Lecture 12: Annotation Wrap, Homework 2 Revisited

LING 1340/2340: Data Science for Linguists

Na-Rae Han

Objectives

- Linguistic annotation
  - Annotation formats
  - Evaluation and quality control
- Homework 2 revisited

Homework 2 pitfalls

	Total token #	Total sentence #	Avg sentence length
Low	300172	13349	22.48
Medium	2206853	94177	23.43
high	1678805	71067	23.62

1. Obtaining averages from flattened groups →
   - We did this back in LING 1330…
2. Producing per-essay measurements, but results are saved as a *separate series*, not inserted into essay_df.
   - Problem? The values become detached from their original essays: you can no longer connect them to additional attributes (L1, Prompt, token count…)
   - You should build out essay_df with per-essay measurements, and use it as your exploration base. Use .groupby() and filters to zone in on particular attributes and further narrow down… on the fly!
3. Only ever looking at three average numbers (for low, medium, high) in drawing conclusions. (And ANOVA.)
   - Remedy? Look at overall DISTRIBUTION.
   - Use .describe(), boxplots.
4. For-loops for processing each row.
   - This is NOT the pandas way!

We must adapt to the new **pandas way**, which at long last allows us a proper statistics treatment.

Per-sample measurements are the ground truth! You then closely examine their **distribution**.

An anatomy of annotation project

- Suppose you are tasked to start up an annotation project:
  - Error annotation of a set of essays written by ESL learners
  - Audio files of sociolinguistic interviews
  - A set of videos featuring ASL content
- What should you be figuring out?
  1. Annotation scheme
  2. Physical representation
  3. Annotation process
  4. Evaluation and quality control
  5. Usage

I agree greatly this topic mainly because I think that English becomes an official language in the not too distant. Now, many people can speak English or study it all over the world, and so more people will be able to speak English. Before the Japanese fall behind other people, we should be able to speak English, therefore, we must study English not only junior high school students or over but also pupils. Japanese education system is changing such a program. …

Adapted from p.9 of Ide & Pustejovsky eds. (2017), *Handbook of Linguistic Annotation*

Annotation scheme

- Error annotation of a set of essays written by ESL learners
- Audio files of sociolinguistic interviews
- A set of videos featuring ASL content

1. Is there an underlying theory? What is it?
2. What features should be targeted and how should they be organized?
3. What is the process of annotation scheme development?
4. Should the potential use of the annotations inform development of the annotation scheme?
5. Will development of the scheme inform the development of linguistic theories or knowledge?

Physical representation

- Error annotation of a set of essays written by ESL learners
- Audio files of sociolinguistic interviews
- A set of videos featuring ASL content

1. How is the annotation represented? What **format**? Standards?
2. What are the reasons for the particular representation chosen?
   - What are the advantages/disadvantages of the chosen representation that may have come to light through its use?
   - Is the chosen format easily convertible into some other format down the line?
3. What **annotation software tools** are capable of handling them?

Linguistic annotation format: standardize?

- Ad-hoc formats mean different linguistic annotations are often incompatible
- Converting back and forth between them wastes resource
- Solution: Standardized format for linguistic annotation
- FoLiA: Format for Linguistic Annotation
  - https://proycon.github.io/folia/
  - XML-based architecture
  - Software support, Python libraries etc.!

Example: semantic role

- https://folia.readthedocs.io/en/latest/semrole_annotation.html

```
      <provenance>
        <processor xml:id="p1" name="proycon" type="manual" />
      </provenance>
    </metadata>
    <text xml:id="example.text">
      <p xml:id="example.p.1">
        <s xml:id="example.p.1.s.1">
          <t>The Dalai Lama greeted him.</t>
          <w xml:id="example.p.1.s.1.w.1"><t>The</t></w>
          <w xml:id="example.p.1.s.1.w.2"><t>Dalai</t></w>
          <w xml:id="example.p.1.s.1.w.3"><t>Lama</t></w>
          <w xml:id="example.p.1.s.1.w.4"><t>greeted</t></w>
          <w xml:id="example.p.1.s.1.w.5" space="no"><t>him</t></w>
          <w xml:id="example.p.1.s.1.w.6"><t>.</t></w>
          <semroles>
           <predicate class="greet">
            <semrole class="agent">
                <wref id="example.p.1.s.1.w.2" />
                <wref id="example.p.1.s.1.w.3" />
            </semrole>
            <semrole class="patient">
                <wref id="example.p.1.s.1.w.5" />
            </semrole>
           </predicate>
          </semroles>
        </s>
      </p>
    </text>
</FoLiA>
```

Annotation format

- To XML or not to XML?
  - Gina Peirce's Russian learner corpus:

```
▼<essay>
 ▼<tunit>
    Россия является частью Европы потому-что Россияни одеваются обычно по моде, так-же как дру
    страны Европы, и так-же многие считают что они более подобны белой Европе чем Азии.
  </tunit>
 ▼<tunit>
    Политика в России отличается от Китая и например Индии.
  </tunit>
 ▼<tunit>
    У нас нет систем
    <err cf="каст" pos="nn" gnd="fm" cs="g" num="pl" t="cs">касты</err>
    .
  </tunit>
 ▼<tunit>
    Даже если Россия чуть опаздывает от Европы по моде или например
    <err cf="восточным" pos="adj" gnd="ms" num="pl" cs="d" t="cs num">восточныя</err>
    услугам, у нас все равно есть просвещение в отлицие от предедущих времён.
  </tunit>
 ▼<tunit>
    Язык у нас так-же полнастью не похож на те-же Азиатские эроглифы.
  </tunit>
 ▼<tunit>
    К мнению что основная часть России в Азии все равно не повод не считать Россиян Европейцам
  </tunit>
</essay>
```

