An Annotated Dataset of Literary Entities

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

Sejal Popat
School of Information
UC Berkeley
sejal@berkeley.edu

Sheng Shen
Computer Science Division
UC Berkeley
sheng.s@berkeley.edu

Abstract

We present a new dataset comprised of 210,532 tokens evenly drawn from 100 different English-language literary texts annotated for ACE entity categories (person, location, geo-political entity, facility, organization, and vehicle). These categories include non-named entities (such as "the boy", "the kitchen") and nested structure (such as [[the cook]'s sister]). In contrast to existing datasets built primarily on news (focused on geo-political entities and organizations), literary texts offer strikingly different distributions of entity categories, with much stronger emphasis on people and description of settings. We present empirical results demonstrating the performance of nested entity recognition models in this domain; training natively on in-domain literary data yields an improvement of over 20 absolute points in F-score (from 45.7 to 68.3), and mitigates a disparate impact in performance for male and female entities present in models trained on news data.

1 Introduction

Computational literary analysis works at the intersection of natural language processing and literary studies, drawing on the structured representation of text to answer literary questions about character (Underwood et al., 2018), objects (Tenen, 2018) and place (Evans and Wilkens, 2018).

Much of this work relies on the ability to extract entities accurately, including work focused on modeling (Bamman et al., 2014; Iyyer et al., 2016; Chaturvedi et al., 2017). And yet, with notable exceptions (Vala et al., 2015; Brooke et al., 2016), nearly all of this work tends to use NER models that have been trained on non-literary data, for the simple reason that labeled data exists for domains like news through standard datasets like ACE (Walker et al., 2006), CoNLL (Tjong Kim Sang and De Meulder, 2003) and OntoNotes (Hovy et al., 2006)—and even historical non-fiction (DeLozier et al., 2016; Rayson et al., 2017)—but not for literary texts.

This is naturally problematic for several reasons: models trained on out-of-domain data surely degrade in performance when applied to a very different domain, and especially for NER, as Augenstein et al. (2017) has shown; and without in-domain *test* data, it is difficult to directly estimate the severity of this degradation. At the same time, literary texts also demand slightly different representations of entities. While classic NER models typically presume a flat entity structure (Finkel and Manning, 2009), relevant characters and places (and other entities) in literature need not be flat, and need not be named: *The cook's sister ate lunch* contains two `PER` entities ([The cook] and [The cook's sister]).

We present in this work a new dataset of entity annotations for a wide sample of 210,532 tokens from 100 literary texts to help address these issues and help advance computational work on literature. These annotations follow the guidelines set forth by the ACE 2005 entity tagging task (LDC, 2005) in labeling all nominal entities (named and common alike), including those with nested structure. In evaluating the stylistic difference between the texts in ACE 2005 (primarily news) and the literary texts in our new dataset, we find considerably more attention dedicated to people and settings in literature; this attention directly translates into substantially improved accuracies for those classes when models are trained on them. The dataset is freely available for download under a Creative Commons ShareAlike 4.0 license at `https://github.com/dbamman/litbank`.

Proceedings of NAACL-HLT 2019, pages 2138–2144
Minneapolis, Minnesota, June 2 - June 7, 2019.

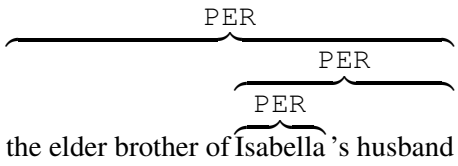
2 Corpus

We draw our corpus from the public-domain texts on Project Gutenberg, selecting individual works of fiction (both novels and short stories) that include a mix of high literary style (e.g., Edith Wharton's *Age of Innocence*, James Joyce's *Ulysses*) and popular pulp fiction (e.g., H. Rider Haggard's *King Solomon's Mines*, Horatio Alger's *Ragged Dick*). All texts are published before 1923 (the current threshold for public domain in the United States), with the majority falling between 1852 and 1911.

From each text, we select approximately the first 2,000 words as a sample, yielding a total dataset of 210,532 tokens.

