Building an Annotated Dataset of Literary Entities and Events

David Bamman
School of Information, UC Berkeley
dbamman@berkeley.edu

In collaboration with Matt Sims, Jerry Park, Sejal Popat and Sheng Shen
Computational Humanities

Ted Underwood (2018), “Why Literary Time is Measured in Minutes”

Ryan Heuser, Franco Moretti, Erik Steiner (2016), The Emotions of London

Richard Jean So and Hoyt Long (2015), “Literary Pattern Recognition”


Franco Moretti (2005), Graphs, Maps, Trees

Holst Katsma (2014), Loudness in the Novel


Plot

http://www.matthewjockers.net/2014/06/05/a-novel-method-for-detecting-plot/
Reagan et al. (2016), "The emotional arcs of stories are dominated by six basic shapes"
Plot

We’re decomposing plot into structured elements

<table>
<thead>
<tr>
<th>Element</th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characters</td>
<td>Entity recognition</td>
</tr>
<tr>
<td>Events</td>
<td>Event detection</td>
</tr>
<tr>
<td>Setting</td>
<td>Entity recognition, setting coreference</td>
</tr>
<tr>
<td>Objects</td>
<td>Object detection/coreference</td>
</tr>
<tr>
<td>Time</td>
<td>Temporal processing, event ordering</td>
</tr>
<tr>
<td>NLP Task</td>
<td>Accuracy</td>
</tr>
<tr>
<td>--------------------------</td>
<td>---------------------------</td>
</tr>
<tr>
<td>Tokenization</td>
<td>100%</td>
</tr>
<tr>
<td>Part-of-speech tagging</td>
<td>98.0% [Bohnet et al. 2018]</td>
</tr>
<tr>
<td>Named entity recognition</td>
<td>93.1 [Akbik et al. 2018]</td>
</tr>
<tr>
<td>Syntactic parsing</td>
<td>95.1 F [Kitaev and Klein 2018]</td>
</tr>
<tr>
<td>Coreference resolution</td>
<td>73.0 F [Lee et al. 2018]</td>
</tr>
</tbody>
</table>
Data in NLP
Syntax

Long is the way and hard, that leads up to light.
Active work

- Domain adaptation
  [Chelba and Acero, 2006; Daumé and Marcu, 2006; Daumé 2009; Duong et al. 2015; Glorot et al. 2011, Chen et al. 2012, Yang and Eisenstein 2014, Schnabel and Schütz 2014]

- Contextualized word representations
  [Peters et al. 2018; Devlin et al. 2018; Howard and Ruder 2018; Radford et al. 2019]

- Data annotation. 200,000 tokens from 100 different novels, annotated for:
  - Entities (person/place, etc.)
  - Events
  - Coreference
  - Quotation attribution
"TOM!"
No answer.

"TOM!"
No answer.

"What's gone with that boy, I wonder? You TOM!"
No answer.

The old lady pulled her spectacles down and looked over them about the room.
Literary entities

Mr. Knightley, a sensible man about seven or eight-and-thirty, was not only a very old and intimate friend of the family, but particularly connected with it, as the elder brother of Isabella's husband.
Mr. Knightley, a sensible man about seven or eight-and-thirty, was not only a very old and intimate friend of the family, but particularly connected with it, as the elder brother of Isabella's husband.

Austen, Emma
Mr. Knightley, a sensible man about seven or eight-and-thirty, was not only a very old and intimate friend of the family, but particularly connected with it, as the elder brother of Isabella's husband.
Nested entity recognition

- Recognize spans of text that correspond to categories of entities (whether named or not).

...the elder brother of Isabella's husband
Dataset

- 100 books from Project Gutenberg
- Mix of high literary style (e.g., Edith Wharton’s *Age of Innocence*, James Joyce’s *Ulysses*) and popular pulp (Haggard’s *King Solomon’s Mines*, Alger’s *Ragged Dick*).
- Select first 2000 words from each text
Entity classes

- **Person.** Single person with proper name (Tom Sawyer) or common entity (the boy); set of people (her daughters).

- **Organization.** Formal association (the army, the Church as an administrative entity).

- **Vehicle.** Devices primarily designed to move an object from one location to another (ships, trains, carriages).
Entity classes

- **GPE.** Entities that contain a population, government, physical location and political boundaries (New York, the village)

- **Location.** Entities with physicality but w/o political status (New England, the South, Mars), including natural settings (the country, the valley, the forest)

- **Facility.** Functional, primarily built structure designed for habitation (buildings), storage (barns), transportation (streets) and maintained outdoor space (gardens).
Metaphor

• Only annotate phrases whose types denotes an entity class.