Annotation format

- Inline or stand-off?
  - **Inline annotation** has annotations occurring alongside the text. Often used for describing a single structural element (ex. per-token)
    - Example: The Brown corpus, Gina Peirce's corpus
    - Pros: simple, self-contained. An XML parser is all you need.
    - Cons: May not be suitable for multi-layer annotations.
    - Folia page on In-line annotation: https://folia.readthedocs.io/en/latest/inline_annotation_category.html
  - **Stand-off annotation** has an annotation existing in a separate layer, typically as a separate file. Annotation points to an *offset* or a *span*.
    - Folia page on Span annotation: https://folia.readthedocs.io/en/latest/span_annotation_category.html

Stand-off annotation: an example

- Original text: "Mia visited Seoul to look me up yesterday."

```
<maf xmlns:"http://www.iso.org/maf">
<seg type="token" xml:id="token1">Mia</seg>
<seg type="token" xml:id="token2">visited</seg>
<seg type="token" xml:id="token3">Seoul</seg>
<seg type="token" xml:id="token4">to</seg>
<seg type="token" xml:id="token5">look</seg>
<seg type="token" xml:id="token6">me</seg>
<seg type="token" xml:id="token7">up</seg>
<seg type="token" xml:id="token8">yesterday
</seg>
<pc>.</pc>
</maf>
```

Word tokens: inline segmentation

```
<isoTimeML xmlns:"http://www.iso.org./isoTimeML">
<TIMEX3 xml:id="t0" type="DATE" value="2009-10-20"
functionInDocument="CREATION_TIME"/>
<EVENT xml:id="e1" target="#token2" class="OCCURRENCE" tense="PAST"/>
<EVENT xml:id="e2" target="#token5 #token7"class="OCCURRENCE"
tense="NONE" vForm="INFINITIVE"/>
<TIMEX3 xml:id="t1" type="DATE" value="2009-10-19"/>
<TLINK eventID="#e1" relatedToTime="#t0" relType="BEFORE"/>
<TLINK eventID="#e1" relatedToTime="#t1" relType="ON_OR_BEFORE"/>
<TLINK eventID="#e2" relatedToTime="#t1" relType="IS_INCLUDED"/>
</isoTimeML>
<tei-isoFSR xmlns:"http://www.iso.org./tei-isoFSR">
<fs xml:id="t0"><f name="Type" value="2009-10-20"/></fs>
</tei-isoFSR>
```

Time Event Annotation: stand-off annotation

TimeML annotation standard

Annotation process

1. Will the annotation be done *manually*, *automatically*, or via some combination of the two?
2. Manual annotation:
   - How many annotators? Their background?
   - What annotation environment/platform will be used?
   - What are the exact steps? Multiple passes involving multiple annotators? Pipeline?
   - How will inter-annotator agreement be computed?
3. Automatic annotation:
   - What software will be used to generate the annotations?
   - How well does this software generally perform? Will it be a good fit with your data?

Evaluation and quality control

1. Systematic scaffolding to minimize human error?
2. By what method(s) will the quality of the annotations evaluated?
   - Inter-annotator agreement (IAA)
3. What is the threshold for the quality of annotations?

Inter-annotator agreement

- An important part of quality control
- Necessary to demonstrate the **reliability** of annotation.
- Common practices:
  - Create **"gold" annotation** (deemed "correct") to evaluate individual annotators' output against
  - Designate a portion of data to be annotated by **multiple annotators**, then measure **inter-annotator agreement**
  - **Pre-** and **post-adjudication** agreement: do disagreements persist after an adjudication process?

Inter-annotator agreement: factors

- Agreement rate depends on two main factors:
  - Quality of annotators: how well-trained the annotators are
  - Complexity of task: how difficult or abstract the annotation task at hand is, how easy it is to clearly delineate the category

🡐 IMPORTANT because human agreement (esp. post-adjudication) is considered a **CEILING** for performance of machine-learning!

How much will humans agree?

- POS tagging
  - Via Universal Dependency POS tagset?
  - Using the Penn Treebank tagset?
- Syntactic tree bracketing for Penn Treebank
  - Reported to be about 88% (F-score)
- Scoring TOEFL essays, 0 to 5
  - Reported to be about 80% (Cohen's kappa)

🡐 Is there hope for automated essay grading?

Cohen's kappa

- Good or bad level of agreement?
  - **Case A**: Movie reviews are annotated as "rotten" or "fresh". Two annotators agree 70% of the time.
  - **Case B**: Student essays are rated from 0 to 5. Two annotators agree 70% of the time.
- **Cohen's kappa (K) coefficient** is one of the most widely used measures of inter-annotator agreement.
  - Accounts for "chance" agreement.

$$\kappa \equiv \frac{p_o - p_e}{1 - p_e}$$

P_o: observed agreement
P_e: probability of chance agreement

P_e is 0.5 in Case A, 0.17 in Case B.
Case A:
K = (0.7 - 0.5) / (1 - 0.5) = **0.4**
Case B:
K = (0.7 - 0.17) / (1 - 0.17) = **0.64**

Wrapping up

- Your project
  - Progress Report #1 due this Friday!