UNSC-NE: A Named Entity Extension to the UN Security Council Debates Corpus

Luis Glaser, Ronny Patz, Manfred Stede

August 31, 2021

1 Description

This dataset contains Named Entity (NE) annotations created via Named Entity Linking (NEL) for the base UN Security Council Debates corpus [1] The base corpus was transformed into a graph presentation and then annotated using DBpedia-spotlight. These annotations were subsequently linked to the wikidata knowledge graph. It is recommended to use the add-on in combination with the base corpus

1.1 Descriptives

Type	**Count**
speeches	82,165
sentences	1,921,352
entities	2,377,371
wikidata classes	40,916
distinct entity classes	4,907
distinct entity superclasses	10,989

Table 1: Descriptives for UNSC-NE

2 Citation

When using our dataset, please use the citation indication on the Dataverse's website.

Glaser, Luis; Ronny, Patz; Manfred, Stede, 2021, "UNSC-NE: A Named Entity Extension to the UN Security Council Debates Corpus", https://doi.org/10.7910/DVN/OV1FLX, Harvard Dataverse, V1

3 Files

unsc-ne.jsonl Contains all data in the corpus add-on in jsonlines (.jsonl) format. The format contains a valid json value on each line. See fig. 1 for a deeper explanation and an example.

unsc-ne.dump We also distribute the corpus as a simple neo4j dump, that can be loaded into a neo4j graph database using the admin tool:

```
neo4j-admin load --from=unsc-ne.dump --database=<NAME>
```

4 `.jsonl` format

Each json object either represents a node or a relationship between two nodes. Nodes are identified with an id, have one or multiple labels and may have properties in form of a dictionary. Relationships are identified with their own id and the ids of the two nodes that are connected. Relationships may also contain properties in form of a dictionary.

```
{ // Node 1
  "type":"node",
  "id": "1",
  "labels": ["Sentence"]
}
{ // Node 2
  "type":"node",
  "id":"2",
  "labels":["DBConcept"],
  "properties": {
    "uri":"dbpedia.org/Wartime_sexual_violence"
  }
}
{ // Relationship between Node 1 and Node 2
  "id":"4",
  "type":"relationship",
  "label":"CONTAINS",
  "start":{"id":"1", "labels":["Sentence"]},
  "end":{"id":"2", "labels":["DBConcept"]}
}
```

Figure 1: Data sample from UNSC-NE corpus addon (newlines between json objects do not exist in the corpus, they are inserted hear for readability.

5 Graph Structure

5.1 Nodes

The following list shows the different node types the UN Security Council debates NEL add-on contains. We also provide a small explanation of each property that a node has. The two node types *Meta* and *Speaker* can be used as links to the foundational corpus.

- AgendaItem
  - name: the name of the agenda item
- Country
  - name: the name of the country
- DBConcept
  - uri: the DBpedia uri this node represents
- Institution
  - name: the name of the institution
- Meta *Represents an entry in* `meta.tsv` *of the fundamental UN Security Council debates corpus*
- Paragraph
  - index: the index within the speech it's contained in
- Sentence
  - index_in_speech: the index within the speech it's contained in
  - index: the index within the paragraph it's contained in
  - text: the text of the sentence itself
- Speaker
  Represents an entry in `speaker.tsv` *of the fundamental UN Security Council debates corpus*
- Speech
- WDConcept
  - uri: the Wikidata Unique Resource Identifier (URI) this node represents
  - label: the English string label of this node (taken from property `rdfs:label`)

5.2 Relationships

The following list contains all relationship that link the nodes above with each other. If a relationship has properties, these are also enumerated and explained shortly.

- AGENDA
  - Speech → AgendaItem
- CONTAINS:
  - Speech → Paragraph
  - Speech → Sentence
  - Paragraph → Sentence
- HAS_METADATA
  - Speech → Meta
- MENTIONS
  - Sentence → DBConcept
- NEXT
  - Sentence → Sentence
  - Speech → Speech
  - Paragraph → Paragraph
- owl_sameAs: links a URI in the DBpedia knowledge graph to a URI in the wikidata knowledge graph it corresponds to
  - DBConcept ↔ WDConcept
- wd_P279: points from a class to a superclass
  - WDConcept → WDConcept
- wd_P31: points from an instance to a class
  - WDConcept → WDConcept
  - surfaceForm: the string that has been annotated
  - offset: the character offset within the sentence
- REPRESENTS
  - Speaker → Institution
  - Speaker → Country

- SPOKE
  - Speaker → Speech
  - Speaker → Paragraph
  - Speaker → Sentence

References

[1] Mirco Schoenfeld, Steffen Eckhard, Ronny Patz, Hilde Van Meegdenburg, and Antonio Pires. The UN Security Council Debates. `https://dataverse.harvard.edu/citation?persistentId=doi:10.7910/DVN/KGVSYH`, 2019.