**Datasets**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Dataset** | **Task** | **#book** | **#token** | **#annotation** | **#type** | **by** | **Time range** |
| LitBank | Entity | 100 works of English-language fiction | 210,532 from first 2000 words of each work | 13,912 | 6 | Human | 1719 - 1923, mainly 1852-1911 |
| Event | 7,849 | 1 |
| Coreference | Need to check dataset | 3 |
| Speaker attribution | 1765 | 1 |

**Tasks from** [**BookNLP**](https://github.com/booknlp/booknlp#entity-annotations)**:**

[**Entity annotations**](https://people.ischool.berkeley.edu/~dbamman/pubs/pdf/naacl2019_literary_entities.pdf)

Challenge compared with existing datasets:

* Model trained on out-of-domain data degrade in performance when applying to literary data
* Severity of degradation is hard to be estimated
* Non-named entities and nested structure are included in literary text(different with existing dataset built mainly on news, e.g. ACE)

Annotations (generally follow ACE annotation guidelines):

* Restricted to proper and common noun phrases
* Label: PER, FAC, GPE, LOC, VEH, ORG
* Specific rules for figurative language: metaphor, personification, metonymy

Experiment:

* Layered bi-LSTM-CRF trained and tested on ACE/LIT, performance of consistent dataset is the best
* Terms with high positive difference of frequency of tagged by each model, show potential gender bias in ACE

[**Event annotations**](https://people.ischool.berkeley.edu/~dbamman/pubs/pdf/acl2019_literary_events.pdf)

Why important: events are important in how literary narratives progress

Challenge compared with news:

* Complex narrative structure both locally (individual scenes) and globally (plot)
* Causality connecting events is not hard coded into literary event sequences

Annotations

* Categories of relationship between verbs and time: activities, achievements, accomplishments, changes of states
* Rules:
  + 4 dimensions for tagging events following ACE/Light ERE: positive polarity, past/present tense, specific events with particular place and time, asserted events
  + Common standards: event trigger is a single word, limited to verbs, adj and nouns. No restrictions on the types of events
  + Standards for stative events:

o Explicit change of states has changed and it can be determined solely by context

o Cause can be deduced, cause and resulting state occurred at same location

o Potential event trigger refers to a mental state which is inherently acute

Experiments

* Event detection
  + Neural models for event trigger detection (LSTM, BiLSTM/with document context/sentence CNN/subword CNN)
  + Featurized models
* Relationship between depiction of realis events and prestige/popularity of author
  + Compare ratio and distance of high/low prestige/popularity
  + High-prestige author use less concrete events, high degree of freedom

[**Character name clustering and coreference**](https://arxiv.org/pdf/1912.01140.pdf)

Gap:

* Benchmark OntoNotes datasets for coreference resolution focus on news, conversation, bible, web, English literature lacks robust data for coreference
* Different behaviors of coreference in literary texts

Annotations (styles largely follows OntoNotes):

* Markable mentions (Departures from OntoNotes):
  + All singleton mentions are markable
  + Markable entity types: annotated entities only
  + Entity categories: proper names, common phrases, all personal pronouns (common forms, historical forms, transcriptions of speech)
  + All quantified and negated noun phrases are included
  + Maximal extent of a span is marked, including honorifics
* Criteria of establishing coreference
  + Coreference: links a mention in text to a discourse entity
    - between generic mentions and others is not allowed
    - Near-identity: neutralization/compression and refocusing/decompression
    - Annotated from reader’s point of view, at any point in the narrative
  + Copulae: links an attribute mention to another mention, only those asserted as currently holding true
  + Apposition: links an appositive expression to another mention
* Process:
  + Link 29103 mentions to entities in text by a custom command-line interface
  + Transfer annotations to a GUI interface for checking

Experiment

* Gold mentions
* Mention identifier

Supersense tagging

This paper outlines a compact version of the Word Sense Disambiguation task (WSD) and an extension of the NER task.

NER: The task of identifying/tagging words that fall under some of the following fixed categories: “person”, “location”, “organization” etc. NER tagging usually doesn’t have more than 5-6 categories. It is a sequence tagging task.

