==========

Documentation for MPQA Opinion Corpus version 1.2

==========

Contents:

---

1. Introduction

This corpus contains news articles manually annotated
using an annotation scheme for opinions and other private
states (i.e., beliefs, emotions, sentiments, speculations,
etc.). Version 1.0 of the corpus was collected and annotated
as part of the summer 2002 NRRC Workshop on Multi-Perspective
Question Answering (MPQA) (Wiebe et al., 2003) sponsored by ARDA.

---

2. Overview of Changes

2.1 Changes in terminology

Since the first release of the corpus, we have updated the terminology used in the annotation scheme. We believe that the new terminology is clearer. The two terminologies are equivalent, and the representations are homomorphic. See Wiebe, Wilson, Cardie (2005) for a complete description of the annotation scheme using the new terminology.

For this version of the corpus, we are releasing the annotations using both the old and new terminologies. Future releases will use only the new terminology.

The files containing the annotations in this release are named as follows:

gateman.mpqa.lre.1.2 - annotation files using new terminology
gateman.mpqa.sigdial.1.2 - annotation files using old terminology

For more information on these files, see section 6.

Section 3 gives a brief overview of the annotations using the new terminology. For an overview of the annotations using the old terminology, see the Database.1.0.README.

Below is a mapping between the old and new terminologies. An arrow (-->) indicates a change for that particular annotation type or attribute.

Old Terminology		New Terminology
expressive-subjectivity:		expressive-subjectivity:
strength	-->	intensity
on:	-->	objective-speech-event:
onlyfactive="yes"	-->	(attribute no longer used)
is-implicit=""	-->	implicit="true"
not-significant-particular	-->	insubstantial
onlyfactive-uncertain	-->	objective-uncertain
on-uncertain	-->	annotation-uncertain
on:	-->	direct-subjective:
onlyfactive="no"	-->	(attribute no longer used)
is-implicit=""	-->	implicit="true"
not-significant-particular	-->	insubstantial
overall-strength	-->	intensity
on-strength	-->	expression-intensity
onlyfactive-uncertain	-->	subjective-uncertain
on-uncertain	-->	annotation-uncertain

The most important change to note is that 'on' annotations with the attribute onlyfactive="yes" are now called 'objective-speech-event' annotations. The 'on' annotations with onlyfactive="no" are now called 'direct-subjective' annotations. 'Agent' annotations and 'inside' annotations remain unchanged.

2.2 Corrected sentence splits

Incorrect sentence splits produced by the sentence splitter have been manually corrected, along with any annotations that were affected. Corrections/refinements were also made to a few other annotations.

2.3 Addition of contextual polarity annotations

Contextual polarity annotations have been added to the subjective expressions marked in the corpus. Using the terminology of the annotation scheme described below, subjective expressions are all expressive subjective elements and all direct subjective annotations that have an expression intensity other than neutral.

The new contextual polarity annotations replace the attitude-type annotations on expressive subjective elements.

See the description of the annotation scheme below for more information about the contextual polarity annotations.

For more information about contextual polarity in general see (Wilson, Wiebe, Hoffmann, 2005).

---

3. Annotation Scheme

This section contains an overview of the types of annotations that you will see marked in the documents of this corpus. The complete annotation instruction used to produce the annotations are available at http://www.cs.pitt.edu/~wiebe/pubs/pub1.html.

In the descriptions below, note that some annotation attributes are marked as "Exploratory". These attributes were added later into the annotation scheme, and are not marked fully or in depth throughout the corpus.

3.1 agent annotation

Marks phrases that refer to sources of private states and speech events, or phrases that refer to agents who are targets of an attitude.

Possible attributes:

id - Unique identifier assigned by the annotator to the first meaningful and descriptive reference to an agent.

There are two agent annotations with a 0,0 byte span in every document. These two annotations are to give an id for the writer of the document ('w') and for an implicit agent ('implicit'). Private states and speech events are sometimes attributed to

implicit agents.

nested-source - Used when the agent reference is the source of a private state/speech event. The nested-source is a list of agent ids beginning with the writer and ending with the id for the immediate agent being referenced.

