2010 i2b2 / VA Challenge Evaluation

Annotation File Formatting

Ground truth annotations and participating team system output will be contained in three files per report, one of the files will contain the information necessary for concepts, the second for assertions, and the third for relations. Output files should be named the same as the report text with a .con extension for concepts, .ast for assertions, and .rel extension for relations.

Report file:	123.txt
Annotation files:	123.con
	123.ast
	123.rel

Participants competing in the concept task will read in the report files and output corresponding concept annotation files.

Participants competing in the assertion task will read in the report files and concept annotation files and output corresponding assertion annotation files.

Participants competing in the relation task will read in the report files, concept annotation files, and assertion annotation files and output corresponding relation annotation files.

Report Files

Reports will be released as plain text files that have been formatted to ensure consistent interpretation of annotation offsets (the line and word numbers that span the annotated text). Reports may be named according to some pseudo-identifier, the type of document, and/or incrementally and as such the report names may contain alphanumeric characters, underscores, and dashes.

Reports are formatted to have one sentence per line. Blank lines have been removed and punctuation has been automatically split from text. Thus a line is defined as the text delimited by newline characters and a word is defined as the text in a line delimited by space characters.

Concept Annotation Files

Each concept mention will be output on a separate line in a concept annotation file formatted as follows:

```
c=”concept text” offset||t=”concept type”
```

where

c represents a mention of a concept. *concept text* is replaced with the actual text from the report.

offset represents the beginning and end line and word numbers that span the concept text. An offset is formatted as the line number followed by a colon followed by the word number. The starting and ending offset are separated by a space. The first line of a report starts as line number 1. The first word in a line is counted as word number 0.

t represents the semantic type of concept mentioned. *concept type* is replaced with **problem**, **treatment**, or **test**.

Examples:

- `c=”prostate cancer” 5:7 5:8||t=”problem”`
- `c=”chemotherapy” 5:4 5:4||t=”treatment”`
- `c=”chest x-ray” 6:12 6:13||t=”test”`

Assertion Annotation Files

Only concepts that are medical problems will have assertions. Each assertion mention will be output on a separate line in an assertion annotation file formatted identical to the concept annotation file with the addition of an assertion value as follows:

```
c=”concept text” offset||t=”concept type”||a=”assertion value”
```

where

c, *offset*, and *t* are defined as above.

a represents the assertion of the concept mentioned. *assertion value* is replaced with **present**, **absent**, **possible**, **conditional**, **hypothetical**, or **associated with someone else**.

Examples:

- `c=”prostate cancer” 5:7 5:8||t=”problem”||a=”present”`
- `c=”diabetes” 2:14 2:14||t=”problem”||a=”absent”`
- `c=”pain” 7:3 7:3||t=”problem”||a=”conditional”`

Relation Annotation Files

Each relation will be output on a separate line formatted as follows:

```
c=”concept text” offset||r=”relation type”||c=”concept text” offset
```

where

c and *offset* are defined as above.

r represents the type of relation the two concepts share. *relation type* is replaced with one of the following values:

- **TrIP** for treatment improves medical problem relations
- **TrWP** for treatment worsens medical problem relations.
- **TrCP** for treatment causes medical problem relations.
- **TrAP** for treatment is administered for medical problem relations.
- **TrNAP** for treatment is not administered because of medical problem relations.
- **PIP** for medical problem indicates medical problem relations.
- **TeRP** for test reveals medical problem relations.
- **TeCP** for test conducted to investigate medical problem relations.

The second *c* and *offset* represent the other concept in the relation and are defined as above.

Examples:

- `c=”discomfort” 3:4 3:4||r=”PIP”||c=”acute MI” 3:8 3:9`
- `c=”chemotherapy” 5:4 5:4||r=”TrAP”||c=”prostate cancer” 5:7 5:8`
- `c=”chest x-ray” 6:12 6:13||r=”TeRP”||c=”pneumonia” 6:15 6:15`

The order of the concepts is not important, meaning that the following lines are considered equivalent by the evaluation program.

- `c=”discomfort” 3:4 3:4||r=”PIP”||c=”acute MI” 3:8 3:9`
- `c=”acute MI” 3:8 3:9||r=”PIP”||c=”discomfort” 3:4 3:4`