COSI-230B: Natural Language Annotation for Machine Learning

Lecture 5: Annotation Tools: brat — Events & Emotion Annotation

Jin Zhao

Brandeis University
Computational Linguistics Program

January 28, 2025

Review

- **Review Questions:**
  - What is the difference between offset annotations vs. in-line annotations?
  - What is crowd-sourcing?
  - What is CC-BY-NC?
  - What is an API?

Always Annotate with a Tool

Golden Rule

Always use a tool to annotate!

Why?

- Difficult or impossible to enforce restrictions without a task-specific schema
- Errors easily crop up if annotators just input labels in a spreadsheet or text file
- Quickly becomes difficult to enforce annotators use the same specification
- Maintains consistency across annotators and annotation sessions

Intro to brat

brat rapid annotation tool

`https://brat.nlplab.org`

Installation:

- Clone the brat repository from GitHub: `https://github.com/nlplab/brat`
- Run `standalone.py` with Python 3.8+ to start a local brat instance

Useful Resources:

- brat manual: `https://brat.nlplab.org/manual.html`
- brat standoff annotation format: `https://brat.nlplab.org/standoff.html`

NER Example in brat

Two important configuration files for span annotations:

1. `annotation.conf` — defines the annotation schema
2. `visual.conf` — defines visual appearance of annotations

With these files, we can annotate spans for Named Entity Recognition (NER) tasks:

- Persons, Organizations, Locations
- Dates, Times, Numbers
- Custom entity types for your task

annotation.conf Example

```
# This is a minimal example configuration

[entities]

# Definition of entities.
# Format is a simple list with one type per line.
# Hierarchy can be imposed via indentation with tabs.

Person
Organization
Location
Date
```

File: ∼/brat/data/your-project/annotation.conf

visual.conf Example

```
[labels]

# Label definitions for display.
# Labels are separated by pipe characters "|".

Person | PER
Organization | ORG
Location | LOC
Date | DATE

[drawing]

Person    bgColor:#ffcccc
Organization  bgColor:#ccffcc
Location  bgColor:#ccccff
Date      bgColor:#ffffcc
```

File: ∼/brat/data/your-project/visual.conf

Homework 1 Preview

You'll need to:

1. Set up brat with a schema (`annotation.conf` file)
2. Load in the data assigned to you (under your `.../brat/data` directory)

Two annotation tasks:

- **Task 1 (Sequence labeling):** Events in English news articles
- **Task 2 (Sentence classification):** Emotion in Reddit comments

You'll use the guidelines we develop together in class today!

Events: What Are They?

- Often events are **verbs**:
  - “He **ran** down the street.”
- But they can also be **nouns**:
  - “The **election** was fiercely contested.”
- Or even **adjectives** (representing changed states):
  - “The volcano was **dormant** for centuries before the eruption.”

Focus: For this task, we focus on **event triggers** — the key words that refer to an event.

Events: Existing Resources

Existing event annotation datasets:

- **TimeML and ACE:** Event triggers, owned by LDC (not freely available)
- **LitBank:** Events in literature, not bounded by certain types
  - "Literary Event Detection" — ACL Anthology

Our task goal:

- Detect events in **news text**
- Potential downstream use: linking events across articles
- Following a coreference or event tracking pipeline

Event Specification

Frame it as a span labeling problem:

- Only focusing on event “trigger” spans
  - Key words or multi-word expressions that refer to an event
- **Not** worried about: actors, temporal information, location, etc.

Questions to consider (discuss in groups):

- What types of events should we annotate?
- What events should *not* be annotated?
- How to handle boundary cases?

Event Specification: Details

Additional considerations:

- **Temporal information:** Include or exclude?
- **Nested events:** How to handle events within events?
- **Participants:** Who are the actors involved?
- **Related words:** Include relevant context words?
- **Conditional events:**
  - e.g., “If the election results...” — is this an event?

Class Activity: Event Guidelines

Group Discussion

What types of events should we propose?
What are examples of each type?

We will develop annotation guidelines as a class and use them for HW 1.

Emotion Classification

Framing as a comment classification task:

- One emotion label per comment
- Classification (not span labeling)

Question: What emotion classes should we use?

- Basic emotions (happy, sad, angry, etc.)?
- Fine-grained categories?
- Domain-specific emotions?

Class Activity: Emotion Guidelines

Group Discussion

What emotion labels should we decide on?
What are examples of each label we choose?

We will develop annotation guidelines as a class and use them for HW 1.

Opinion Mining: Problem Definition

An opinion is a quadruple: (g, s, h, t)

- $g =$ the opinion/sentiment **target** (what is being evaluated)
- $s =$ the **sentiment** about the target
- $h =$ the opinion **holder**
- $t =$ the **time** when the opinion was expressed

Example:

"Posted by: John Smith, Date: September 10, 2011
(1) I bought a Canon G12 camera..."

Opinion Mining: Entity Definition

Entity: A product, service, topic, issue, person, organization, or event.

An entity e is described with a pair $e(T, W)$:

- T = a hierarchy of **parts**, sub-parts, and so on
- W = a set of **attributes** of e

Example:

- Entity: “Canon G12 camera”
- Parts: lens, sensor, battery, screen
- Attributes: picture quality, size, weight, price

Opinion Mining: Refined Definition

An opinion is a quintuple: $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$

- e_i = the name of an **entity**
- a_{ij} = an **aspect** of e_i
- s_{ijkl} = the **sentiment** on aspect a_{ij} of entity e_i
- h_k = the opinion **holder**
- t_l = the **time** when the opinion was expressed

Goal: Given a document d, discover all opinion quintuples in d.

Sentiment Analysis Tasks

Key subtasks:

1. **Entity extraction:** Category and mention identification
2. **Aspect extraction:** Category and expression identification
    - Explicit: “The **picture quality** of this camera is great.”
    - Implicit: “This camera is **heavy**.” (implies weight aspect)
3. **Sentiment classification:** Positive, negative, neutral
4. **Holder identification:** Who expressed the opinion?
5. **Time extraction:** When was it expressed?

Aspect-Based Sentiment Analysis

Example:

"Posted by: big John, Date: Sept. 15, 2011
(1) I bought a Samsung camera and my friends brought a Canon camera yesterday.
(2) In the past week, we both used the cameras a lot.
(3) The photos from my Samsung are not that great, and the battery life..."

Multiple entities, aspects, and sentiments in one document:

- Samsung camera — photos — negative
- Samsung camera — battery life — ?
- Canon camera — ? — ?

Emotion Classification Datasets

Existing emotion classification resources:

- **SemEval 2019 Task 3: EmoContext**
  - Classes: Happy, Angry, Sad, Other
- **Sentence and Clause Level Emotion (Multi-Genre Corpus)**
  - 10 classes: Joy, Trust, Anticipation, Surprise, Sadness, Fear, Anger, Disgust, Other-emotion, No-emotion
- **SenTube: YouTube Comments**
  - Positive or negative about the video

Wrap-up

Key Takeaways:

- Always annotate with an annotation tool when possible
- Need to **model the domain** before we annotate
- Initial modeling may have flaws, but we can **iteratively improve**
- Developed class guidelines for Events and Emotions

Homework:

- **HW 0** due Friday, January 26th at midnight
- **HW 1** will be posted on Monday, January 29th

Next Class: Continue annotation specifications