Homework 2

Concepts and Applications in NLP

December 6, 2024

1 Extracting information from dependency-parsed data

Universal Dependencies (UD) is a project that is developing cross-linguistically consistent treebank annotation for many languages.[1] As corpus to work with, get the English portion of the Parallel Universal Dependencies (PUD) treebanks: `https://github.com/UniversalDependencies/UD_English-PUD`. It contains 1000 sentences, taken from the news domain (sentence id starts in 'n') and from Wikipedia (sentence id starts with 'w').

1.1 Assignment 1(a)

Get familiar with the structure and the annotation of the UD data. You can find an overview of the format here: `https://universaldependencies.org/format.html`. The different dependency relations are explained here: `https://universaldependencies.org/format.html`.

Consider a sentence from the corpus, for example this one:

```
# sent_id = n01005031
# text = The feasibility study estimates that it would take passengers about four minutes
to cross the Potomac River on the gondola.
 1  The          the          DET   DT  Definite=Def|PronType=Art  3   det       3:det
 2  feasibility  feasibility  NOUN  NN  Number=Sing                3   compound  3:compound
 3  study        study        NOUN  NN  Number=Sing                4   nsubj     4:nsubj
 4  estimates    estimate     VERB  VBZ Mood=Ind|Number=Sing|
                                        Person=3|Tense=Pres|
                                        VerbForm=Fin               0   root      0:root
 5  that         that         SCONJ IN  _                          8   mark      8:mark
 6  it           it           PRON  PRP Case=Nom|Gender=Neut|
                                        Number=Sing|Person=3|
                                        PronType=Prs               8   expl      8:expl
 7  would        would        AUX   MD  VerbForm=Fin               8   aux       8:aux
 8  take         take         VERB  VB  VerbForm=Inf               4   ccomp     4:ccomp
 9  passengers   passenger    NOUN  NNS Number=Plur                8   iobj      8:iobj
 10 about        about        ADV   RB  _                          11  advmod    11:advmod
 11 four         four         NUM   CD  NumForm=Word|NumType=Card  12  nummod    12:nummod
 12 minutes      minute       NOUN  NNS Number=Plur                8   obj       8:obj
 13 to           to           PART  TO  _                          14  mark      14:mark
 14 cross        cross        VERB  VB  VerbForm=Inf               8   csubj     8:csubj
 15 the          the          DET   DT  Definite=Def|PronType=Art  17  det       17:det
 16 Potomac      Potomac      PROPN NNP Number=Sing                17  compound  17:compound
 17 River        River        PROPN NNP Number=Sing                14  obj       14:obj
 18 on           on           ADP   IN  _                          20  case      20:case
 19 the          the          DET   DT  Definite=Def|PronType=Art  20  det       20:det
 20 gondola      gondola      NOUN  NN  Number=Sing                14  obl       14:obl:on
 21 .            .            PUNCT .   _                          4   punct     4:punct
```

[1] `https://universaldependencies.org/`

- What is the root?
- What pairs of verb-(in)direct object and verb-subject can you find?

1.2 Assignment 1(b)

Write a (python) script that extracts pairs of verbs and direct objects (only nouns).

Input: the file *en_pud-ud-test.conllu*

Output:
sentence-id1 *-tab-* V-Obj1, V-Obj2, ...
sentence-id2 *-tab-* V-Obj1, V-Obj2, ...

Always output lemmas, not the inflected forms.

Please submit your homework by mail on or before **07. January 2025**