QUOTATIONS ANNOTATED FOR SENTIMENT MULTILINGUAL SUBJECTIVITY ANALYSIS: GOLD STANDARD DATA SET

1 BASIC INFORMATION

1.1 The data set composition

The data represents a set of 1590 of language quotations (reported speech) manually annotated for sentiment (POSitive, NEGative, OBJective/neutral) towards entities mentioned inside the quotation. Objective is the default, meaning that the absence of a label can be interpreted as OBJ. Each of the quotations is characterized by:

- the news ID (e.g. dailymail-be691eeeedee8e7d24eca8299e84937c)
- the quotation itself (note that the sentiment value refers to the entity mentioned inside the quotation, and not to the entire text of the quotation)
- source name
- source ID
- target ID
- target name
- the sentiment mark-up by a pair of annotators (Ann1...Ann4)
- An agreement label (TRUE or FALSE) between the annotators

Source Name refers to the person who issued the quotation (e.g. Angela Merkel said: "..."). Target Name refers to the entity mentioned inside the quotation, i.e. the entity whose sentiment value we are interested in (e.g. " ... Tony Blair ..."). Ann1 to Ann4 refers to the four different human annotators. The Agreement column simply shows whether the pairs of annotators agreed or not.

The data set is accompanied by the annotation guidelines the authors used to annotate the examples.

1.2 Representation of the data set (flat files, database, markup)

The data set is provided as an Excel file with three sheets (the first sheet contains reference information, the second contains the data set itself and the third sheet contains the annotation guidelines.

1.3 Character encoding

The characters are UTF8 encoded.

2 ADMINISTRATIVE INFORMATION

2.1 Contact person

Name: Ralf Steinberger
Address: Joint Research Centre, ISPRA, Italy
Affiliation: Institute for the Protection and Security of the Citizen (IPSC)
Position: Head of Language Technology Group
Telephone: +39 - 0332 78 6271 + 5648
Fax: +39 - 0332 78 5154
e-mail: Ralf.Steinberger@jrc.ec.europa.eu

2.2   Delivery medium (if relevant; description of the content of each piece of medium)

The data set can be freely downloaded from the address below:
http://langtech.jrc.ec.europa.eu/JRC_Resources.html .

2.3   Copyright statement and information on IPR

The resource is free, with requested citation of the relevant papers (see below)

3   TECHNICAL INFORMATION

3.1 Directories and files

Not relevant

3.2 Data structure of an entry

The data set is represented in a Excel sheet, one quotation snippet per line, followed by the annotation information (see above section 1.1)

3.3 Corpora size (nmb. of tokens, MB occupied on disk)

1590 annotated snippets, 0.6 MB on disk

4   CONTENT INFORMATION

4.1 Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)

Monolingual annotated snippets

4.2 The natural language(s) of the corpus

English

4. 3 Domain(s)/register(s) of the corpus

Journalism

4.4 Annotations in the corpus (if an annotated corpus)

4.4.1 Types of annotations (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up)

Human labeling on positivity or negativity of the snippet (towards the target person or organization)
Agreement between annotators, labeled automatically

4.4.2 Tags (if POS/WSD/TIME/discourse/etc –tagged or parsed),
POS(itive), NEG(ative), empty (objective) and TRUE or FALSE (annotators agreement)

4.4.3Alignment information (if the corpus contains aligned documents: level of alignment, how it was achieved)
Not relevant

4.4.4 Attributes and their values (if annotated)
Not relevant

4.5 Intended application of the corpus
The purpose of the annotation was to produce a gold standard collection of news snippets to be used in training/evaluation opinion mining crawlers, sentiment classifiers.

4.6 Reliability of the annotations (automatically/manually assigned) – if any
Manual mark-up of two annotators (most of the time)

5 RELEVANT REFERENCES AND OTHER INFORMATION

Balahur-Dobrescu Alexandra & Ralf Steinberger (2009). Rethinking sentiment analysis in the news: from theory to practice and back. 'Workshop on Opinion Mining and Sentiment Analysis' (WOMSA), held at the 2009 CAEPIA-TTIA 13th Conference of the Spanish Association for Artificial Intelligence, pp. 1-12. Sevilla, Spain, 13.11.2009. Available from:
http://langtech.jrc.ec.europa.eu/Documents/09_WOMSA-WS-Sevilla_Sentiment-Def_printed.pdf

Balahur Alexandra, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik van der Goot, Matina Halkia, Bruno Pouliquen & Jenya Belyaeva (2010). *Sentiment Analysis in the News*. In: Proceedings of the 7th International Conference on Language Resources and Evaluation (**LREC'2010**), pp. 2216-2220. Valletta, Malta, 19-21 May 2010. Available from:
http://langtech.jrc.it/Documents/2010_03_LREC_Sentiment-analysis.pdf

Balahur Alexandra, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik van der Goot, Matina Halkia, Bruno Pouliquen & Jenya Belyaeva (2010). Sentiment Analysis in the News. In: Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'2010), pp. 2216-2220. Valletta, Malta, 19-21 May 2010. Available from:
http://langtech.jrc.ec.europa.eu/Documents/2010_03_LREC_Sentiment-analysis.pdf