Readability of legal texts

Class #11, April 26 2022

Jana Plaňavová Latanowicz  j.planavova-latanowicz@uw.edu.pl
Barbora Hladká  hladka@ufal.mff.cuni.cz
Part I :: Natural Language Processing and Legal Domain

- HW #5 evaluation
- Czech Legal Text Treebank
HW #5 :: assignment

Universal Dependencies and UDPipe
see Lecture #5: video (from 00:58:00)

Subject annotation

Annotation task

Annotate subjects in the sentences of the preamble of the EU regulation 2020/2092 on a general regime of conditionality for the protection of the Union budget. Our motivation to organize this task is twofold:

- evaluate UDpipe system on manually annotated data
- get experience with readability of legal texts

Once a candidate country becomes a Member State, it joins a legal structure that is based on the fundamental premiss

The homework should be completed by April 19.

Annotation instructions

- To log in to Brat editor use the credentials sent via e-mail.
- Read carefully the preamble and identify subject(s) in each sentence. We follow the Universal Dependencies annotation guidelines where the basic units of annotation are (syntactic) words, which means that the subject is exactly one word in our annotation task. Typically, it is a noun, pronoun or relative pronoun. Mark all subjects standing in a coordinated construction separately.
Subject annotation as a machine learning task

- Feature vectors
- Real world objects
- Words in texts
- True target function $f$
- Approximation of $f$
- Target values
- Subject 0/1

$x = <x_1, \ldots, x_m>$
Subject annotation as a machine learning task

- Words in texts
- Feature vectors
- True target function $f$
- True prediction
- Approximation of $f$
- $x = <x_1, ..., x_m>$

HW #5
Examples (gold data)

its preamble (29 items)
What we have - manually annotated subjects in CS, PL, FR versions, each version by 2 annotators

<table>
<thead>
<tr>
<th>Languages, formats and link to OJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>HTML</td>
</tr>
<tr>
<td>CS</td>
</tr>
<tr>
<td>PL</td>
</tr>
<tr>
<td>FR</td>
</tr>
<tr>
<td># of tokens</td>
</tr>
<tr>
<td>CS 2,285</td>
</tr>
<tr>
<td>PL 2,259</td>
</tr>
<tr>
<td>FR 3,160</td>
</tr>
</tbody>
</table>
What we have - a complex syntactic analysis of each version using UDPipe
Inter-Annotator Agreement :: CS

Confusion matrix for 2 annotators treating A1 CS as gold standard and A2 CS as classifier

<table>
<thead>
<tr>
<th></th>
<th>A1 CS</th>
<th>A2 CS</th>
<th>(\sum)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1 CS</td>
<td>1</td>
<td>0</td>
<td>116</td>
</tr>
<tr>
<td>A2 CS</td>
<td>113</td>
<td>33</td>
<td>146</td>
</tr>
<tr>
<td>(\sum)</td>
<td>116</td>
<td>N-116</td>
<td>N</td>
</tr>
</tbody>
</table>

F1 = \(\frac{2 \times 0.97 \times 0.77}{0.97 + 0.77}\) = 0.85
Precision = \(\frac{113}{113+3}\) = 0.97
Recall = \(\frac{113}{111+33}\) = 0.77

N = total number of words in Czech version

Data Analytics for Students of Social Studies and Humanities  [https://ufal.mff.cuni.cz/courses/npfl134](https://ufal.mff.cuni.cz/courses/npfl134)
### Inter-Annotator Agreement :: CS

Confusion matrix for 2 annotators treating A2 CS as gold standard and A1 CS as classifier

<table>
<thead>
<tr>
<th>A1 CS</th>
<th>A2 CS</th>
<th>(\Sigma)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>113</td>
<td>33 146</td>
</tr>
<tr>
<td>0</td>
<td>3</td>
<td>N-149 N-146</td>
</tr>
<tr>
<td>(\Sigma)</td>
<td>116</td>
<td>N-116 N</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A2 CS</th>
<th>A1 CS</th>
<th>(\Sigma)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>113</td>
<td>3 116</td>
</tr>
<tr>
<td>0</td>
<td>33</td>
<td>N-149 N-116</td>
</tr>
<tr>
<td>(\Sigma)</td>
<td>146</td>
<td>N-146 N</td>
</tr>
</tbody>
</table>

**F1** = \(2 \times 0.97 \times 0.77 / (0.97 + 0.77) = 0.85\)

Precision = 113/(113+3) = 0.97

Recall = 113/(111+33) = 0.77

N = total number of words in Czech version

N = total number of words in Czech version

Data Analytics for Students of Social Studies and Humanities  [https://ufal.mff.cuni.cz/courses/npfl134](https://ufal.mff.cuni.cz/courses/npfl134)
Inter-Annotator Agreement

A1 A2 CS: 0.86
A1 A2 PL: 0.83
A1 A2 FR: 0.72
UDPipe evaluation
Accounting units shall take inventory of their assets and liabilities pursuant to section 29 and 30. The relation of obligation is present: what accounting units have to do – take inventory.

Czech Legal Text Treebank  [http://hdl.handle.net/11234/1-2498](http://hdl.handle.net/11234/1-2498)

is a collection of 1,133 manually annotated sentences from two Czech legal documents: The Accounting Act (563/1991 Coll., as amended) and Decree on Double-entry Accounting for undertakers (500/2002 Coll., as amended):

- morphologically
- syntactically
- entities = terms from the accounting domain
- relations of definition, right, obligation
Example

(3) Účetní jednotky, …, tvoří opravné položky a rezervy …

(3) Accounting units, …, create fixed items and reserves …

(accounting units, create, fixed items)

(accounting units, create, reserves)
A survey on Subject annotation

1. How much time did you spend on the annotation task?
2. Do you consider reading and understanding the text difficult?
Part II :: Readability of legal texts
Why readability of legal texts and policy documents is important?

