The Bielefeld University Sentiment Analysis Corpus for German and English (USAGE)

Short Manual

Roman Klinger*

June 18, 2015

License

Introduction

This file archive contains the USAGE corpus, a resource of annotations for product reviews in German and English. A detailed discription is available in the paper

> Roman Klinger and Philipp Cimiano.
> The USAGE review corpus for fine-grained, multi-lingual opinion analysis
> In: *Proceedings of the Language Resources and Evaluation Conference.*
> Reykjavík. Iceland. May 2014

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

```
@InProceedings{Klinger2014,
  author =   {Roman Klinger and Philipp Cimiano},
  title =   {The USAGE review corpus for fine-grained, multi-lingual opinion analysis},
  booktitle = {Proceedings of the Language Resources and Evaluation Conference (LREC)},
  year =   {2014},
  address =   {Reykjavik, Iceland},
  month =   {May},
  organization = {ELRA},
}
```

The original location of this corpus is
http://dx.doi.org/10.4119/unibi/citec.2014.14.
Additional information might be made available in the future at
https://code.google.com/p/usage-corpus-data-tools/
and/or
http://www.roman-klinger.de/usagecorpus.

In the following, receiving the reviews and the corpus format is explained.

*rklinger@cit-ec.uni-bielefeld.de

File overview

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

- `USAGE-corpus/files`
  the corpus itself
- `USAGE-corpus/documents`
  the original annotation guidelines provided and discussed with the annotators and the original paper preprint
- `USAGE-corpus/crawler`
  a software tool which helps to retrieve the textual reviews from Amazon.com and Amazon.de.

Description of the corpus files

The folder `USAGE-corpus/files` contains the corpus, *i.e.*, the annotations of product reviews in a set of files. The corpus is separated into German and English, each for different products. The prefixes of the files are for German `de-coffeemachine`, `de-cutlery`, `de-microwave`, `de-toaster`, `de-trashcan`, `de-vacuum`, `de-washer`, and for English `en-coffeemachine`, `en-cutlery`, `en-dishwasher`, `en-microwave`, `en-washer`, `en-trashcan`, `en-vacuum`, and `en-toaster`.

For each of these prefixes, files with the extensions `.txt`, `-a1.csv`, `-a2.csv`, `-a1.rel`, and `-a2.rel` exist. The formats of these files are explained in the following.

.txt

Each `.txt` consists of three tabular-separated columns:

```
Internal-ID    Amazon-Product-ID    Amazon-Review-ID
```

The Internal-ID is unique throughout the whole corpus. The Amazon product ID is structured such that the product site can be opened by `http://www.amazon.de/gp/product/<Internal-ID>` for German and `http://www.amazon.com/gp/product/<Internal-ID>` for English, *e.g.*, `http://www.amazon.de/gp/product/B00ET00LEC` and `http://www.amazon.de/gp/product/B000ALVUM6`[1].

The Amazon review ID is structured such that the review site can be opened by `http://www.amazon.de/review/<AmazonReviewID>` or `http://www.amazon.com/review/<Amazon-Review-ID>/`, *e.g.* `http://www.amazon.de/review/R1HQ73Q7QOX4E` or `http://www.amazon.com/review/R3VLXAC5XYQSMT`.

During the preparation of the corpus for publication, three reviews disappeared from Amazon (either by deletion of the author or by Amazon).[2]. These are marked with the review ID `[disappeared]`.

.csv

Each of the `.csv` file contains the offset information for subjective (evaluative) phrases and aspect phrases. Each file contains 8 tabular-separated columns:

```
class   Internal-ID   left   right   string   phrase-ID   polarity   relation
```

The class is either `aspect` or `subjective` The internal ID refers to the internal ID in the `.txt`-file. The offsets in the plain text representation of the Amazon review are denoted by left and right. The column string is the exact string representation of the phrase. The phrase-ID is a unique identifier of the phrase. The columns polarity can have the values `negative`, `neutral`, `positive` or `unknown` for `subjective` phrases and `unknown` for `aspect` phrases. The column relation denotes if an aspect is `foreign` or `related` (*i.e.*, it is an aspect of the product under review or of another product or similar.)

[1] as of March 28, 2014
[2] The review IDs are available as of March 18, 2014

.rel

Each of the `.rel` file contains the information which of the phrases in the `.csv` files are in relation. It consists of XX tabular-separated columns:

```
Relation-Type    Internat-ID    Phrase-ID1    Phrase-ID2    string1    string2
```

The relation type can be of `TARG-SUBJ` if an aspect and a subjective phrase are denoted to be in relation or `COREF` if one aspect phrase is referencing another aspect phrase. The internal ID is refering to the `.txt` and `.csv` files. The Phrase-ID1 and 2 are refering to the phrases in the `.csv` file. The columns string1 and string2 are the string representations as mentioned in the `.csv` file.

Retrieving the textual content of the reviews

This publication does not contain the Amazon reviews themselves. You need to retrieve them in your own responsibility. We provide a crawler which downloads them for you.[3] The folder `USAGE-corpus/crawler` contains this small piece of software (implemented in Java and partly in Scala). Future changes to this software might be made available at `https://code.google.com/p/usage-corpus-data-tools/`.

Please install Maven and run (in the folder in which the `pom.xml` is located:

```
$ mvn compile
$ mvn assembly:single
```

After that, the folder `USAGE-corpus/crawler/target` contains the file `recrawling-0.5-jar-with-dependencies.jar`. You can start this file directly via

```
$ java -jar recrawling-0.5-jar-with-dependencies.jar
```

or via

```
$ bin/crawl.sh
```

As parameters, please specify the input file (which is one of the `.txt` files mentioned above), the domain (`de` for German and `com` for English), and an output file. An example call is therefore (from the folder `USAGE-corpus/crawler`

```
$ bin/crawl.sh ../files/de-vacuum.txt de ../files/de-vacuum-text.txt
```

After calling that for each of the input files, for both German and English, the output files are augmented by additional columns containing the text of the reviews.

We assume that small changes in the retrieved files might occur. We therefore provide a small tool which tries to find the correct offsets based on the string representation in the `.txt` file. You can call that via

```
$ bin/correctOffsets.sh txtfile csvfileWithText
```

We propose that you use this tool only in the case that you recognize small offset issues in the `.csv` file.

In any case of issues with the workflow, do not hesitate to contact us.

[3] Provided under the GPL 2.0, working as is of March 18, 2014