Annotation Guideline

A Cross-Document Coreference Resolution Approach to Low-Resource Languages

Nathanon Theptakob, Thititorn Seneewong Na Ayutthaya, Chanatip Saetia,
Tawunrat Chalothorn, and Pakpoom Buabthong

Objective

Annotate all entities and connect all coreference entities, both within-document and cross-document together.

Annotation instruction

1. In INCEpTION tool, highlight the entity. In this guideline, we only focus on entity, not event. For example,

   Somchai is a pro driver who joined a car rally. Over 100 people join this event.

   Only annotate the blue part (entity), not the red part (event)
2. After annotating all entities, drag one entity onto another to connect 2 entities as a coreference. In this part, pronouns are also coreference to the corresponding nouns. For example

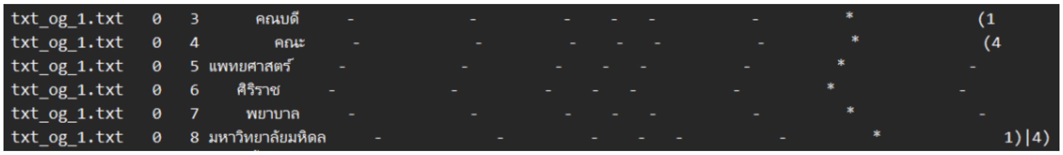
   Somchai is a pro driver who joined a car race. He must compete with 100 contestants.

   In this sentence, "Somchai", "a pro driver" are in coreference. "He", must be included into the cluster since the word refers to Somchai. For more example regarding the configuration of the annotation tools, visit INCEpTION - Projects (tu-darmstadt.de)
3. After annotation and coreference connection, export the results in CoNLL 2012 format and record in a separate file.
4. For cross-document coreference entities, record their connections in the following with the following specific data format
   1. For entities that does not exist in the record, write the entity's name in the first column "Entity". You do not need to specify the full name.
   2. The next columns will be document names. For each document name column, the data will be coded as

      <Entity code> D <Document number>

In which entity code and document number can be found in the exported .conll file from INCEpTION tool. For example

c.



```
txt_og_1.txt   0   3      คณบดี     -            -         -    -   -          -         *            (1
txt_og_1.txt   0   4         คณะ     -            -         -    -   -          -         *            (4
txt_og_1.txt   0   5 แพทยศาสตร์     -            -         -    -   -          -         *            -
txt_og_1.txt   0   6     ศิริราช     -            -         -    -   -          -         *            -
txt_og_1.txt   0   7      พยาบาล     -            -         -    -   -          -         *            -
txt_og_1.txt   0   8 มหาวิทยาลัยมหิดล     -            -         -    -   -          -         *         1)|4)
```

“คณบดีคณะแพทยศาสตร์ศิริราชพยาบาล มหาวิทยาลัยมหิดล” will be coded as 1D1