2011 i2b2/VA Challenge Output formats

Output files for the 2011 i2b2/VA challenge will follow similar formatting from previous i2b2 challenges. The 2011 i2b2/VA challenge is on co-reference resolution; therefore, the output files capture co-reference chains (.chains). The input for the systems is patient reports (.txt) and the markables (.con) that should be studied for co-reference. Pairs (.pair) are provided for reference and development. Evaluation will be on chains (.chain). Specifications for these file formats are as follows:

1) Ground truth markables (.con)

The .con files hold all of the markables, i.e., concepts, that should be studied for co-reference in a single document from the corpus. Concepts in this context fit within the following classes:

- Treatment
- Problem
- Test
- Person
- Pronoun

Ground truth markables from the 2010 i2b2/VA challenge corpus
Ground truth markables representing medical problems (aka problems), treatments and tests are included in the .con file. These are retained from the ground truth annotations released from the 2010 i2b2/VA challenge corpus.

"Person" and "pronoun" markables
Person markables include mentions of person names that could have been redacted by de-identification systems, or realistic surrogates added as part of the de-identification resynthesis process.

The .con files contain markables, one markable per line, and adhere to the following format:

```
c= "<Markable>" <StartLineOffset>:<StartWordOffset>
<EndLineOffset>:<EndWordOffset>||t="<Class>"
```

For the markable "`Diabetes`" located on the second line and as the third word:
`c="Diabetes" 2:3 2:3||t="problem"`

For the markable "`Oral glucose tolerance test`" located on the fifth line and spanning fourth through seventh tokens:
`c="Oral glucose tolerance test" 5:4 5:7||t="test"`

For the pronoun markable "`he`" located on the fifth line and as the fourth word:
`c="he" 5:4 5:4||t="pronoun"`

For the person markable "`the patient`" located on the eighth line and spanning the third and fourth tokens:
`c="The Patient" 8:3 8:4||t="person"`

For the treatment markable located on the 59th line and spanning the sixth and seventh tokens:
`c="dvt prophylaxis" 59:6 59:7||t="treatment"`

2) Files containing co-reference pairs (.pairs)

Each .txt file will have a corresponding *.pairs* file that contains all the co-reference pairs identified in the document. Each entry in the *.pairs* files represents a pair of co-referring markables, using the following format:

`c="concept1" offset||t=coreference_type||c="concept2" offset`

Where the first concept has the smaller offset span start (appears earlier in the text).
The type of co-reference could be:

- "`coref person`" for person co-references
- "`coref problem`" for problem concept co-references
- "`coref treatment`" for treatment concept co-references
- "`coref test`" for test concept co-references

Here are a few examples:

```
c="her" 26:18 26:18||t="coref person"||c="the patient" 27:7 27:8

c="pneumonia" 23:22 23:22||t="coref problem"||c="her right lower lobe
pneumonia" 24:9 24:13

c="the procedure" 54:3 54:4||t="coref treatment"||c="surgery" 87:14
87:14

c="his" 69:0 69:0||t="coref person"||c="he" 71:3 71:3
```

Note that there is no "coref pronoun" type. If a pair involves 2 pronoun concepts, the type of the co-reference for the pair should be the type of the associated class (person, treatment, problem, or test). Here are 2 examples:

Assuming that "`he`" and "`his`" below refer to "`the patient`", we could have:
`c="he" 63:0 63:0||t="coref person"||c="his" 64:0 64:0`

Assuming that two instances of "`they`" refer to "`pain medications`", we could have:
`c="they" 51:4 51:4||t="coref treatment"||c="they" 52:0 52:0`

3) Files containing co-reference chains (.chains)

Each .txt file should have a corresponding *.chains* file that contains all the co-reference chains identified in the document. Each entry in the .chains files represents a chain of co-referring markables, in ascending order of span offset, terminated by the type of co-reference (t="coref type").

A chain of N markables will have the following format:
`c=<concept1> offset ||c=<concept2> offset|| … ||c=<conceptN> offset ||t=<coreference_type>`
where concept1 start offset < concept2 start offset < … < conceptN start offset

Note that co-reference chains should include at least one markable that is NOT a pronoun to determine their type.

Here are a few examples:

```
c="dr. J" 87:6 87:7||c="dr. J" 92:4 92:5||c="T J , m.d." 95:2
95:6||c="T J, m.d." 95:10 95:14||c="T J , m.d." 103:0 103:4||t="coref
person"

c="right hip osteoarthritis" 21:0 21:2||c="advanced osteoarthritis of
his right hip" 49:3 49:8||c="severe osteoarthritis of the right hip"
51:6 51:11||t="coref problem"

c="coumadin" 91:4 91:4||c="anticoagulation" 91:6 91:6||t="coref
treatment”

c="the patient" 22:0 22:1||c="she" 23:0 23:0||c="she" 24:0
24:0||c="her" 26:0 26:0||c="she" 27:0 27:0||c="she" 29:0 29:0||t="coref
person"

c="this 34 year old female" 37:0 37:7||c="this" 37:0 37:0||c="she" 39:0
39:0|| t="coref person"

c="pain medications" 51:4 51:4||c="they" 52:0 52:0|| c="they" 53:0
53:0||t="coref treatment"
```