WSD: WSD is inherently the task of differentiating words with ambiguous meaning. There are tens of thousands of word senses (Imagine a classification task with tens of thousands of classes). Obtaining a dataset robust enough to differentiate such level of granularity is challenging.

Based on Wordnet’s supersenses (lexicographers classes) this paper uses a tagset of **41 supersenses**: “act”, “animal” “artifact” “shape” “contact” “emotion” “weather” etc.

Corpus used to train: **SemCor Corpus** ([A SEMANTIC CONCORDANCE](https://dl.acm.org/doi/pdf/10.3115/1075671.1075742))

This is a tagged corpus which takes the text from the Brown Corpus and the tags from WordNet synset (synonym set).

The brown corpus is one of the oldest corpus containing 500 samples, each approx. 2000 words long for a total of 1 million words. The genres of these samples range from newspaper reporting to technical writing, to fiction to philosophical essays.

**Referential gender inference**

No specific paper mentioned for this task. The task is to tag the referential gender (he/him etc.) and not the gender identity (Male/Female etc.).

Book NLP uses prior information (text corpus) for the alignment of names with referential gender. Meaning, they use historical information, but they acknowledge that contextual information is also important for referential gender inference. (e.g., "Tom" in the book "Tom and Some Other Girls", where Tom is aligned with she/her).

The prior information used is [The Project Gutenberg’s Fifteen Thousand Useful phrase](https://www.gutenberg.org/cache/epub/18362/pg18362-images.html)s. This can be ignored and BookNLP will annotate based on just the context of the input corpus.

**[Speaker attribution](https://arxiv.org/pdf/2004.13980.pdf)**

This simply attributes the speakers to the speeches in the book. The model is trained on the same LitBank dataset used in Entity, Event and Character Name Clustering tasks.

The paper referenced for this task measures information propagation in literary social networks. The speaker attribution task is just one segment of the entire pipeline.

Previous works in Speaker Attribution are limited in scope (e.g. 6 texts by Austen, Dickens etc.)

Annotations carried out by [brat: a Web-based Tool for NLP-Assisted Text Annotation](https://aclanthology.org/E12-2021.pdf) which is a combination of Human + web-interface to manually annotate texts.

Information propagation modeling process

* Identify a set of unique characters using coreference resolution (existing in litbank)
* Attribute dialogue to characters:
  + Quotation identification: simple regular expression “”
  + Quotation attribution: map quotes to nearest speaker, link identified speaker mentions to character entity
* Build a graph for speakers - listeners interaction based on conversation
* Identify information:
  + Identify quoted speech referencing >=1 character, extract tuple (subject, verb, object)
  + select tuples with term in 4 categories (amorous, hostile, juridical vital), each category includes synonyms presenting in top 100 frequent words
* Identify implicit propagation:
  + Propositional tuple passes by ABC, where character AB are co-present and B repeats proposition to C in a different conversation block
* Identify explicit propagation: ‘[character-id] said/declared/mentioned, etc’
* 6 network measures which captures topological properties in network

Annotation process: quotation identification & attribution in modeling

**Tasks from** [**David Bamman**](https://scholar.google.com/citations?hl=en&user=RkA1y54AAAAJ&view_op=list_works&sortby=pubdate):

[Annotating Character Relationships in Literary Texts](https://arxiv.org/pdf/1512.00728.pdf)

[The Labeled Segmentation of Printed Books](https://aclanthology.org/D17-1077.pdf)

[New Alignment Methods for Discriminative Book Summarization](https://arxiv.org/pdf/1305.1319.pdf)

[Discovering Multilingual Text Reuse in Literary Texts](http://www.perseus.tufts.edu/publications/2009-Bamman.pdf)

[A Framework for (Under)specifying Dependency Syntax without Overloading Annotators](https://arxiv.org/pdf/1306.2091.pdf)

[An Ownership Model of Annotation: The Ancient Greek Dependency Treebank](https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.211.9736&rep=rep1&type=pdf#page=17)