John is a doctor

the young man was not really a poet; but surely he was a poem

Chesterton, *The Man Who Was Thursday*
Personification

• Person includes characters who engage in dialogue or have reported internal monologue, regardless of human status (includes aliens and robots as well).

As soon as I was old enough to eat grass my mother used to go out to work in the daytime, and come back in the evening.

Sewell, Black Beauty
Metonymy

• Describing one concept by a closely related one (e.g., “the White House said…”)

• Annotate the evoked entity class (PER here rather than FAC).

“Them men would eat and drink if we was all in our graves,” said the indignant cook, who indeed had a real grievance; and the outraged sentiment of the kitchen was avenged by a bad and hasty dinner

Oliphant, *Miss Marjoribanks*
## Data

<table>
<thead>
<tr>
<th>Cat</th>
<th>Count</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>PER</td>
<td>9,383</td>
<td>my mother, Jarndyce, the doctor, a fool, his companion</td>
</tr>
<tr>
<td>FAC</td>
<td>2,154</td>
<td>the house, the room, the gardne, the drawing-room, the library</td>
</tr>
<tr>
<td>LOC</td>
<td>1,170</td>
<td>the sea, the river, the country, the woods, the forest</td>
</tr>
<tr>
<td>GPE</td>
<td>878</td>
<td>London, England, the town, New York, the village</td>
</tr>
<tr>
<td>VEH</td>
<td>197</td>
<td>the ship, the car, the train, the boat, the carriage</td>
</tr>
<tr>
<td>ORG</td>
<td>130</td>
<td>the army, the Order of Elks, the Church, Blodgett College</td>
</tr>
</tbody>
</table>
How well can find these entity mentions in text as a function of the training domain?
Data

- ACE (2005) data from newswire, broadcast news, broadcast conversation, weblogs
Prediction


- Evaluate performance difference when altering the training/test domain.
Prediction

• F1 by category (train on ACE/Literature) and test on Literature.
Analysis

- Tag entities in 1000 new Gutenberg texts (78M tokens) using the two models (ACE vs. LIT) and analyze the difference in frequencies with which a given string is tagged as **PER** under both models.
MOSCOW, April 17 (AFP)

Silence is golden -- especially when your hand is weak -- top Moscow policy analysts said in an assessment of the fallout from Russia's vocal opposition to what turned out to be a swift US-led campaign in Iraq.

Several top diplomacy experts told a Kremlin-run forum that countries like China and India that said little about the conflict before its March 20 launch were already reaping the benefits.

Some suggested that Russian President Vladimir Putin will now be scrambling to contain the damage to his once-budding friendship with US President George W. Bush because he was poorly advised by his intelligence and defense aides.
Analysis

- How well does each model identify entities who are men and women?

- We annotate the gender for all PER entities in the literary test data and measure the recall of each model with respect to those entities.

<table>
<thead>
<tr>
<th>Training</th>
<th>Women</th>
<th>Men</th>
<th>Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACE</td>
<td>38.0</td>
<td>49.6</td>
<td>-11.6</td>
</tr>
<tr>
<td>Literary</td>
<td>69.3</td>
<td>68.2</td>
<td>1.1</td>
</tr>
</tbody>
</table>
LitBank

LitBank is an annotated dataset of 100 works of fiction to support tasks in natural language processing and the computational humanities, described in more detail in:


LitBank is licensed under a Creative Commons Attribution 4.0 International License.
Events

- Event trigger detection, slot filling
  [MUC, ACE, DEFT ERE]

- Veridicality, factuality & committed belief
  [Saurí and Pustejovsky 2009; de Marneffe et al. 2012, Werner et al. 2015, Lee et al. 2015, Rudinger et al. 2018]

- Temporal grounding
  [Pustejovsky et al. 2003]

- Narrative event chains, schemas
  [Chambers and Jurafsky 2008, Cheung et al. 2013]
Events

Realist view: events are things that happen.

Aristotle, Russell, Whitehead, Quine, Vendler, Montague, Davidson, Dowty, etc.
Events in language

• Events are typically realized through verbs (and some nominalizations)
  
  • He walked down the street
  
  • He had a nice walk.
Polarity

John walked by Frank and didn’t say hello.

<table>
<thead>
<tr>
<th>Value</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>positive</td>
<td>depicted as taking place</td>
</tr>
<tr>
<td>negative</td>
<td>depicted as not taking place</td>
</tr>
</tbody>
</table>
Tense

I walked to the store and will buy some groceries

past

future

Value

past

present

future
## Specificity

<table>
<thead>
<tr>
<th>Value</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>specific</td>
<td>singular occurrence at a particular place and time</td>
</tr>
<tr>
<td>general</td>
<td>claim about groups, abstractions</td>
</tr>
</tbody>
</table>

My son just watched *Frozen*

Kids like *Frozen*
I walked to the store to buy some groceries
Modality

Beliefs: Rumors of my demise have been greatly exaggerated.