3 Annotation

We adopt the ACE 2005 guidelines for entity annotation, focusing on the subset of people (PER), natural locations (LOC), built facilities (FAC), geo-political entities (GPE), organizations (ORG) and vehicles (VEH).[1] While traditional named entity recognition presumes a flat structure in which entity labels cannot be embedded within each other, we allow for nested structure, as in the following (from Jane Austen's *Emma*):



This nested structure is in fact quite common in our data, with entities that contain at least one level of nesting accounting for 13.8% of the annotations—86.2% contain no nesting (as in *Isabella* above), 12.5% contain one level (*Isabella's husband*), 1.2% contain two (*the elder brother of Isabella's husband*), and 0.1% contain three. The dataset contains a total of 13,912 entity annotations.

3.1 Entity types

We generally follow the ACE annotation guidelines for each of the entity classes and restrict our annotations to proper and common noun phrases (i.e., excluding pronouns or WH-question words); table 1 illustrates examples for each class.

[1]We exclude the ACE category of weapons (WEA), since that class is rarely attested in our data.

PER. By person we describe a single person indicated by a proper name (Tom Saywer) or common entity (the boy); or set of people, such as *her daughters* and *the Ashburnhams*.

FAC. ACE guidelines define a facility as a "functional, primarily man-made structure" designed for human habitation (buildings, museums), storage (barns, parking garages), transportation infrastructure (streets, highways), and maintained outdoor spaces (gardens) (LDC, 2005). We adopt the ACE threshold for taggability here as well, and rooms and closets within a house as the smallest possible facility.

GPE. Geo-political entities are single units that contain a population, government, physical location, and political boundaries (LDC, 2005). In literary data, this includes not only cities that have known geographical locations within the real world (London, New York), or nations (England, the United States), but also both named and common imagined entities as well (the town, the village).

LOC. Locations describe entities with physicality but without political entities. In our dataset, this includes named regions without political organization (New England, the South) and planets (Mars). The most common class, however, are geologically designated areas describing natural settings, such as the sea, the river, the country, the valley, the woods, and the forest.

VEH. Literary texts include a number of vehicles defined as "a physical device primarily designed to move an object from one location to another" (LDC, 2005); ships, trains, and carriages dominate since nearly all texts were written before the rise of automobiles.

ORG. Organizations are defined by the criterion of formal association and are relatively rare in literary data, comprising the least frequently occurring entity class. The most frequent organizations include the army and the Church (as an administrative entity, distinct from the church as a facility with a physical location).

3.2 Figurative language

Literary language in particular presents several unique challenges to entity annotation, including metaphor, personification and metonymy.

Entity type	Count	Examples
PER	9,383	my mother, Jarndyce, the doctor, a fool, his companion
FAC	2,154	the house, the room, the garden, the drawing-room, the library
LOC	1,170	the sea, the river, the country, the woods, the forest
GPE	878	London, England, the town, New York, the village
VEH	197	the ship, the car, the train, the boat, the carriage
ORG	130	the army, the Order of Elks, the Church, Blodgett College

Table 1: Entity classes annotated in literary data.

Metaphor. For non-figurative texts, predicative structures like *John is a doctor* nearly always entail the two comparands to be identical in their entity type (here, *John* and *a doctor* are both PER). Literary texts, however, are awash in figurative metaphor, such as "the young man was not really a poet; but surely he was a poem" (Chesterton, *The Man Who Was Thursday*). In such cases where the metaphor takes a predicative structure of *x is y*, we annotate only those phrases whose type describes an entity class (in this case, labeling *a poet* as a PER, but not *a poem*).

Personification. Several works, such as London's *The Call of the Wild* and Sewell's *Black Beauty*, feature personified animals as main characters, with dialogue and evident cognition. We expand the criteria for PER to include such characters who engage in dialogue or have reported internal monologue, regardless of their human status (this includes depicted non-human life forms in science fiction, such as aliens and robots, as well).

Metonymy. Metonymy is a rhetorical device of describing one aspect of a concept by a closely related one (such as the *White House* to refer to the organization of government it houses). We see many examples of metonymy in literature, such as the following:

> '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)

Following ACE, we annotate such examples by annotating the evoked entity class; in this case, annotating *the kitchen* as a PER (describing a set of cooks who feel outrage) rather than as a FAC.

