The Sentiment Corpus of App Reviews with Fine-grained Annotations in German (SCARE)

Short manual

Mario Sänger[1]

1 Introduction

This file archive contains the SCARE corpus, a set of fine-grained annotations of mobile application reviews from the Google Play Store. In total, the corpus consists of annotations for 1,760 application reviews with 2,487 aspects and 3,959 subjective phrases annotated. Furthermore, a data set with over 800,000 user reviews from 11 application categories is available. A detailed description of the corpus is available in the paper

> Mario Sänger, Ulf Leser, Peter Adolphs, Steffen Kemmerer, Roman Klinger
> SCARE – The Sentiment Corpus of App Reviews with Fine-grained Annotations in German. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. Portorož, Slovenia. May 2016

Please cite the paper as follows when using the corpus in your work:

```
@inproceedings{Saenger2016,
  author    = {Mario S\"anger and Ulf Leser and Steffen Kemmerer
               and Peter Adolphs and Roman Klinger},
  title     = {{SCARE -- The Sentiment Corpus of App Reviews with
                Fine-grained Annotations in German}},
  booktitle = {Proceedings of the Tenth International Conference on
               Language Resources and Evaluation (LREC'16)},
  year      = {2016},
  month     = {May},
  address   = {Portoro\v{z}, Slovenia},
  publisher = {European Language Resources Association (ELRA)}
}
```

The original location of this corpus is at `http://www.romanklinger.de/scare/`. If you have any questions regarding the corpus please send an email to *scare@romanklinger.de*.

[1] saengerm@informatik.hu-berlin.de

2 File overview

After you extracted the downloaded archive, the folder structure is as follows:

1. `SCARE-corpus/annotations`
   the files with the annotations of corpus
2. `SCARE-corpus/documents`
   the original annotation guidelines (in German) provided to the annotators and the original paper preprint
3. `SCARE-corpus/dictionaries`
   the emoticon, smiley and negation word dictionary used to provide the prediction baseline within the original paper

The text of the reviews are not part of this data publication. Please refer Section 5 for further information.

3 Annotations

The folder `SCARE-corpus/annotations` contains the corpus, i.e., the annotations of mobile application reviews in a set of files. The corpus is separated into the 11 application categories the review originate from. The categories are *instant messengers*, *fitness tracker*, *social networks*, *games*, *news applications*, *alarm clocks*, *navigation apps*, *office tools*, *weather apps*, *sport news* and *music players*.

For each of these categories, three files with the extensions `.txt`, `.csv`, `.rel` exist. The formats of these files are explained in the following:

<category>.txt
The `.txt` files contain the review texts. Each file consist of two tabular-separated columns:

#	Name	Description
1	`Internal-ID`	The internal identificator of the review
2	`Review text`	The complete text of the review, i.e. the title of the assessment as well as the evalution text concatenated with "||"

<category>.csv
Each of the .csv file contains the offset information for subjective (evaluative) phrases and aspect mentions. Each file contains eight tabular-separated columns:

#	Name	Description
1	`Class`	The class of the entity. This can either be *subjective* or *aspect*.
2	`Review-ID`	The id of the review (from the `.txt` file) the phrase originates from
3	`Left`	The left offset of the phrase / aspect within the review text
4	`Right`	The right offset of the phrase / aspect within the reviewtext
5	`String`	The exact string representation of the phrase / aspect
6	`Phrase-ID`	A unique (internal) identifier of the phrase / aspect
7	`Polarity`	The polarity of a subjective phrase. This can either be *Positive*, *Negative* or *Neutral*. For aspects this column is always set to *Neutral*.
8	`Relation`	This column describes the relationship of an aspect to the main application discussed in the review. The relationship can either be *Related* or *Foreign*. For subjective phrases this column is always set to *Related*.

<category>.rel
Each of the `.rel` files contains the information which of the subjective phrases in the `.csv` files are related to which aspect (also from the `.csv` files). It consists of five tabular-separated columns:

#	Name	Description
1	`Relation-ID`	A unique (internal) identificator of the relation.
2	`Aspect-ID`	The internal identificator of the target aspect, that is evaluated
3	`Subj-ID`	The internal identificator of the subjective phrase, which gives an evaluation of the target aspect
4	`Aspect-String`	The exact string representation of the target aspect
5	`Subj-String`	The exact string representation of the subjective phrase

4 Dictionaries

The folder `SCARE-corpus/dictionaries` contains the smiley, emoticon and negation word dictionaries we used to provide the prediction baseline. We considered a smiley to be the representation of a facial expression or mood by several characters. For a emoticon, however only a single font or UTF-8 character is used. Both dictionaries are separated in positive and negative expression forms. The smiley dictionary consists of 114 positive and 118 negative representations. Within the emoticon lexicon 28 positive and 57 negative forms are gathered. The negation word dictionary contains a plain list of 14 German terms, which generally carry a negation or imply the absence of certain matters.

5 Licence