Hypotheticals: If you visit Rockridge, try Zachary’s.

Commands: John was ordered to return the book or face a fine.

Threats: AMI threatened to release the photos if Bezos did not comply.

Desires: She wants to go to Rome.
Dataset

• 100 books from Project Gutenberg (same as for entity annotations)

• First 2000 words from each text

• 7,892 events
My father’s eyes had closed upon the light of this world six months, when mine opened on it.

Dickens, *David Copperfield*

Call me Ishmael

Melville, *Moby Dick*
Event detection

Can we detect events that have occurred?

0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0

My father’s eyes had closed upon the light of this world six months, when mine opened on it.

Call me Ishmael

Dickens, *David Copperfield*

Melville, *Moby Dick*
Baseline

All (and only) verbs are events

Call me Ishmael
<table>
<thead>
<tr>
<th>Feature</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word identity</td>
<td>Call</td>
</tr>
<tr>
<td>Part of speech</td>
<td>VB</td>
</tr>
<tr>
<td>Context</td>
<td>L:\∅, R:“me Ishmael”, L_POS:\∅, R:“PRP NNP”</td>
</tr>
<tr>
<td>Wordnet synset</td>
<td>[name, call]</td>
</tr>
<tr>
<td>Dependency label + head</td>
<td>root</td>
</tr>
</tbody>
</table>
Featurized model

- Syntactic information: is subject a bare plural?
- Is the subject countable?

My son just watched Frozen

Kids like Frozen

Reiter and Frank (2010)
Featurized model

L2-regularized logistic regression
Neural

- Bidirectional LSTM + CNN (with subword CNN)
- Word embeddings learned from 15,000 Project Gutenberg texts
- Sentence-level CNN (Feng et al. 2016)
Bidirectional LSTM

Word embedding + subword character CNN

CNN + max pooling (Feng et al. 2016)

Call me Ishmael.
Neural

- Bidirectional LSTM + CNN (with subword CNN)
- Word embeddings learned from 15,000 Project Gutenberg texts
- Sentence-level CNN (Feng et al. 2016)
BERT

- Contextualized word representations from BERT base model + bidirectional LSTM.
- No fine-tuning on domain or task.
Setting coreference

• Which events take place at the same physical location in the narrative?
Questions

• Measuring the abstractness of a novel by the density of realis events.

<table>
<thead>
<tr>
<th>Events</th>
<th>Book</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.109</td>
<td>The Quest of the Silver Fleece</td>
</tr>
<tr>
<td>0.091</td>
<td>The Invisible Man</td>
</tr>
<tr>
<td>0.083</td>
<td>Of Human Bondage</td>
</tr>
<tr>
<td>0.081</td>
<td>The Man of the Forest</td>
</tr>
<tr>
<td>0.078</td>
<td>Gulliver's Travels</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>0.006</td>
<td>The Legend of Sleepy Hollow</td>
</tr>
<tr>
<td>0.004</td>
<td>The Mysteries of Udolpho</td>
</tr>
<tr>
<td>0.003</td>
<td>Middlemarch</td>
</tr>
<tr>
<td>0</td>
<td>The Magnificent Ambersons</td>
</tr>
</tbody>
</table>
Questions

• Measuring the abstractness of a novel by the density of realis events.

• Modeling the distribution of settings over narrative time
What are the textual signals of authorial prestige?

“Prestige” =

- Inclusion in ODNB, MLA, Stanford exam lists (Algee-Hewitt et al. 2016)
- Reviewed by elite literary journals (Underwood 2019)
Prestige

- 100 authors identified in Underwood (2019) with highest and lowest prestige (as measured by the number of times their works were reviewed by elite literary journals); 150 high-prestige novels + 188 low-prestige novels.

- Metric = frequency of *realis* events (normalized by number of tokens)

<table>
<thead>
<tr>
<th>Class</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>High prestige</td>
<td>4.3 [4.2-4.5]</td>
</tr>
<tr>
<td>Low prestige</td>
<td>5.0 [4.9-5.1]</td>
</tr>
</tbody>
</table>
• Low prestige novels have less variability in depiction of realis events — something is always happening.
Thanks!

David Bamman
dbamman@berkeley.edu

Matt Sims, Jerry Park and David Bamman, “Literary Event Detection” (ACL 2019)

David Bamman, Sejal Popat and Sheng Shen, “An Annotated Dataset of Literary Entities” (NAACL 2019)

https://github.com/dbamman/litbank