3.3 Annotation process

Two co-authors annotated all 100 texts with a single pass between them after an initial phase of discussions about the annotation process, difficulties encountered and formalizing annotation decisions specific to literary texts. At the end of annotating, the inter-annotator agreement was calculated by double-annotating the same five texts and measuring the F1 score. We find that agreement rate to be high (86.0 F), likely due to the existence of thorough previous guidelines in ACE that both annotators were able to reference during the process of annotation.[2]

4 Comparison with ACE

We can compare the properties of this dataset to those of the ACE 2005 annotated data. To enable an apples-to-apples comparison, we filter the ACE data to exclude entity labels for tokens that are marked with a mention type of pronoun (PRO) or WH-question (WHQ) and remove all weapon (WEA) labels; we consider only the subsets for broadcast conversation (bc), broadcast news (bn), newswire (nw) and weblog (wl), as in past work (Lu and Roth, 2015; Muis and Lu, 2017; Ju et al., 2018).

Figure 2 plots the difference in entity label distributions between the ACE 2005 data and our literary data: literature has a proportionally higher ratio of person and facility mentions, and much lower mentions for GPEs and organizations.

4.1 Prediction

To understand how this different distribution of entity types impacts the performance of models trained on these different sources, we evaluate the performance of a state-of-the-art model for nested

[2]Note we report F-score since we are measuring the agreement rates between annotators not only in their choice of labels (for which a categorical chance-corrected measure like Cohen's κ would be appropriate), but also the *spans* in text to which they apply.

Train → Test	Precision	Recall	F
ACE → ACE	75.3 [72.7–77.9]	63.3 [60.5–66.1]	68.8 [66.2–71.2]
ACE → Literature	57.8 [54.2–61.5]	37.7 [35.1–40.6]	45.7 [42.8–48.6]
Literature → Literature	75.1 [72.1–77.8]	62.6 [59.7–65.4]	68.3 [65.5–70.8]

Table 2: Performance on nested entity recognition using the layered BiLSTM-CRF of Ju et al. (2018) with different training → test combinations. All metrics are reported with 95% bootstrap confidence intervals.

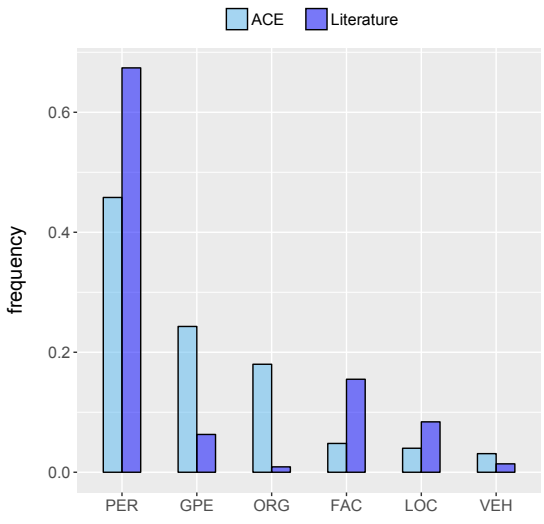


Figure 1: Distribution of entity types

entity recognition (Ju et al., 2018). We create training, development and test splits on the 100 literary books by stratifying at the document level, with 80 books in training, 10 books in development and 10 books in test.

To preprocess ACE, we tokenize and split sentences using the Stanford tokenizer (Manning et al., 2014), and create training, development and test partitions again stratified by document, so that sentences from the same document do not appear in both train and test. As above, we adapt the ACE annotations to our format by removing pronoun (PRO) and WH-question (WHQ) annotations and remove all weapon (WEA) labels, and consider only the subsets for broadcast conversation, broadcast news, newswire and weblogs. We present results with 95% confidence intervals using the bootstrap.

When trained on ACE and tested on ACE, the layered bidirectional LSTM-CRF of Ju et al. (2018) achieves an F-score of 68.8. When that same model (trained on ACE) is evaluated on our literature data, performance drops precipitously (23 absolute points in F-score). This alone—that cross-domain performance can be so strikingly worse—is a significant result, providing the first estimate of how performance degrades across these domains for this task.

However, when we train an identically parameterized model on the training partition of the literary data and evaluate it on the literary test partition, performance naturally improves substantially to an F-score of 68.3. As table 2 shows, performance improves dramatically for nearly all entity classes; the classes with the most statistically significant improvement are PER and FAC—both of which improve by 20 absolute points.

