The experiencer category is second best, however, the recall for this category is still very low. This can be explained by the fact that experiencers are expressed in the text mostly as personal pronouns. As far as the number of personal pronouns in our texts is relatively low (13% of all tokens in a sentence on average), and only a small fraction of them act as experiencers (< 1% of all tokens in a sentence on average), the classifier cannot learn when an entity is an experiencer or not.

**Experiment 3** The goal of this experiment was to estimate if joint modelling of emotion and roles is feasible. We observe that, for the emotion category, $F_1$ increases by 5 pp in strict and by 9 pp in fuzzy evaluation if we provide the classifier with the information, which sequence of tokens is an experiencer. For experiencer prediction, $F_1$ increases by 20 pp in strict and by 22 pp in fuzzy evaluation if we tell the classifier which word or sequence is labelled as emotion. These results indicate the complementarity of both categories. A qualitative study on a subsample of linguistic properties of emotions and experiencers shows that when the emotion expression and experiencer are parts of the same phrase (verb or adjectival phrase), the emotion word serves as a head to the word that represents an experiencer. Hence, the classifier is able to partially learn that any phrase that is a part of the emotion phrase, whose head is a personal pronoun or a proper name, is a potential experiencer.

The same applies to experiencer: if the head of the governing phrase is an emotion, then the head of the current phrase is a potential experiencer. However, due to variability of emotion expressions, this cannot always be the case.

7 Discussion, Conclusion, and Future Work

As evaluation of inter-annotator agreement and sequence labelling results of the baseline model show, the task of annotating emotions and corresponding roles, as well as their subsequent prediction is a difficult one. A high variability of emotion expressions (see Table 5) and a variability of cause and target expressions make it hard. At the same time, the resource we present provides interesting and valuable insights in the language of emotion expression and, therefore, is useful to the community of linguists who are interested in the study of linguistic properties of emotions.

However, we also note that developing such a resource has its limitations: Due to the subjective nature of emotions, it is challenging, if not impossible, to come up with an annotation methodology that would lead to less disparate annotations, especially if in addition to emotion, other categories should be annotated together with roles. That is in line with previous research. For instance Schuff et al. (2017) and Russo et al. (2011) found that aggregating labels by multiple annotators without a majority vote procedure but by merging is easier to model computationally.

We tackle this problem by employing a multi-step procedure that helps to improve the agreement of the relation annotation. This does not help in the emotion annotation itself, but helps in the role assignment. The introduction of our multi-step annotation procedure lead to an increased inter-annotator agreement for experiencer and cause annotations by 13 pp and 5 pp in strict evaluation. This indicates that the task seems easier to annotators if they perform role assignment with predefined emotion annotations.

Another difficulty arises from the nature of the texts we work with. Fictional texts are highly metaphoric and full of allusions and metonymies, which requires thoughtful reading (often reading between the lines) and a broader context. However, this is something that our annotators do not have: all the context they have at their disposal is a triple of sentences, each of which can rely on information that is available in other parts of the book, but not in the annotation unit. Therefore, it is not always possible to annotate the
cause, target, or even the experiencer. This is a trade-off: On the one side, we did not want to annotate full books to have a representative corpus. On the other side, we might not have provided sufficient context. Future work will therefore aim at better understanding how to preselect the relevant context that is needed for reliable annotation and secondly use such knowledge for a follow-up annotation project.

Nonetheless, we are confident that the dataset we present is useful to linguists and digital humanities scholars, as it contains valuable information about complex interactions of emotions, characters, and events in fictional texts, and gives interesting insights into the language of emotion expression in general.

Last but not least, the dataset constitutes a difficult task for structured prediction, as our baseline analysis has shown. Our experiments suggest that the prediction of emotions with their roles is a task that should be tackled with joint models. Therefore, this corpus adds an interesting relation extraction task to the set of existing challenges.

Acknowledgements

This research has been conducted within the CRETA project (http://www.creta.uni-stuttgart.de/) which is funded by the German Ministry for Education and Research (BMBF) and partially funded by the German Research Council (DFG), projects SEAT (Structured Multi-Domain Emotion Analysis from Text, KL 2869/1-1). We thank Laura-Ana-Maria Bostan, Sebastian Padó and the CRETA consortium for fruitful discussions.

References


## A Concepts used for Phrase Annotation

Table 5 provides a list of concepts defined for the annotation of emotions, modifiers, and entities. The Examples column contains examples of annotations from the final corpus. Table 6 provides examples of linguistic realization of entities along with examples from the annotated dataset.

