Literary Entity Recognition

Nate Stringham

Task: Identifying Literary Entities

The task is to identify **6** types of entities found in ACE 2005

Person (PER) - proper names, common entities, sets of people

- Ishmael, my mother, the Van Rensselaers,

Facility (FAC) - a structure designed for human habitation, storage, transportation infrastructure, maintained outdoor spaces

- the library, the garden, the house

Geo-political entity (GPE) - must have population, government, physical location, political boundaries

- The village, London, England

Task: Identifying Literary Entities

Location (LOC) - must be physical, but no political organization

- Sea, several fields beyond, the Thames

Vehicle (VEH) - entity designed to transport an object between locations (old books -> older kinds of vehicles)

- The ship, a baggage car, The Nellie, a cruising yawl

Organization (ORG) - a formal association between people (rarest category for Lit)

- the army, The Swedish Pathological Sociey, a leaden-headed old corporation

The LitBank Data Set

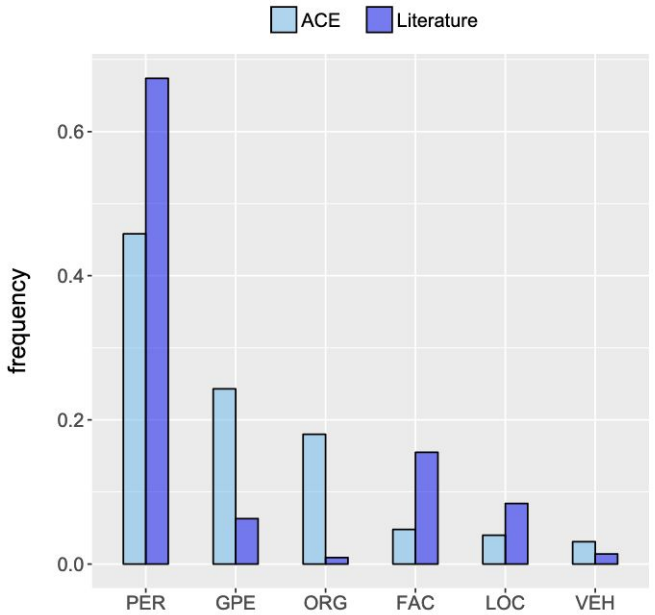
- 100 works of English Language Fiction selected from the Project Gutenberg Corpus annotated with BIO tags (David Bamman, Sejal Popat and Sheng Shen (2019))
- Includes annotations for 6 Entity Types from ACE 2005
- Includes event and coref and quotation annotations

Author	**Title**
Alcott, Louisa May	Little Women
Dickens, Charles	Bleak House
Grey, Zane	Desert Gold
Joyce, James	Ulysses
Stoker, Bram	Dracula

What is Different About Literary Entities?

Easiest explanation is to imagine reading your favorite book compared to a news article.

- Different distribution across entity types
  - Focus on people (PER) and descriptions of setting (FAC/LOC)
- Rich descriptions ⇨ long entities
- Common nominals as well as named entities
- Nested Entity Structure
- Figurative Language



Bamman, David, Sejal Popat, and Sheng Shen. "An annotated dataset of literary entities." *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019.

Nested Entities

LEVEL 0 and 1:

“The short-hand writers, the reporters of the court, and the reporters of the newspapers …”

- PER, ORG

“Fog up the river, where it flows among green aits and meadows …”

- LOC, LOC

Figurative Language

Personification - when main character is a personified animal annotate PER

- “... a wise old horse …” - Black Beauty
- “... a huge St. Bernard …” - Call of the Wild

Metonymy - annotate examples as the evoked entity class

- “... the kitchen was outraged…” -> kitchen annotated as PER

Dataset Statistics

First ~ 2,000 tokens from 100 books = 210,532 total tokens

8562 total sentences (~24 tokens per sentence)

13,912 total entity annotations

We split the dataset into three parts (train, dev, test) by randomly assigning books to each split.

Resulting dataset is 80-10-10 train, dev, test split by books (books aren't divided between splits)

Entity Type breakdown in Gold

	PER	FAC	GPE	LOC	VEH	ORG	Total
Train:	7598	1723	684	924	143	81	11153
Dev:	883	170	71	89	18	20	1251
Test:	902	261	123	157	36	29	1508

Evaluation

Model takes in a sentence and outputs an entity tuple of the form:

(entity, label, [start, end])

Where **entity** is the extracted string, **label** is the predicted entity type, and **[start, end]** is the character indices corresponding to the start and end of the entity.

