Ruler 1.2 Manual

July 10, 2015

Mateusz Kopeć

Institute of Computer Science, Polish Academy of Sciences
m.kopec@ipipan.waw.pl

About

The current version of the program facilitates the automatic clustering of mentions into coreferent clusters using a rule-based method.

Homepage: http://zil.ipipan.waw.pl/Ruler
Contact person: Mateusz Kopeć [mateusz.kopec@ipipan.waw.pl]
Author: Mateusz Kopeć

1 Requirements

Java Runtime Environment (JRE) 1.7 or newer.

2 Input data format

Input texts must be in TEI format used in the National Corpus of Polish (TEI NKJP, see [1] or [2] for reference). That means they must contain at least the following layers:

- `text_structure.xml` – containing the text structure,
- `ann_segmentation.xml` – with segmentation,
- `ann_morphosyntax.xml` – with morphosyntactic information,
- `ann_mentions.xml` – with mentions to cluster (this layer is not in National Corpus of Polish, see it's description below).

Additional layers may or may not be present:

- `ann_groups.xml` – with syntactic groups,
- `ann_words.xml` – with syntactic words,
- `ann_named.xml` – with named entites.

All files can be gzipped if necessary.

2.1 Format of ann_mentions.xml

This file contains mentions (represented by `<seg>` tags), which are simple a set of pointers to morphosyntax layer segments. Structure of the text is also kept, mentions are grouped into sentences and paragraphs, corresponding to ones in morphosyntax.

In the example figure 1, each mention is preceded with a comment with its orthographical form, however it's not obligatory. All `<ptr>` elements target tokens, which form the mention. Feature `<f>` with name `semh` shows, which token of the mention is it's semantic head.

Zero subjects are distinguished from other mentions by having an additional feature `<f name="zero" fVal="true" />`.

3 Output data format

Ruler builds on TEI NKJP format, adding a new layer:

- `ann_coreference.xml`

This layer stores the information about groups of mentions. Each group is supposed to contain only mentions referring to the same entity, i.e. they should be coreferent.

3.1 Format of ann_coreference.xml

This file stores information about coreference clusters. Each cluster is represented by `<seg>` tag and contains pointers to it's elements – mentions, referring to `ann_mentions.xml` file. The comment with orthographical forms of cluster elements before each `<seg>` tag is not obligatory. Value `ident` in `type` of coreference means identity (currently it's the only type **Ruler** produces). Value of `dominant` feature is the orthographical form of mention decided to be a best representative of a cluster.

This file doesn't contain paragraphs and sentences, because clusters can span across them. The only `<p>` tag is artificial, to fit the requirements of the TEI format. Example file is presented in figure 2.

4 Usage

Standalone jar doesn't need any installation. To run it, simply execute:

```
java -jar ruler-1.1-SNAPSHOT.one-jar.jar <dir with input texts> <dir
for output texts>
```

All texts recursively found in `<dir with input texts>` are going to be annotated with coreference layer and saved in `<dir for output texts>`.

References

1. Piotr Bański and Adam Przepiórkowski. The TEI and the NCP: the model and its application. In *LREC 2010 Workshop on Language Resources: From Storyboard to Sustainability and LR Lifecycle Management*, Valletta, Malta, 2010. ELRA.
2. Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, editors. *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warsaw, 2011.

```
<teiCorpus xmlns="http://www.tei-c.org/ns/1.0">
<TEI>
  <text>
    <body>
      <p xml:id="mentions_p-1" corresp="morph_1-p">
        <s xml:id="mentions_p-1.1-s" corresp="morph_1.1-s">
          <!-- Europejskiego Króla Kurkowego  -->
          <seg xml:id="mention_6">
            <fs type="mention">
              <f name="semh" fVal="ann_morphosyntax.xml#morph_1.1.24-seg"/>
            </fs>
            <ptr target="ann_morphosyntax.xml#morph_1.1.23-seg"/>
            <ptr target="ann_morphosyntax.xml#morph_1.1.24-seg"/>
            <ptr target="ann_morphosyntax.xml#morph_1.1.25-seg"/>
          </seg>
          ...
        </s>
        <s xml:id="mentions_p-1.2-s" corresp="morph_1.2-s">
          <!-- był -->
          <seg xml:id="mention_11">
            <fs type="mention">
                <f name="semh" fVal="ann_morphosyntax.xml#morph_1.1.4-seg"/>
                <f name="zero" fVal="true" />
            </fs>
            <ptr target="ann_morphosyntax.xml#morph_1.1.4-seg"/>
          </seg>
          ...
        </s>
      </p>
      <p xml:id="mentions_p-2" corresp="morph_2-p">
      ...
      </p>
      ...
    </body>
  </text>
</TEI>
</teiCorpus>
```

Fig. 1. Example `ann_mentions.xml` file

```
<?xml version="1.0" ?>
<teiCorpus xmlns="http://www.tei-c.org/ns/1.0">
  <TEI>
    <text>
      <body>
        <p>
          <!--  udział; udział; udziale -->
          <seg xml:id="coreference_0">
            <fs type="coreference">
              <f name="type" fVal="ident"/>
              <f name="dominant" fVal="udział"/>
            </fs>
            <ptr target="mention_1"/>
            <ptr target="mention_8"/>
            <ptr target="mention_21"/>
          </seg>
          ...
          <seg ...
          </seg>
        </p>
      </body>
    </text>
  </TEI>
</teiCorpus>
```

Fig. 2. Example `ann_coreference.xml` file