Corpus transformation into RDF - the daily struggle

Martin Brümmer
University of Leipzig
bruemmer@informatik.uni-leipzig.de
Corpus transformation into RDF - Motivation

- Corpora are central for many NLP tasks
- Corpus formats are plenty and often loosely defined or used (TEI, CoNLL, LMF, ...)
- Different from NLP tool in-/output formats
- Constitutes the need for corpus and NLP tool output conversion into various formats
NIF

• The NLP Interchange Format (NIF) is an RDF/OWL-based format
• achieve interoperability between NLP tools, language resources and annotations
• Way of annotating text as well as NLP tool output
What's NIF?

“My favourite actress is Natalie Portman.”
What's NIF?

“My favourite actress is Natalie Portman.”

Tokenizer

#char=3,12
a nif:String, nif:RFC5147String, nif:Word;
nif:anchorOf “favourite”;
nif:referenceContext #char=0,>

nif:beginIndex “3”
nif:endIndex “12”.

Martin Brümmer RDF corpora – the struggle is real
What's NIF?

“My favourite actress is Natalie Portman.”

Create RDF resources for strings based on their string offsets
Same in a corpus, where annotation already exists
What's NIF?

“My favourite actress is Natalie Portman.”

Tokenizer

```r
a nif:String, nif:RFC5147String, nif:Word;
nif:anchorOf "favourite";
nif:referenceContext <#char=0,>;
nif:beginIndex "3";
nif:endIndex "12".
```

Snowball Stemmer

```r
nif:stem "favourit".
```

Stanford Core NLP

```r
nif:oliaLink <http://purl.org/olia/penn.owl#JJ>;
nif:oliaCategory <http://purl.org/olia/olia.owl#Adjective>;
nif:lemma "favorite". [sic]
```

DBpedia Spotlight

```r
itsrdf:taldentRef <http://dbpedia.org/resource/Favourite>;
itsrdf:taConfidence "0.10"^^xsd:decimal.
```
What's NIF?

“My favourite actress is Natalie Portman.”

Tokenizer

```
#char=3,12
a nif:String, nif:RFC5147String, nif:Word;
nif:anchorOf “favourite”; nif:referenceContext #char=0,>
nif:beginIndex “3”;
nif:endIndex “12”.
```

Snowball Stemmer

```
#char=3,12
nif:stem “favourite”.
```

Stanford Core NLP

```
#char=3,12
nif:oliaLink <http://purl.org/olia/penn.owl#JJ>;
nif:oliaCategory <http://purl.org/olia/olia.owl#Adjective>;
nif:lemma “favorite”. [sic]
```

DBpedia Spotlight

```
#char=3,12
itsrdf:taldentRef <http://dbpedia.org/resource/Favourite>;
itsrdf:taConfidence "0.10"^^xsd:decimal.
```

Integration through merged RDF

```
#char=3,12
a nif:RFC5147String, nif:String;
a nif:Word;
nif:anchorOf “favourite”; nif:referenceContext #char=0,>
nif:beginIndex “3”;
nif:endIndex “6”;

nif:stem “favourit”;

nif:oliaLink <http://purl.org/olia/penn.owl#JJ>;
nif:oliaCategory <http://purl.org/olia/olia.owl#Adjective>;
nif:lemma “favorite”;

itsrdf:taldentRef <http://dbpedia.org/resource/Favourite>;
itsrdf:taConfidence "0.10"^^xsd:decimal.
```

@base <http://example.org/prefix>
NIF as pivot format

NIF can serve as a pivot format for corpora

- CoNLL
- TEI
- GATE XML

NLP Tool
Parser
Framework
NIF as pivot format

NIF can serve as a pivot format for corpora

CoNLL → NLP Tool

TEI → Parser

GATE XML → Framework

CoNLL → NIF

TEI → Parser

GATE XML → Framework

Martin Brümmer

RDF corpora – the struggle is real
The struggle

- NIF / RDF produces overhead
- Brown corpus TEI: 56MB; NIF: 1GB
The struggle

- NIF / RDF produces overhead
- Brown corpus TEI: 56MB; NIF: 1GB

- Original order gets lost if corpus is converted into a graph
- Makes it hard to stream sentences in order
- Corpus needs to be fully parsed or indexed first
- Not performant or feasible on very large corpora
The struggle

- NIF / RDF produces overhead
- Brown corpus TEI: 56MB; NIF: 1GB

- Original order gets lost if corpus is converted into a graph
- Makes it hard to stream sentences in order
- Corpus needs to be fully parsed or indexed first
- Not performant or feasible on very large corpora

- New use cases need patching of ontology
The struggle

- RDF not easy to parse
- Entry barrier of RDF, Linked Data and ontologies to be considered
- Linked Data provides few benefits for monolithic resources like corpora
- Most of these challenges are inherent to RDF and not easily overcome.
Do the benefits outweigh the problems? Can we do better?