Example: w, Foreign Ministry, US State Dept

nested-target - (Exploratory) Used when an agent annotation is the target of a negative or positive attitude. The nested-target, like the nested-source, is a list of agent ids beginning with the writer and ending with the id for the agent being targeted.

agent-uncertain - Used when the annotator is uncertain whether the agent is the correct source of a private state/speech event

Possible values: somewhat-uncertain, very-uncertain

3.2 expressive-subjectivity annotation

Marks expressive-subjective elements, words and phrases that indirectly express a private state. For example, 'fraud' and 'daylight robbery' in the following sentence are expressive-subjective elements.

"We foresaw electoral fraud but not daylight robbery," Tsvangirai said.

Possible attributes:

nested-source - List of agent ids beginning with the writer and ending with the id for the immediate agent that is the source of the private state being expressed by the expressive-subjective element.

nested-source-uncertain - Used when an annotator is uncertain as to whether the agent is the correct nested source.

Possible values: somewhat-uncertain, very-uncertain

intensity - Indicates the intensity of private state being expressed by the expressive-subjective element.

Possible values: low, medium, high, extreme

polarity - Indicates the contextual polarity of the private state.

Possible values: positive, negative, both, neutral, uncertain-positive, uncertain-negative, uncertain-both, uncertain-neutral

3.3 direct-subjective annotation

Marks direct mentions of private states and speech events (spoken or written) expressing private states.

Possible attributes:

nested-source - List of agent ids, beginning with the writer and ending with the id for the immediate agent that is the source of the private state or speech event.

annotation-uncertain - Used when an annotator is uncertain as to whether the expression marked is indeed a direct private state or a speech event.

Possible values: somewhat-uncertain, very-uncertain

implicit - The presence of this attribute indicates that the speech event is implicit. This attribute is used when there is not a private state or speech event phrase on which to actually make an annotation. For example, there is no phrase "I write" for the writer of the sentence.

subjective-uncertain - Used when an annotator is uncertain as to whether a private state is being expressed.

Possible values: somewhat-uncertain, very-uncertain

intensity - Indicates the overall intensity of the private state being expressed, considering the 'direct-subjective' phrase and everything inside its scope.

Possible values: low, medium, high, extreme

expression-intensity - Indicates the intensity of the speech event or private state expression itself.

Possible values: neutral, low, medium, high, extreme

polarity - Indicates the contextual polarity of the private state. Only included when expression-intensity is not neutral.

Possible values: positive, negative, both, neutral, uncertain-positive, uncertain-negative, uncertain-both, uncertain-neutral

attitude-type - (Exploratory) Indicates the type of attitude being expressed.

Possible values: negative, positive, both

attitude-toward - (Exploratory) Agent id of who the attitude-type is directed toward.

insubstantial - Used when the private state or speech event is not substantial in the discourse

Possible values are combination of: c1, c2, c3

These possible values correspond to criteria necessary for a private state or speech event to be substantial. Please see the annotation instructions for a complete description of these criteria. The criteria listed for this attribute are the criteria that the private state or speech speech event fails to meet.

3.4 objective-speech-event annotation

Marks speech events that do not express private states.

Possible attributes:

nested-source - List of agent ids, beginning with the writer and ending with the id for the immediate agent that is the source of the private state or speech event.

annotation-uncertain - Used when an annotator is uncertain as to whether the expression marked is indeed a speech event.

Possible values: somewhat-uncertain, very-uncertain

implicit - The presence of this attribute indicates that the speech event is implicit. This attribute is used when there is not a speech event phrase on which to actually make an annotation. For example, there is no phrase "I write" for the writer of the sentence.

objective-uncertain - Used when an annotator is uncertain as to whether the speech event is objective.

Possible values: somewhat-uncertain, very-uncertain

insubstantial - Used when the speech event is not substantial in the discourse

Possible values are combination of: c1, c2, c3

These possible values correspond to criteria necessary for a private state or speech event to be substantial. Please see the annotation instructions for a complete description of these criteria. The criteria listed for this attribute are the criteria that the private state or speech event fails to meet.

3.5 inside annotation

The term 'inside' refers to the words inside the scope of a direct private state or speech event phrase ('on'). The annotators did not mark 'inside' annotations. However, 'inside' annotations were created automatically for each writer 'on' annotation. Each writer 'inside' corresponds to a GATE sentence.

