Lecture 15: Linguistic Annotation

LING 1340/2340: Data Science for Linguists

Na-Rae Han

Objectives

- Linguistic annotation project: considerations for planning
- Annotation standards
  - Format
  - Inter-annotator agreement

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

  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?
2. What are the reasons for the particular representation chosen?
3. What are the advantages/disadvantages of the chosen representation that may have come to light through its use?
4. What software or system was used to generate the annotated data?

Annotation process

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

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

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

1. By what method(s) will the quality of the annotations evaluated?
   - Inter-annotator agreement (IAA)
2. What is the threshold for the quality of annotations?

Usage

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

1. By what means and under what conditions will the data be available to users?
2. What are the expected usages of the annotated data?
3. Will the data be used for machine learning, and if so what types of task?

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 другие
    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 Европейцами.
  </tunit>
</essay>
```

Annotation format

- Inline or stand-off?
  - **Inline annotation** has annotations occurring alongside the text.
    - Example: The Brown corpus, Gina Peirce's corpus
    - Pros: simple, self-contained. An XML parser is all you need.
    - Cons: Text-annotation relation is contextually determined. May not be suitable for multi-layer annotations.
  - **Stand-off annotation** has an annotation existing in a separate layer, typically as a separate file. Annotation points to an offset or a span.

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

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

- Happy Thanksgiving!
- 3rd progress report due 11/28 (Tue).
- 11/28 (Tue)
  - Multimodal annotation
  - Project presentation: margeret
- Presentation schedule
  - 11/28 (Tue) Margaret
  - 11/30 (Thu) Katherine, Paige, Andrew
  - 12/5 (Tue) Alicia, Chris, Ben
  - 12/7 (Thu) Dan, Robert Kyle