- rule of law element ([https://worldjusticeproject.org/](https://worldjusticeproject.org/))
  - accountability
  - law is clear, publicized, stable and is applied evenly
  - accessibility of law and justice (incl. enforcement)
  - part of good administration
- economic impact (legal counselling is not attainable)
- need to understand and translate law into technical specifications (e.g. EU regulations on the technological sectors like digital single market legislation)
Why measuring readability of legal texts and policy documents is important?

- decline in literature skills (incl. understanding, evaluating and using written texts)
- intentional obfuscation (e.g. lawyers gobbledygook)
- complexity of topics covered
- multilinguality of EU legislation is taken into account during the law-making processes
How to measure readability of EU law? (State of the art)


sample of 201 EU documents on Single Digital Market

5 readability indices:

- mostly based on the grade level of education (eg. score 9,3 - ninth grader would be able to read the text, score 22 for a university graduate)
- [https://pypi.org/project/textstat/](https://pypi.org/project/textstat/)

Conclusion: hypothetical grade level around 30

Criticism: does not take into account different types of EU acts and documents
QuitUp comparison and Ruohonen research

Caveats:

- stylometry (QuitUp) vs readability (Ruohonen)
- multilingual vs monolingual approach
QuitaUp

- [Link](https://korpus.cz/quitaup/)
- pre-processes input texts using UDPipe
  - i.e., tokenization, lemmatization, POS tagging, syntactic parsing
You cannot end a sentence with because because because is a conjunction.

- tokens = smaller units in a text
- token length = number of characters (because 7)
- word types = different tokens in a text (9)
- function words (synsemantic words) have little lexical meaning (prepositions, conjunctions, pronouns, …) you, a, with, is, cannot
- content words (autosemantic words) possess semantic content (nouns, adjectives, most verbs, most adverbs) end, sentence, conjunction
QuitaUp :: Comparison

What we have compared:

- 6 official translations of a preamble of the Regulation 2020/2092 of 16 December 2020 on a general regime of conditionality for the protection of the Union budget
- mechanical translation of the Regulation from English into Russian by using https://lindat.mff.cuni.cz/services/translation
- A.A. Milne “Winnie-the-Pooh” (in original language, about 1,5 chapter, number of characters roughly similar to the one of Regulation’s preamble in English)
Token-related measures

Number of tokens

![Graph showing token-related measures for different languages. The x-axis represents the number of tokens, ranging from 0 to 4000. The y-axis represents the languages: EN, FR, IT, DE, CS, PL, RU, and Winnie. The graph displays the token counts for each language.](https://ufal.mff.cuni.cz/courses/npfl134)
Number of different tokens

- EN
- FR
- IT
- DE
- CS
- PL
- RU
- Winnie
Average token length
Word types

Vocabulary richness

- EN
- FR
- IT
- DE
- CS
- PL
- RU
- Winnie

Data Analytics for Students of Social Studies and Humanities  https://ufal.mff.cuni.cz/courses/npfl134
Words that appear just once in the text (hapaxes)
Content words

Activity (verbs/(verbs+adjectives)) vs descriptivity

---

Data Analytics for Students of Social Studies and Humanities  [https://ufal.mff.cuni.cz/courses/npfl134](https://ufal.mff.cuni.cz/courses/npfl134)
Thematic concentration

Thematic concentration (how much the text focuses on the main topic or topics, while the main topic is detected using th...
## Thematic Concentration :: CS and EN

<table>
<thead>
<tr>
<th>Word</th>
<th>POS</th>
<th>TW (primary TC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>unie</td>
<td>NOUN</td>
<td>0.0314</td>
</tr>
<tr>
<td>státu</td>
<td>NOUN</td>
<td>0.0189</td>
</tr>
<tr>
<td>komise</td>
<td>NOUN</td>
<td>0.0145</td>
</tr>
<tr>
<td>opatření</td>
<td>NOUN</td>
<td>0.0145</td>
</tr>
<tr>
<td>měla</td>
<td>VERB</td>
<td>0.0102</td>
</tr>
<tr>
<td>právního</td>
<td>ADJ</td>
<td>0.0069</td>
</tr>
<tr>
<td>zásad</td>
<td>NOUN</td>
<td>0.0028</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Word</th>
<th>POS</th>
<th>TW (primary TC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>law</td>
<td>NOUN</td>
<td>0.0063</td>
</tr>
<tr>
<td>union</td>
<td>NOUN</td>
<td>0.0056</td>
</tr>
<tr>
<td>member</td>
<td>PROPN</td>
<td>0.0033</td>
</tr>
<tr>
<td>measures</td>
<td>NOUN</td>
<td>0.0025</td>
</tr>
<tr>
<td>rule</td>
<td>NOUN</td>
<td>0.0025</td>
</tr>
<tr>
<td>commission</td>
<td>NOUN</td>
<td>0.0016</td>
</tr>
<tr>
<td>financial</td>
<td>ADJ</td>
<td>0.0009</td>
</tr>
</tbody>
</table>
Conclusions

- Readability of legal acts is language-specific (it should be measured for each language)
- Measuring readability of legal documents should reflect the type of the legal document in question (e.g. EU regulation, national law, contracts, policy documents)
- Readability of legal documents should be defined against “average user” of the given type of document (generally applied rules vs individually addressed rules)
Further reading

  https://journals.sagepub.com/doi/10.1177/002194368101800405
- In Czech: Jak umělá inteligence chápe občanský zákoník (Noc vědců, 2020)
  https://www.youtube.com/watch?v=3EtQIzQ7CsE (from 3:00:00)