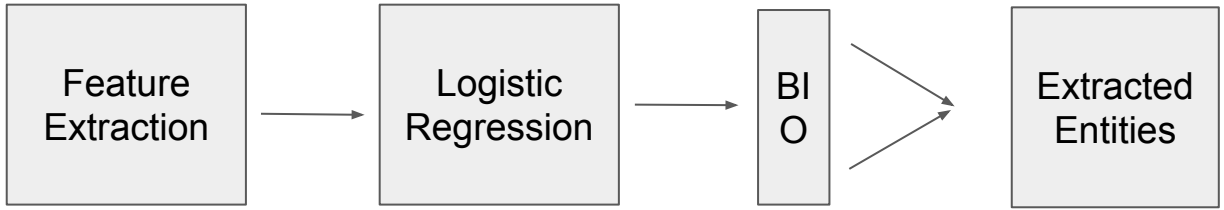
We compute these tuples for the gold data and then compare them to the predicted tuples.

- Compute Precision, Recall, F-Score
- macro and micro averaged

Macro average weights classes evenly while micro weights instances evenly.

Phase 1: Logistic Regression

Motivation: simple baseline for sequence labeling



Features:

- Word, Word-1, Word+1
- POS, POS-1, POS +1
- Notably forgot to include capitalization as a feature

Phase 1 Results

Table 1: ALL Nesting Levels - 100% of the entities in gold

	Precision			Recall			F-Score		
	tr	dev	test	tr	dev	test	tr	dev	test
PER	0.39	0.34	0.28	0.54	0.40	0.32	0.46	0.37	0.30
FAC	0.31	0.32	0.25	0.35	0.27	0.18	0.33	0.29	0.21
GPE	0.60	0.45	0.48	0.64	0.38	0.39	0.62	0.41	0.43
LOC	0.29	0.25	0.25	0.32	0.20	0.19	0.31	0.22	0.21
VEH	0.27	0.00	0.52	0.18	0.00	0.25	0.22	0.00	0.33
ORG	0.15	0.00	0.00	0.14	0.00	0.00	0.15	0.00	0.00
micro avg	0.38	0.34	0.29	0.49	0.35	0.28	0.43	0.34	0.28
macro avg	0.34	0.23	0.30	0.36	0.20	0.22	0.35	0.21	0.25

Phase 1 Results

Table 2: Nesting Level 0 (86.2% of the entities in gold)

	Precision			Recall			F-Score		
	tr	dev	test	tr	dev	test	tr	dev	test
PER	0.39	0.34	0.28	0.64	0.47	0.37	0.48	0.39	0.32
FAC	0.31	0.32	0.25	0.41	0.31	0.23	0.35	0.32	0.24
GPE	0.60	0.45	0.48	0.67	0.41	0.40	0.63	0.43	0.44
LOC	0.29	0.25	0.25	0.36	0.23	0.25	0.32	0.24	0.25
VEH	0.27	0.00	0.52	0.20	0.00	0.25	0.23	0.00	0.34
ORG	0.15	0.00	0.00	0.16	0.00	0.00	0.15	0.00	0.00
micro avg	0.38	0.33	0.29	0.57	0.41	0.33	0.46	0.37	0.31
macro avg	0.33	0.23	0.30	0.41	0.24	0.25	0.36	0.23	0.26

Phase 1 Results

Table 3: This is the table for nesting level 1 (12.5% of the entities in gold)

	Precision			Recall			F-Score		
	tr	dev	test	tr	dev	test	tr	dev	test
PER	0.0006	0.0028	0.006	0.0067	0.0230	0.0472	0.0012	0.0051	0.0106
FAC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GPE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LOC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VEH	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ORG	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
micro avg	0.0004	0.0022	0.0041	0.0051	0.0178	0.0283	0.0008	0.0040	0.0073
macro avg	0.0001	0.0004	0.001	0.0011	0.0038	0.0078	0.0002	0.0008	0.0017