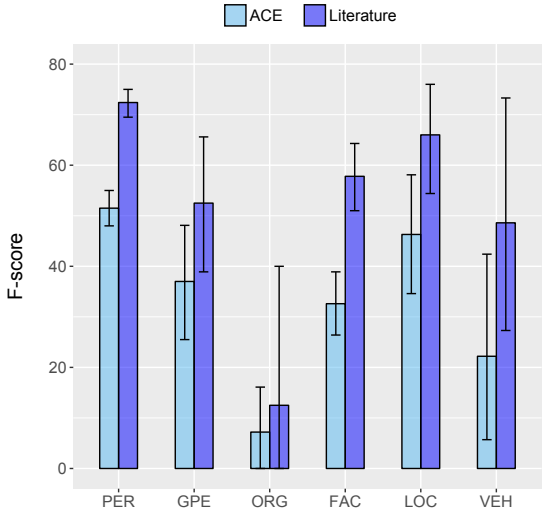


Figure 2: F-scores with 95% bootstrap confidence intervals by entity type when evaluating on literature test data with different training sources.

4.2 Analysis

To better understand the ways in which a model trained on ACE data differs in its predictions from an identically parameterized model trained on literary data, we used the two models described above to generate predictions for nested entities in a random sample of 1,000 full-text books from Project Gutenberg not in our training, development or test data (a total of 78M tokens). We

then analyzed a simple difference in frequencies between the predictions of the two models on that same data; for a given entity e (e.g., *the boy*) and category t (e.g., PER), we calculate the frequency f as the number of times e was tagged by each model as t, and measure the difference:

$$f_{LIT}(e,t) - f_{ACE}(e,t)$$

The ten terms with the strongest positive difference in frequencies for the PER class—those phrases that are found significantly more often in a model trained on literary data than a model trained on ACE—are *Mrs., Miss, Lady, Aunt, Sir, Captain, no one, Mr, Madame* and *nobody*, suggesting a potential gender bias in the predictions; indeed, while ACE 2005 contains 47 instances of *Mr.*, it contains no mentions of *Mrs.* or honorific *Miss* (and only three instances of *Ms.*). While other work has demonstrated the gender bias present in word embeddings (Bolukbasi et al., 2016; Caliskan et al., 2017; Zhao et al., 2019) and in such NLP tasks as coreference resolution (Rudinger et al., 2018; Zhao et al., 2018), sentiment analysis (Kiritchenko and Mohammad, 2018), and speech recognition (Tatman, 2017), we can investigate the same phenomenon here: does a model trained on ACE result in a disparate impact in its performance recognizing men and women entities in text?

To answer this question, we annotate the gender for all PER entities in the literary test data (a total of 969 entities) and measure the recall of each model as a function of the gender of the true entity (measuring, for example, how many women in the gold literary data each model was able to identify, and how many men).

Training	Women	Men	Diff	p
ACE	38.0	49.6	−11.6	0.0009
Literature	69.3	68.2	1.1	0.7459

Table 3: Recall on literary test data by the gender of PER entity for models trained on ACE and literary data.

Table 4 lists these results: while a model trained on literary data recognizes male and female entities at roughly equal rates, ACE data shows a strong disparate performance, with female entities recognized at a rate over 11 points worse than male entities. This difference is significant at $p < 0.001$ under a permutation test (randomly shuffling the gender labels assigned to entities to generate a non-parametric null distribution, with 100,000 permutations).

If we remove the obvious entities from the gold data that begin with *Mrs.* and *Miss* (the honorifics that are rarely attested in ACE) along with those that begin with *Mr.*, we still see a sizable disparity in performance, suggesting that this result is more pervasive than the simple absence of those honorifics from the training data.

