TESTACCORD Databank

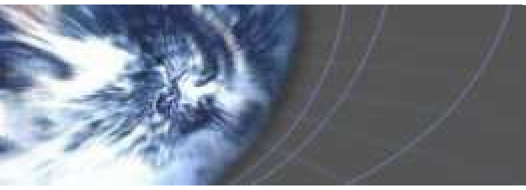
OPINION data set : presentation

Jean-Yves Antoine[1], **Jeanne Villaneau**[2]

[1]**LI – Université François Rabelais de Tours** (EA 6300)
[2]**IRISA** (UMR 6074)

Université François Rabelais Tours

http://www.info.univ-tours.fr/~antoine/parole_publique/



Introduction

TESTACCORD.OPINION is a data set of *TestAccord*, a data bank of various annotation sets dedicated to the experimental study of data reliability and inter-coders agreement in the framework of Natural Language Processing.

TESTACCORD.OPINION addresses the annotation of opinion annotation on a corpus of film reviews made by ordinary people on French websites. More precisely, it concerns the subjective annotation of the polarity and the sentiment strength conveyed by every sentence of the film reviews. This document presents into details the data set.

TestAccord.Opinion : annotated corpus

This data set results from the annotation of a corpus of film reviews. The reviews were written by ordinary people and correspond to relatively short texts selected from two dedicated French websites (www.senscritique.com and www.allocine.fr). The corpus contains 183 sentences. All reviews concern the same French movie. We asked 27 subjects to attribute independently an opinion value to every sentence through a 5-items scale of values (sse table just below).

Coding value	Meaning	Opinion polarity	Sentiment strength
-2	strongly negative	negative	Strong
-1	moderately negative	negative	moderate
0	neutral opinion	neutral	None
1	moderately positive	positive	moderate
2	strongly positive	positive	Strong

This opinion scale encompasses the *polarity* and *sentiment strength* dimensions. It enables to compare without methodological bias an annotation with 3 coding categories (*polarity*: negative, positive, neutral) and the original 5-categories (*polarity+strength*) annotation. The annotation guide only gave the coders the coding value and its associated meaning.

The subjects (11 men / 14 women) were adult people (average age: 31.6 years). All the coders have a superior level of education (at least, high-school diploma), they did not know each other and worked separately during the annotation task. Only four of them had a prior experience in corpus annotation.

The annotation was conducted as follows: the coders were not trained but were given precise annotation guidelines providing some explanations and examples on the emotional values they had to use. The coders achieve the annotation once, without any restriction on time. They had to rely on their own judgment, without considering any additional information. Sentences were given in a random order to investigate an out-of-context perception of emotion.

We conducted a second experiment where the order of the presented sentences was those of the original film review, in order to study the influence of the discourse context. As a result, two different annotations are found in this data set.

The criterion of data significance – at least five chance agreements per category – proposed by (Krippendorff, 2004) is greatly satisfied for the valence annotation (3 coding categories). It is approached on the complete annotation where we can assure 4 chance agreements per category.

TestAccord.Opinion : distributed data

The data set is distributed as an Calc Open Office file (`.ods`). Two sets of test data are actually distributed, which correspond to two different sheets:

▪ Hors contexte	annotations on the sentences in a random ordrer
▪ Contexte	annotations with the sentence order of the film review

Every sheet is made of 26 columns and 183 lines. Each line corresponds to a specific sentence which is presented in first column. The next 25 columns are corresponds to the coders' annotations.

The annotations are respecting the 5-classes annotation scheme. You can translate it directly to a valence annotation (3-classes) by merging the -2 and -1 classes into a negative category as well as the 1 and 2 classes in a positive one.

Creative Commons Licence

TestAccord is freely distributed under a *Creative Commons* CC-BY-SA licence.

This means that you must respect the following conditions of use :

- *BY : attribution* - Licensees may copy, distribute, display and perform the work and make derivative works based on it only if they give the author or licensor the credits in the manner specified by these..

- *SA : share-alike* - Licensees may distribute derivative works only under a license identical to the license that governs the original work.

References

- Krippendorff K. (2004). *Content Analysis: an Introduction to its Methodology*. Chapter 11. Sage: Thousand Oaks, CA.