From Text to Networks

Tutorial @ DH 2018, Montreal

Nils Reiter, Sandra Murr, Max Overbeck, Evgeny Kim
Today’s Goals

• Generate networks from raw texts
  • Nodes represent entity annotations in the text
  • Edges represent co-occurrence counts
    • Co-occurrence within a certain text segment
• Focus on methods, not tools
  • Tools serve as an example, but other tools could be used instead
  • Technical steps are relatively simple
    • Possible for “early-career” programmers
• Meta-Goals
  • Showcasing a modularized research workflow
<table>
<thead>
<tr>
<th>Time</th>
<th>Agenda</th>
</tr>
</thead>
<tbody>
<tr>
<td>09:30</td>
<td>Lecture: Annotation theory, guidelines, entity annotations</td>
</tr>
<tr>
<td>09:45</td>
<td>Hands-on: Annotation, inter-annotator agreement</td>
</tr>
<tr>
<td>10:45</td>
<td>Discussion</td>
</tr>
<tr>
<td>11:00</td>
<td>Coffee Break</td>
</tr>
<tr>
<td>11:30</td>
<td>Lecture: Tool-support, segment annotation</td>
</tr>
<tr>
<td>12:00</td>
<td>Hands-on: Text Segmentation and Network Exports</td>
</tr>
<tr>
<td>12:45</td>
<td>Discussion</td>
</tr>
<tr>
<td>13:00</td>
<td>End</td>
</tr>
</tbody>
</table>
Annotation

Theory

• Explicit assignment of categories to text spans
  • Text spans are explicitly bounded (begin, end)

• Annotation by humans vs. annotation by computers
  • Focus on humans (in the first part)
Annotation
Inter-subjective Annotations

• Linguistic annotations: Inter-subjective annotations
  • E.g., part of speech
  • Different annotators create the same annotations
  • Inter Annotator Agreement (IAA)

• Annotation guidelines
  • Mediator between theory and annotation
  • Guidebook for annotators (who may not be experts)
    • What is to be annotated?
    • Which categories are used in which cases?
    • How to deal with borderline cases?
    • How have we dealt with difficult cases in the past?
    • Tons of examples
# Annotation Guidelines

## Contents

1. Introduction .................................................. 1
2. List of parts of speech with corresponding tag ............. 1
3. List of tags with corresponding part of speech .......... 6
4. Confusing parts of speech .................................. 7
5. Specific words and collocations ............................. 22
6. General tagging conventions ................................. 31
   6.1 Part of speech and syntactic function .................. 31
   6.2 Vertical slash convention ................................ 31
   6.3 Capitalized words ........................................ 31
   6.4 Abbreviations ............................................ 32

TOC of Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision)
Annotation
Iterative Workflow

corpus

- pilot annotations
- analysis

- guidelines v1

- annotations
- analysis

- guidelines v2

theory
Annotation
Analysis

• Parallel annotation of the same text by multiple annotators
  • Without talking to each other

• Manual comparison
  • By the annotators themselves
  • By third person(s) – “adjudication”

• Quantitative comparison
  • Inter-Annotator Agreement
  • Kappa $\kappa$: “agreement of the annotators above chance” (Fleiss, 1971)
  • Upper limit for automatic annotation

• Yes, it’s time-consuming
Annotation

Why

• Empirical validation of theories
  • Discovering phenomena not covered by a theory
  • Strengthening definitions in a theory
    • Often confused categories might be overlapping or at least unclear
  • Uncovering implicit assumptions

• Data creation
  • Manually annotated data can be analyzed
    • Which categories are how frequent in what context?
  • Automatic tools can be evaluated
    • How well do machines do this task?
  • Supervised tools can be trained
Entities and Entity References
Entities

- Entities are specific objects that are distinguishable by naming in a real or fictional world.
- “Object” does not (necessarily) imply a physical object
- Entity references can be:
  - **Proper names:**
    - Justin Trudeau is the current Prime Minister of Canada
  - **Generic names (appellative noun phrases):**
    - The three Ministers with the blue suits.
  - **Pronouns (we do not annotate):**
    - He will run for president.
Entities and Entity References

Text

Persons

Locations

World

University of Stuttgart
Annotation Guidelines

• Our aim is to achieve full annotation of relevant entities in a given corpus

• Reference expressions are *maximum nominal phrases* (NPs), that is, nouns with preceding/subsequent text parts which further specify the noun
  • E.g., relative clauses
    • [The maid who had most responsibility] was Anna
CRETA Entity Classes

• Six entity classes
  • Person (PER)
  • Places (LOC)
  • Organizations (ORG)
  • Works (WRK)
    • [The Treaties of Rome]_{WRK} were signed in Rome in 1957.
  • Events (EVT)
    • [September 11]_{EVT} has changed everything.
  • Abstract concepts (CNC)
    • [Our common European values]_{CNC} have to be defended.
Hands On Session 1

Tasks

1. Parallel, but individual annotation of the same text by two participants (10 minutes)
2. Comparison and discussion of the parallel annotations (15 minutes)
3. Define additional rules to improve the guidelines (15 minutes)
4. Application of the improved guidelines on a different text (10 minutes)
Hands On Session 1
Annotation Texts

• Narrative texts
  • Around the world in 80 days (Jules Verne)
  • Alice’s Adventure in Wonderland (Lewis Carroll)
  • Huckleberry Finn (Mark Twain)
  • Narrative of the Captivity and Restoration of Mrs. Mary Rowlandson (Mary Rowlandson)

• News articles from Montreal Gazette
  • Straight shooters: Women outfoxing men in DodgeBow
  • In praise of science 'superclusters’

• Several speeches from the European Parliament