3.6 anchoring implicit sources

When the source is implicit, this means there is no speech or private state phrase to anchor the objective speech or direct subjective annotation to. When the objective speech or direct subjective is implicit and the source is the writer, we opted to anchor the annotation to the beginning of the sentence, typically a 0-byte or 1-byte span. Occasionally, there are other implicit objective speech or direct subjective annotations with a source that is not the writer. These annotations are typically anchored at the beginning of the text that is attributed to the source. Note that all implicit objective speech and direct subjective annotators are marked with the attribute "implicit=true".

---

4. Subjective Sentences

The annotations described in section 3 are expression-level annotations, performed below the level of the sentence. We ask annotators to identify all subjective expressions in a sentence, which gives us very fine-grained, detailed annotations. Although the annotators sometimes differ over which particular expressions are subjective, and how many subjective expressions are in a sentence, they have very good agreement as to whether there is subjectivity in a sentence (see (Wiebe, Wilson, Cardie (2005)).

For the work using this data that appeared in CoNLL03 (Riloff et al., 2003) and EMNLP03 (Riloff & Wiebe, 2003) the following definition of a subjective sentence was used. The definition is in terms of the annotations.

A sentence was considered subjective if 1 OR 2:

1. the sentence contains a "GATE_direct-subjective" annotation WITH attribute intensity NOT IN ['low', 'neutral'] AND NOT WITH attribute insubstantial.

2. the sentence contains a "GATE_expressive-subjectivity" annotation WITH attribute intensity NOT IN ['low']

Otherwise, a sentence was considered objective.

The file, test_setCoNLL03, contains the list of files used for evaluation in (Riloff et al, 2003).

NOTE: Since the experiments performed in (Riloff et al., 2003) and (Riloff & Wiebe, 2003), some annotation errors and errors in the sentence splits have been corrected.

---

5. Data

The corpus contains 535 documents, a total of 11,114 sentences.

The articles in the corpus are from 187 different foreign and U.S. news sources. They date from June 2001 to May 2002.

The articles were identified by human searches and by an information retrieval system. The majority of the articles are on 10 different topics, but a number of additional articles were randomly selected (more or less) from a larger corpus of 270,000 documents. This last set of articles has topic: misc.

The 10 topics are:

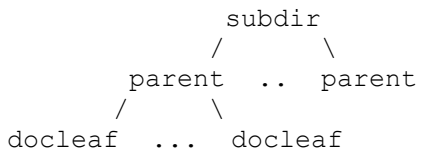
- argentina: economic collapse in Argentina
- axisofevil: reaction to President Bush's 2002 State of the Union Address
- guantanamo: U.S. holding prisoners in Guantanamo Bay
- humanrights: reaction to U.S. State Department report on human rights
- kyoto: ratification of Kyoto Protocol
- mugabe: 2002 presidental election in Zimbabwe
- settlements: Israeli settlements in Gaza and West Bank
- spacestation: space missions of various countries
- taiwan: relations between Taiwan and China
- venezuela: presidential coup in Venezuela

The file, Release-Full-Doclist, lists the documents in the full release of the MPQA corpus. The file, Release-Full-ByTopic, lists the documents along with their topics.

---

6. Database Structure

The database/ contains three subdirectories: docs, meta_anns, man_anns. Each subdirectory has the following structure:



Within each subdirectory, each document is uniquely identified by its parent/docleaf. For example, 20010927/23.18.15-25073, identifies one document. 20010927 is the parent; 23.18.15-25073 is the docleaf.

6.1 database/docs

The docs subdirectory contains the document collection. In this subdirectory, each docleaf (e.g., 23.18.15-25073) is a text file containing one document.

6.2 database/meta_anns

Each docleaf (e.g., 23.18.15-25073) in the meta_anns subdirectory contains information about the document (e.g., source, date). The meta_anns files are in MPQA format, which is described in section 7.

6.3 database/man_anns

This subdirectory contains the manual annotations for the documents. In this subdirectory, each docleaf (23.18.15-25073) is a directory that contains three files: gateman.mpqa.lre.1.2, gateman.mpqa.sigdial.1.2, and gatesentences.mpqa.1.2.