Training	Women	Men	Diff	p
ACE	40.4	48.3	−7.9	0.0358
Literature	63.7	67.1	−3.4	0.3542

Table 4: Recall on literary test data by the gender of PER entity for models trained on ACE and literary data, excluding all gold entities beginning with *Mrs.*, *Miss* and *Mr.*

5 Conclusion

We present in this work a new dataset of nested entity annotations for literature; such data allows us to measure the performance of existing NER systems when evaluated on literary data, train new models optimized for literature as a domain, and explore the stylistic differences in entity attention that help define literature as a genre. In addition to helping advance the state-of-the-art in NLP for literary texts, we provide this dataset to advance modeling for entity recognition generally; as Søgaard (2013) argues, the robustness of performance improvements for methods in NLP is best estimated by performance across a range of domains; we would expect a robust model that shows improvement on news entities in ACE and proteins in GENIA to show improvements on recognizing characters and settings in literature as well.

All data is freely available for public use under a Creative Commons Sharealike license and is available at: https://github.com/dbamman/litbank; code to support this work can be found at: https://github.com/dbamman/NAACL2019-literary-entities.

Acknowledgments

We thank the anonymous reviewers, Matt Sims and Jon Gillick for their valuable feedback. The research reported in this article was supported by an Amazon Research Award, a grant from the Digital Humanities at Berkeley initiative and by resources provided by NVIDIA.

References

Isabelle Augenstein, Leon Derczynski, and Kalina Bontcheva. 2017. Generalisation in named entity recognition. *Comput. Speech Lang.*, 44(C):61–83.

David Bamman, Ted Underwood, and Noah A. Smith. 2014. A Bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 370–379, Baltimore, Maryland. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 4356–4364, USA. Curran Associates Inc.

Julian Brooke, Adam Hammond, and Timothy Baldwin. 2016. Bootstrapped text-level named entity recognition for literature. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 344–350. Association for Computational Linguistics.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Snigdha Chaturvedi, Mohit Iyyer, and Hal Daumé III. 2017. Unsupervised learning of evolving relationships between literary characters. In *Association for the Advancement of Artificial Intelligence*.

Grant DeLozier, Ben Wing, Jason Baldridge, and Scott Nesbit. 2016. Creating a novel geolocation corpus from historical texts. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 188–198. Association for Computational Linguistics.

Elizabeth F. Evans and Matthew Wilkens. 2018. Nation, ethnicity, and the geography of British fiction, 1880–1940. *Cultural Analytics*.

Jenny Rose Finkel and Christopher D Manning. 2009. Nested named entity recognition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 141–150. Association for Computational Linguistics.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: the 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, NAACL-Short '06, pages 57–60, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III. 2016. Feuding families and former friends: Unsupervised learning for dynamic fictional relationships. In *North American Association for Computational Linguistics*.

Meizhi Ju, Makoto Miwa, and Sophia Ananiadou. 2018. A neural layered model for nested named entity recognition. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 1446–1459.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53. Association for Computational Linguistics.

LDC. 2005. Ace (automatic content extraction) English annotation guidelines for entities. `https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-entities-guidelines-v5.6.6.pdf`.

Wei Lu and Dan Roth. 2015. Joint mention extraction and classification with mention hypergraphs. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 857–867.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Aldrian Obaja Muis and Wei Lu. 2017. Labeling gaps between words: Recognizing overlapping mentions with mention separators. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2608–2618. Association for Computational Linguistics.

Paul Rayson, Alex Reinhold, James Butler, Chris Donaldson, Ian Gregory, and Joanna Taylor. 2017. A deeply annotated testbed for geographical text analysis: The corpus of lake district writing. In *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities*, pages 9–15. ACM.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14. Association for Computational Linguistics.

Anders Søgaard. 2013. Estimating effect size across datasets. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 607–611. Association for Computational Linguistics.

Rachael Tatman. 2017. Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain. Association for Computational Linguistics.

Dennis Yi Tenen. 2018. Toward a computational archaeology of fictional space. *New Literary History*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL-2003*, pages 142–147. Edmonton, Canada.

Ted Underwood, David Bamman, and Sabrina Lee. 2018. The transformation of gender in English-language fiction. *Cultural Analytics*.

Hardik Vala, David Jurgens, Andrew Piper, and Derek Ruths. 2015. Mr. Bennet, his coachman, and the Archbishop walk into a bar but only one of them gets recognized: On the difficulty of detecting characters in literary texts. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 769–774, Lisbon, Portugal. Association for Computational Linguistics.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. ACE 2005 multilingual training corpus. LDC.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *NAACL*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20. Association for Computational Linguistics.