Dr. Inventor Multi-layered Annotated Corpus
- Version 1.0 -

Delivery format of the Gold Standard Annotations of scientific discourse, subjective statements, citation purpose, sentence relevance and hand written summaries

The Gold Standard Annotations of the 40 documents of Dr. Inventor Corpus are represented by relying on the set of XML markup elements described hereafter.
The annotated corpus for each document includes four layers / main types of annotations:

- **SCIENTIFIC DISCOURSE**: the scientific discourse category (APPROACH, BACKGROUND, CHALLENGE, etc.) associated to each sentence;
- **CITATION PURPOSE**: the citation purpose (CRITICISM, NEUTRAL, COMPARISON, etc.) of the sentences (or part of sentences) that belong to the context of each in-line citation of each document;
- **SUBJECTIVE STATEMENTS**: the aspect (NOVELTY, COMMON_PRACTICE, ADVANTAGE, etc.) of the sentences of each document;
- **SUMMARY**: the relevance with respect to the creation of a summary of each sentence of each document, rated with 5 grades, from 1.0 (totally irrelevant for a summary) to 5.0 (very relevant for a summary).

In addition to these annotations the corpus includes three hand-written summaries (max 250 words) of each document.

The annotations of the papers of the Corpus are grouped by folders. Each folder includes the annotation layers of a specific document (A01, A02, ..., A40).

The Gold Standard Annotations are represented thanks to proper XML markup elements. For each paper of the Corpus, there is a different XML document to represent the Gold Standard annotations of a specific layer (SCIENTIFIC DISCOURSE, CITATION PURPOSE, SUBJECTIVE STATEMENTS, SUMMARY).
Every XML version of a paper of the Corpus, independently from the type of Gold Standard Annotations it represents, includes the following markup elements:

- `DOCUMENT`: root XML element, with the attribute 'name' specifying the name of the corpus document;
- `TITLE`: the title of the paper;
- `ABSTRACT`: the abstract of the paper;
- `H1, H2, H3, H4`: useful to identify the title of the sections of the document at different nesting levels.

SCIENTIFIC DISCOURSE:
The Gold Standard annotations of sentences with subjective statements are represented by an XML file associated to each document of the Corpus with name ending in '_RHETORICAL_v1.xml'.
In each XML file, each sentence has been characterized by a scientific discourse category and is spotted by means of the XML element *Sentence* that has the attribute **rhetoricalClass** with value equal to the Gold Standard annotation.

A example of *Sentence* annotations is represented by the following XML excerpt:

```
<Sentence rhetoricalClass="DRI_Challenge">
While realism is important, other factors such as intuitiveness, ease of
interaction and computational cost are also of great importance in animation
production.</Sentence>
<Sentence rhetoricalClass="DRI_Background">A simulation method lets the animator
easily create certain effect which could be otherwise almost an impossible
mission with manual manipulations.</Sentence>
```

The first *Sentence* XML element has the gold standard value '`DRI_Challenge`'.

SUBJECTIVE STATEMENTS:
The Gold Standard annotations of sentences with subjective statements are represented by an XML file associated to each document of the Corpus with name ending in '_ASPECT_v1.xml'.
In each XML file, each sentence that has been characterized by an aspect is spotted by means of the XML element *Sentence* that has the attribute **aspectClass** with value equal to the Gold Standard annotation.

A example of *Sentence* annotations is represented by the following XML excerpt:

```
<Sentence aspectClass="COMMON_PRACTICE">
However, although PSD may be used as a compensation to the underlying SSD, and
the animator specifies the PSD examples after the SSD has been performed, it is
generally believed that the examples are best interpolated in the rest pose,
before the SSD has been applied.</Sentence>
<Sentence aspectClass="COMMON_PRACTICE">Therefore the action of the SSD and any
other deformations must be ?inverted? in order to push the example compensation
before these operations.</Sentence>
```

The first *Sentence* XML element has the gold standard value '`COMMON_PRACTICE`'.

CITATION:
The Gold Standard Annotations of citation purposes are represented by an XML files associated to each document of the Corpus with name ending in '_CITATION_PURPOSE_M_v1.xml'.
In each citation-related XML document, each sentence of part of sentence that belongs to a citation context is surrounded by the XML element *Cit_context*. Each element of this type has the attribute **CITid_XX** that has as value the Gold Standard annotation. The XX part of the attribute name constitutes the identifier of the in-line citation the context belongs to.
In-line citations are spotted by means of the XML element *InlineCitation*. This element has the attribute **CITid** that has the value equal to its in-line citation identifier.
An example of *Cit_context* and *InlineCitation* annotations is represented by the following XML excerpt:

```
<Cit_context CITid_3="CRITICISM">
Related research efforts have improved the speed and power of example-based
skinning.</Cit_context>
<InlineCitation CITid="3">[ 2 ]</InlineCitation>
<Cit_context CITid_3="NEUTRAL"> incorporate linear elements into RBF to produce
constant changes between examples.</Cit_context>
```

We can notice that the first *Cit_context* element identifies a citation context sentences that has '`CRITICISM`' as Gold Standard value and **3** as in-line citation identifier. It references the in-line citation identified by the *InlineCitation* element with attribute `CITid="3"`.

SUMMARY:
The Gold Standard annotations representing the relevance of each sentence with respect to the document summary are spotted by means of an XML file associated to each document of the Corpus with name ending in '_SUMMARY_v1.xml'. In each XML file, each sentence has an associated summary relevance score and is identified by means of the XML element *Sentence* that has the attribure **summaryRelevanceScore** with value equal to the Gold Standard annotation.

The summary relevance score of a sentence can be determined by means of 5 grades, from 1.0 (totally irrelevant for a summary) to 5.0 (very relevant for a summary). The Gold Standard value of the summary relevance score is the average of the summary relevance scores of each annotator (rounded to the second decimal digit).

A example of *Sentence* annotation is represented by the following XML excerpt:

```
<Sentence summaryRelevanceScore="3.33">
Therefore the action of the SSD and any other deformations must be ?inverted? in
order to push the example compensation before these operations.</Sentence>
< Sentence summaryRelevanceScore="5.00">
Our goal is to incorporate examplebased skinning into a system having a variety
of other skinning and deformation operations, and to be able to invert these
operations regardless of their nature.</Sentence>
```

The first *Sentence* XML element has a gold standard summary relevance value of *3.33*.

HAND-WRITTEN DOCUMENT SUMMARIES
For each document of the Corpus, three hand written summaries are available. Each annotator has been asked to summarize with his own words the contents of the document, thus generating a summary in maximum 250 words.
Each hand written summary is included in a text file with name DOCID_HandWrittenSummary_ANNID.txt where DOCID is the identifier of the document of the corpus that is summarized (A01, A02, …, A40) while ANNID is the identifier of the annotation that wrote the summary. For instance the summary contained in a file with name *A04_S05__HandWrittenSummary_17.txt* summarizes the document A04 and has been written by the annotator with id 17.