

LFA-11 Corpus with music and smartphones — Documentation

Abstract

This documentation describes the “LFA-11 Corpus”, i.e., a German text corpus with 2,713 promotional texts, professional and user reviews from the music area as well 2,093 blog posts on smartphones. All documents were manually classified by domain experts from the industry with respect to their topic relevance, their language function and their sentiment polarity.

1 Aim of the Corpus

The purpose of the corpus is to provide textual data for the development and evaluation of approaches to language function analysis and the sentiment analysis. Correspondingly, each text in the corpus has been manually classified by language function (personal, commercial, or informational) as well as by sentiment (positive, negative, neutral). These annotations can be learned by a Machine Learning classifier on any kinds and number of features.

2 Download and Installation

The corpus is free for scientific use under the *Creative Commons* license and can be downloaded at <http://tinyurl.com/ijcnlp2011>. The corpus documents are packed in a *tar.gz archive*. For instructions on how to extract such files, see <http://www.gzip.org>.

2.1 File formats

Both the music texts and the smartphone blog posts come together with their annotations in a standard UTF-8 encoded XMI file. These files are preformatted for *Apache UIMA*, which is an implementation of the *Unstructured Information Management Architecture* (Ferrucci and Lally,

Relevance	true	false
<i>music</i>		
Training set	1327 (97.9%)	28 (2.1%)
Validation set	673 (99.1%)	6 (0.9%)
Test set	662 (97.5%)	17 (2.5%)
<i>smartphone</i>		
Training set	561 (53.6%)	486 (46.4%)
Validation set	307 (58.7%)	216 (41.3%)
Test set	287 (54.9%)	236 (45.1%)

Table 1: Distribution of relevance annotations

2004) for the development of natural language processing applications. However, the annotations can be easily imported into arbitrary applications as well.

3 Description

The corpus consists of two separated collections, *music* and *smartphone*. The music collection consists of 2,713 promotional texts, professional and user reviews from the music area, while the smartphone collection contains 2,093 blog posts on smartphones, which were taken from the *Spinn3r* corpus (see <http://www.spinn3r.com>). For both domains, the corpus was split into a training, a validation, and a test set.

For each document, its relevance is annotated, i.e. whether the statements in the text refer to the given topic. The distribution of relevance annotations is given in Table 1. In case of double annotations, the second annotation has been counted. So, the exact frequencies depend on the annotation used.

Further, the language function of a text and the sentiment polarity have been manually classified. Table 2 shows the distribution of language functions over the corpus. The language function annotation is called *Genre* in the corpus documents, because language functions resemble genres with respect to possible applications.

Function	Commercial	Informational	Personal
<i>music</i>			
Training set	127 (9.4%)	707 (52.2%)	521 (38.5%)
Validation set	72 (10.6%)	188 (27.7%)	419 (61.7%)
Test set	68 (10.0%)	269 (39.6%)	342 (50.4%)
<i>smartphone</i>			
Training set	90 (8.6%)	411 (39.3%)	546 (52.1%)
Validation set	36 (6.9%)	208 (39.8%)	279 (53.4%)
Test set	28 (5.4%)	193 (36.9%)	302 (57.7%)

Table 2: Distribution of language function annotations

Sentiment	Positive	Negative	Neutral
<i>music</i>			
Training set	1003 (74.0%)	93 (6.9%)	259 (19.1%)
Validation set	558 (82.2%)	39 (5.7%)	82 (12.1%)
Test set	514 (75.7%)	50 (7.4%)	115 (16.9%)
<i>smartphone</i>			
Training set	205 (19.6%)	104 (9.9%)	738 (70.5%)
Validation set	110 (21.0%)	70 (13.4%)	343 (65.6%)
Test set	84 (16.1%)	80 (15.3%)	359 (68.6%)

Table 3: Distribution of sentiment annotations

Table 3 shows according statistics for sentiment.

4 Annotations

The mentioned document annotations are represented in the XMI files as follows:

- **Metadata:** Meta annotation of the document. It has links to all the following annotations.
- **Relevance:** The relevance of the text with respect to the given topic. Values are true and false.
- **Genre:** The language function of the text. Texts can have a commercial, informational or personal function.
- **Opinion:** The sentiment polarity of the text, i.e. positive, negative or neutral.

If a document was annotated twice (see below), it has two annotations of each kind.

5 Compilation and Annotation Process

The smartphone source documents were retrieved via queries on a self-made Lucene Index,

which was built for the Spinn3r corpus. On the other hand, the music source documents were manually selected and prepared by employees of a company from the digital asset management industry. Afterwards, two employees annotated the plain document texts of both the music and the smartphone part with a appropriate tool. Annotation guidelines had been written before. About 20% of the music documents and 40% of the smartphone documents were annotated twice in order to estimate the interannotator agreement. For the language function annotation, this resulted in $\kappa_m = 0.78$ (music) and $\kappa_s = 0.67$ (smartphone) of Cohen’s Kappa (Carletta, 1996), which constitutes *substantial agreement*.

In a postprocessing step, the documents were exported to an XMI file which is conform to the format of *Apache UIMA*.

References

- Jean Carletta. 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 22: 249–254.
- David Ferrucci and Adam Lally. 2004. UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment. *Natural Language Engineering*, 10(3–4): pages 327–348.