Aligning the Norwegian UD Treebank with Entity and Coreference Information

Tollef Emil Jørgensen and Andre Kåsen
tollef.jorgensen@ntnu.no, ankas9475@oslomet.no

Highlights

- A merged collection of **named entity and coreference data, grounded in the Universal Dependencies (UD) treebanks** for the two written forms of Norwegian: Bokmål and Nynorsk.
- Thorough **conversion and alignment processes for UD treebanks**, analysis of discovered data inconsistencies.
- A system that supports
  1) specialized conversion and alignment of noisy (subsets of) treebank data.
  2) a generalized alignment system between differing versions of the same treebank.

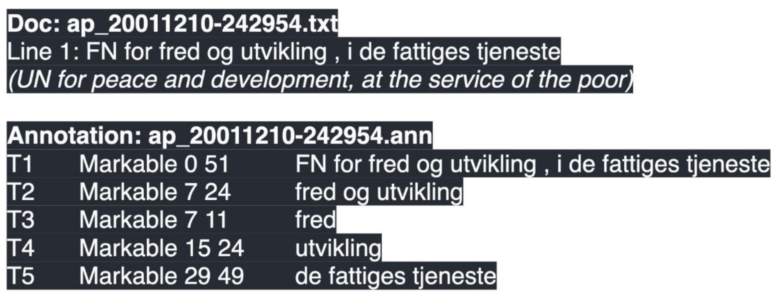
Data

- NDT: The Norwegian Dependency Treebank
- NorNE: Norwegian Named Entities
  - Annotated on an outdated version of UD
- NARC: Norwegian Anaphora Resolution Corpus
  - Differing format, incomplete subsets, document-level IDs
    - UD is sentence-level

Conversion

A coreference conversion tool is developed to handle the BRAT annotation output to JSONLines to CoNLL-U. Discontinuous mentions, bridging, ++

BRAT format (.txt+.ann)

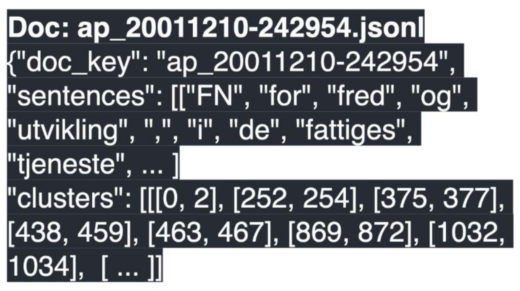


Doc: ap_20011210-242954.txt
Line 1: FN for fred og utvikling , i de fattiges tjeneste
(UN for peace and development, at the service of the poor)

Annotation: ap_20011210-242954.ann

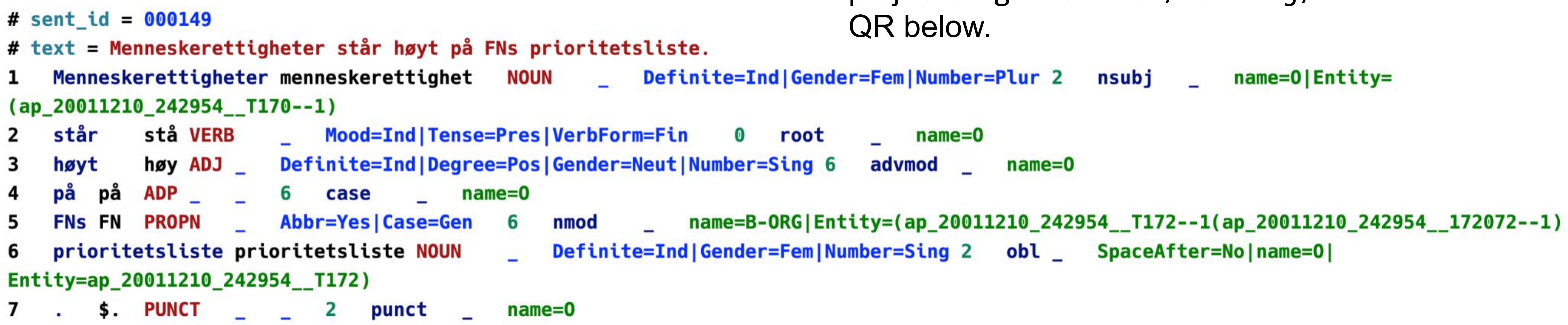
```
T1    Markable 0 51     FN for fred og utvikling , i de fattiges tjeneste
T2    Markable 7 24     fred og utvikling
T3    Markable 7 11     fred
T4    Markable 15 24    utvikling
T5    Markable 29 49    de fattiges tjeneste
```

JSONLine



Doc: ap_20011210-242954.jsonl

```
{"doc_key": "ap_20011210-242954",
"sentences": [["FN", "for", "fred", "og",
"utvikling", ",", "i", "de", "fattiges",
"tjeneste", ... ]
"clusters": [[[0, 2], [252, 254], [375, 377],
[438, 459], [463, 467], [869, 872], [1032,
1034],  [ ... ]]
```



```
# sent_id = 000149
# text = Menneskerettigheter står høyt på FNs prioritetsliste.
1   Menneskerettigheter menneskerettighet   NOUN    _   Definite=Ind|Gender=Fem|Number=Plur 2   nsubj   _   name=O|Entity=
(ap_20011210_242954__T170--1)
2   står    stå VERB    _   Mood=Ind|Tense=Pres|VerbForm=Fin    0   root    _   name=O
3   høyt    høy ADJ _   Definite=Ind|Degree=Pos|Gender=Neut|Number=Sing 6   advmod  _   name=O
4   på  på  ADP _   _   6   case    _   name=O
5   FNs FN  PROPN   _   Abbr=Yes|Case=Gen   6   nmod    _   name=B-ORG|Entity=(ap_20011210_242954__T172--1(ap_20011210_242954__172072--1)
6   prioritetsliste prioritetsliste NOUN    _   Definite=Ind|Gender=Fem|Number=Sing 2   obl _   SpaceAfter=No|name=O|
Entity=ap_20011210_242954__T172)
7   .   $.  PUNCT   _   _   2   punct   _   name=O
```

Alignment

NARC is based on the source material in the treebank, but has several omitted documents/sentences.

`[w1, w2, w3, ..., wN-1, wN]` (UD)
`[w1, w2, w3, ..., wN-1, wN]` (annotated)

We perform direct and lazy matching of the UD and target dataset, attempting to minimize alignment scores. Multiple matched candidates are resolved by the linear sum assignment problem.

$$C_{i,j} = \mathrm{sent\_to\_UD\_dist\_score}(N_i, U_j)$$

Examples of noisy data:

NARC sentence	UD sentence
Illustrasjonsfoto .	Illustrasjonsfoto
Illustrasjonsfoto |	Illustrasjonsfoto
Illustrasjonsofoto |	Illustrasjonsfoto .
Nei !	- Nei ?
Nei !	- Nei .
- Ja .	Ja .

Bokmål	UD	NorNE	NARC	Aligned
Sentences	20,044	20,045	16,461	15,672
Tokens	310,221	310,222	257,646	244,136
Entities	-	20,134	-	16,271
Markables	-	-	55,225	52,815
Mentions	-	-	77,565	73,983
SplitAnte	-	-	140	134
Bridging	-	-	1,060	1,025

How to use it

Align any annotated dataset based on a treebank in your language with a single line of code. It can be merged with the official UD treebank. Visit the project on `github.com/tollefj/UD-NARC`
QR below.

github.com/tollefj/UD-NARC