<table>
<thead>
<tr>
<th>Concept</th>
<th>Value</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emotion</td>
<td>Anger</td>
<td>angry, defend themselves by force, break your little finger, loss of my temper</td>
</tr>
<tr>
<td></td>
<td>Anticipation</td>
<td>want, wish, wholly absorbed, looked listlessly round, wholly absorbed</td>
</tr>
<tr>
<td></td>
<td>Disgust</td>
<td>repellent, cheap excitement, turn away from, beg never to hear again</td>
</tr>
<tr>
<td></td>
<td>Fear</td>
<td>horrified, tense and frightened, shaking fingers</td>
</tr>
<tr>
<td></td>
<td>Joy</td>
<td>cheerful, grateful, boisterous and hilarious, violins moved and touched him</td>
</tr>
<tr>
<td></td>
<td>Sadness</td>
<td>failed, despair, the cloudy thoughts, staring at the floor</td>
</tr>
<tr>
<td></td>
<td>Surprise</td>
<td>perplexing, suddenly, petrified with astonishment, loss for words, with his mouth open</td>
</tr>
<tr>
<td></td>
<td>Trust</td>
<td>honor, true blue, immeasurable patience</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>careful, brave, had but a tongue, break in her voice, bit deeply into his thumb</td>
</tr>
</tbody>
</table>

| Modifier | strong | I loved her the more |
|          | weak   | with a little pity |
|          | negated | could not be content |

| Entity | character | the chairman of the board |
|        | event     | marry a man I did not love, because of his gold |
|        | other     | Lily’s beauty |

### Table 5: Concepts used for the phrase annotation layer.

<table>
<thead>
<tr>
<th>Entity type</th>
<th>Linguistic realiz.</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Character</td>
<td>noun phrase</td>
<td>his son</td>
</tr>
<tr>
<td></td>
<td>adjectival phrase</td>
<td>old man</td>
</tr>
<tr>
<td>Event</td>
<td>verb phrase</td>
<td>Mrs. Walton had got another baby.</td>
</tr>
<tr>
<td></td>
<td>adverbal phrase</td>
<td>Jesus spoke unkindly to his mother when he said that to her.</td>
</tr>
<tr>
<td></td>
<td>prepositional phrase</td>
<td>[...] giving her up.</td>
</tr>
<tr>
<td></td>
<td>clause</td>
<td>[...] what she said to him [...]</td>
</tr>
<tr>
<td></td>
<td>noun phrase</td>
<td>the journey</td>
</tr>
<tr>
<td>Other</td>
<td>adjectival phrase</td>
<td>[...] old age [...]</td>
</tr>
<tr>
<td></td>
<td>noun phrase</td>
<td>[...] the heavens and the earth.</td>
</tr>
<tr>
<td></td>
<td>tense phrase</td>
<td>She was the only treasure on the face of the Earth that my heart coveted.</td>
</tr>
</tbody>
</table>

### Table 6: Typical linguistic realization of entities.
B  Genre and author composition

<table>
<thead>
<tr>
<th>Subject headings</th>
<th>Most frequent author</th>
<th># texts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fiction, Christian fiction</td>
<td>MacDonald George</td>
<td>178</td>
</tr>
<tr>
<td>Historical fiction (translations), Epic literature</td>
<td>Hugo Victor</td>
<td>107</td>
</tr>
<tr>
<td>Social fiction</td>
<td>Dostoevsky Fyodor</td>
<td>63</td>
</tr>
<tr>
<td>Domestic fiction, Single women</td>
<td>Gissing George</td>
<td>45</td>
</tr>
<tr>
<td>Young men, Bildungsroman</td>
<td>Thackeray William</td>
<td>42</td>
</tr>
<tr>
<td>Love stories</td>
<td>James Henry</td>
<td>38</td>
</tr>
<tr>
<td>Didactic fiction</td>
<td>Eliot George</td>
<td>36</td>
</tr>
<tr>
<td>Political fiction</td>
<td>Atherton Gertrude Franklin Horn</td>
<td>35</td>
</tr>
<tr>
<td>Historical fiction (translations), France</td>
<td>Dumas Alexandre</td>
<td>35</td>
</tr>
<tr>
<td>German fiction (translations), Social classes</td>
<td>Freytag Gustav</td>
<td>22</td>
</tr>
</tbody>
</table>

Table 7: Most frequent subject headings and authors in the corpus. Subject headings are taken from Project Gutenberg metadata and are shortened for readability.

C  Excerpt from the corpus file

Figure 4: Excerpt from REMAN corpus.