Teaching computers to understand politicians

or: Semantic Analyses of Swedish Parliamentary Data

Name: Stian Rødven Eide†
Date: May 23, 2019

†stian.rodven.eide@svenska.gu.se
The Swedish parliamentary data is available from
- Riksdagen, the Swedish parliament
  HTML encapsuled in XML/JSON/SQL or plain text
- Språkbanken, University of Gothenburg
  Linguistically annotated XML

Some things are not in Språkbanken
- Ledamöter (data on MPs)
- Voteringar (voting data)
- Anföranden (speeches - with extensive metadata)
<corpus id="rd-prot">
  <dokument (document metadata)>
    <paragraph>
      <sentence>
        <ne (named entity, time - when identified in the text)>
          <w (linguistic annotation)>
        </w>
      </ne>
    </sentence>
  </paragraph>
</dokument>
My work on parliamentary data

* Semantic analyses, including
  * Named entity recognition/resolution (NER)
  * Argumentation mining (AM)
  * Sentiment analysis (SA)

* Also in a diachronic perspective, tracking changes over time
My mostly used sources

* Anföranden (speeches)
  * Contains the same text as the protocols, but with extensive metadata
* Ledamöter (MPs)
  * Provides additional metadata for NER
* Voteringar (votes)
  * Basic feature for SA, useful for AM
* Neither of these are in Språkbanken, but can easily be annotated through the Sparv API (annotation pipeline)
A typical speech document

<table>
<thead>
<tr>
<th>Property</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>dok_hangar_id</td>
<td>Internal document ID</td>
</tr>
<tr>
<td>dok_id</td>
<td>Meeting + speech no.</td>
</tr>
<tr>
<td>dok_titel</td>
<td>Protocol title</td>
</tr>
<tr>
<td>dok_rm</td>
<td>Parliamentary year</td>
</tr>
<tr>
<td>dok_nummer</td>
<td>Number of meeting</td>
</tr>
<tr>
<td>dok_datum</td>
<td>Date of speech</td>
</tr>
<tr>
<td>avsnittsrubrik</td>
<td>Topic title</td>
</tr>
<tr>
<td>underrubrik</td>
<td>Topic subtitle</td>
</tr>
<tr>
<td>kammaraktivitet</td>
<td>Type of debate</td>
</tr>
<tr>
<td>anforande_id</td>
<td>Unique speech ID</td>
</tr>
<tr>
<td>anforande_nummer</td>
<td>Speech number in debate</td>
</tr>
<tr>
<td>talare</td>
<td>Speaker name</td>
</tr>
<tr>
<td>parti</td>
<td>Speaker party</td>
</tr>
<tr>
<td>anforandetext</td>
<td>Full speech text</td>
</tr>
<tr>
<td>intressent_id</td>
<td>Speaker’s ID</td>
</tr>
<tr>
<td>rel_dok_id</td>
<td>Document being debated</td>
</tr>
<tr>
<td>replik</td>
<td>Speech type</td>
</tr>
<tr>
<td>systemdatum</td>
<td>Date of publishing</td>
</tr>
</tbody>
</table>
* I work with Python and Prolog, using JSON from Riksdagen
* JSON can trivially be converted to a Python dictionary
* Prolog for linking and querying data
  * A speech may be replying to a person without using a name
  * With linked data, this can more easily be resolved
  * Also, additional metadata may provide useful features for machine learning
A semantic graph of MPs and debates

Position
- Group-ID
- Role
- Start date
- End date
- Position type
- Assignment
- Status

MP-Name —> MP-ID —> Speech —> Topic

Party-ID —> Wiki-ID —> Meeting —> Date

Wiki-Name