The file gateman.mpqa.lre.1.2 contains the human opinion annotations using the new terminology (Wiebe, Wilson, Cardie (2005)). The file gateman.mpqa.sigdial.1.2 contains the annotations using the old terminology (Wilson and Wiebe (2003)). The file gatesentences.mpqa.1.2 contains spans for sentence, minus junk sentences that contain meta data or other spurious information that was not part of the article. These junk sentences were removed by hand.

All three files, gateman.mpqa.lre.1.2, gateman.mpqa.sigdial.1.2 and gatesentences.mpqa.1.2 are in MPQA format, described in section 7.

---

7. MPQA Annotation Format

The MPQA format is a type of general stand-off annotation. Every line in an annotation file is either a comment line (beginning with a '#") or an annotation line (one annotation per line).

An MPQA annotation line consists of text fields separated by a single TAB. The fields used are listed below, with an example annotation underneath.

```
id      span    data_type       ann_type        attributes
58      730,740 string          GATE_agent       nested-source="w,chinarep"
```

Every annotation has a identifier, id. This id is unique ONLY within a given MPQA annotation file.

The span is the starting and ending byte of the annotation in the document. For example, the annotation listed above is from the document, temp_fbis/20.20.10-3414. The span of this annotation

is 730,740. This means that the start of this annotation is
byte 730 in the file docs/temp_fbis/20.20.10-3414, and byte 740
is the character after the last character of the annotation.

```
     blah, blah, blah, example annotation, blah, blah, blah
                       |                 |
                  start byte          end byte
```

The data_type of all annotations should be 'string'.

The types of annotations in the gateman.mpqa files are
GATE_agent, GATE_expressive-subjectivity, GATE_direct-subjective,
GATE_objective-speech-event, GATE_inside, and GATE_split.
With the exception of GATE_split, these annotation types
correspond to the annotation types described in section 3.

Sentence annotations in the gatesentence.mpqa.1.2 files have
type GATE_sentence.

Each attribute is an attribute_name="attribute_value" pair.
An annotation may have any number of attributes, including
0 attributes. Multiple attributes for an annotation are
separated by single spaces, and they may be listed in any
order. The attributes that an annotation may have depends
on the type of annotation. The set of possible attributes
for each annotation type is listed in section 3.

---

8. Acknowledgements

The development of the MPQA Opinion Corpus version 1.0
was performed in support of the Northeast Regional Reseach
Center (NRRC) which is sponsored by the Advanced Research
and Development Activity (ARDA), a U.S. Government entity
which sponsors and promotes research of import to the
Intelligence Community which includes but is not limited
to the CIA, DIA, NSA, NIMA, and NRO.

The development of version 1.2 was supported in part by the
NSF under grant IIS-0208798 and by the Advanced Research
and Development Activity (ARDA).

---

9. Contact Information

Please direct any questions that you have about this corpus or
the annotation scheme to Theresa Wilson at the University of
Pittsburgh.

Theresa Wilson email: twilson@cs.pitt.edu

10. References

Janyce Wiebe, Eric Breck, Chris Buckley, Claire Cardie, Paul Davis, Bruce Fraser, Diane Litman, David Pierce, Ellen Riloff, Theresa Wilson, David Day, Mark Maybury (2003). REcognizing and Organizing Opinions Expressed in the World Press. 2003 AAAI Spring Symposium on New Directions in Question Answering.

Theresa Wilson and Janyce Wiebe (2003). Annotating Opinions in the World Press. 4th SIGdial Workshop on Discourse and Dialogue (SIG0dial-03). ACL SIGdial.

Ellen Riloff, Janyce Wiebe, and Theresa Wilson (2003). Learning Subjective Nouns Using Extraction Pattern Bootstrapping. Seventh Conference on Natural Language Learning (CoNLL-03). ACL SIGNLL.

Ellen Riloff and Janyce Wiebe (2003). Learning Extraction Patterns for Subjective Expressions. Conference on Empirical Methods in Natural Language Processing (EMNLP-03). ACL SIGDAT.

Janyce Wiebe, Theresa Wilson, and Claire Cardie (2005). Annotating expressions of opinions and emotions in language. Language Resources and Evaluation (formerly Computers and the Humanities) 1(2).

Theresa Wilson, Janyce Wiebe, and Paul Hoffman (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. Proceedings of HLT/EMNLP 2005, Vancouver, Canada.

Theresa Wilson
Janyce Wiebe

version 1.2
last modified 12/06/05