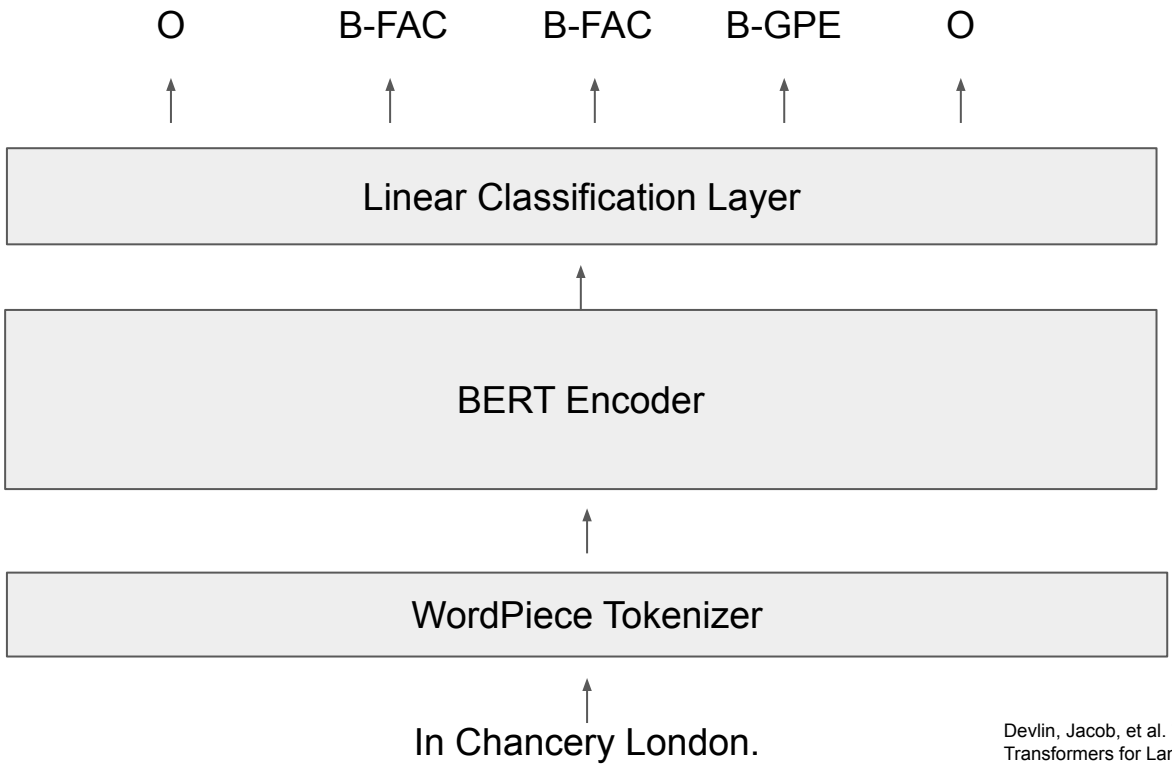
Phase 1 Takeaways

- Nested entities are tricky, but < 14% of entities
  - Don’t get fixated on nested, focus on non-nested
- GPE, VEH, PER are easier, ORG is the hardest
- Naive feature set w/logistic regression is enough to get some performance for entity recognition.

Motivation and Next Steps:

- Try a more complex model that leverages pre-trained information

Phase 2: Fine-tuned BERT



12 Layers
12 Attention Heads
110M params

Devlin, Jacob, et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." (2019).

Phase 2 Results

	Precision			Recall			F-Score		
	tr	dev	test	tr	dev	test	tr	dev	test
PER	0.80	0.67	0.64	0.71	0.65	0.60	0.76	0.66	0.62
FAC	0.66	0.53	0.46	0.63	0.49	0.44	0.65	0.51	0.45
GPE	0.77	0.41	0.56	0.81	0.62	0.55	0.78	0.49	0.56
LOC	0.64	0.33	0.43	0.62	0.28	0.38	0.63	0.30	0.41
VEH	0.68	0.47	0.57	0.69	0.44	0.72	0.68	0.46	0.63
ORG	0.31	0.45	0.00	0.25	0.25	0.00	0.28	0.32	0.00
micro avg	0.76	0.61	0.57	0.70	0.59	0.54	0.73	0.60	0.56
macro avg	0.62	0.46	0.45	0.64	0.48	0.44	0.63	0.46	0.44

Table 2: Results from evaluating the `LER-bert` on ALL Nesting Levels - 100% of the entities in gold.

Phase 2 vs Phase 1

	Precision		Recall		F-Score	
	LER-base	LER-bert	LER-base	LER-bert	LER-base	LER-bert
PER	0.28	0.64	0.32	0.60	0.30	0.62
FAC	0.25	0.46	0.18	0.44	0.21	0.45
GPE	0.48	0.56	0.39	0.55	0.43	0.56
LOC	0.25	0.43	0.19	0.38	0.21	0.41
VEH	0.52	0.57	0.25	0.72	0.33	0.63
ORG	0.00	0.00	0.00	0.00	0.00	0.00
micro avg	0.38	0.57	0.28	0.54	0.28	0.56
macro avg	0.34	0.45	0.22	0.44	0.25	0.44

Table 1: System results on the test set containing entities from ALL nesting levels. **LER-base** is the logistic regression token classifier from Phase1 and **LER-bert** is the fine-tuned model introduced in Phase2.

Phase 2 Takeaways

- Using rich pre-trained representations such as those found in BERT really improves performance
- Micro-averaging seems to be inflated because PER type has so many instances (~68% of entities in train) and our model is relatively good at that category.
- Still not getting ORG entities. (< 1 % of entities in train)

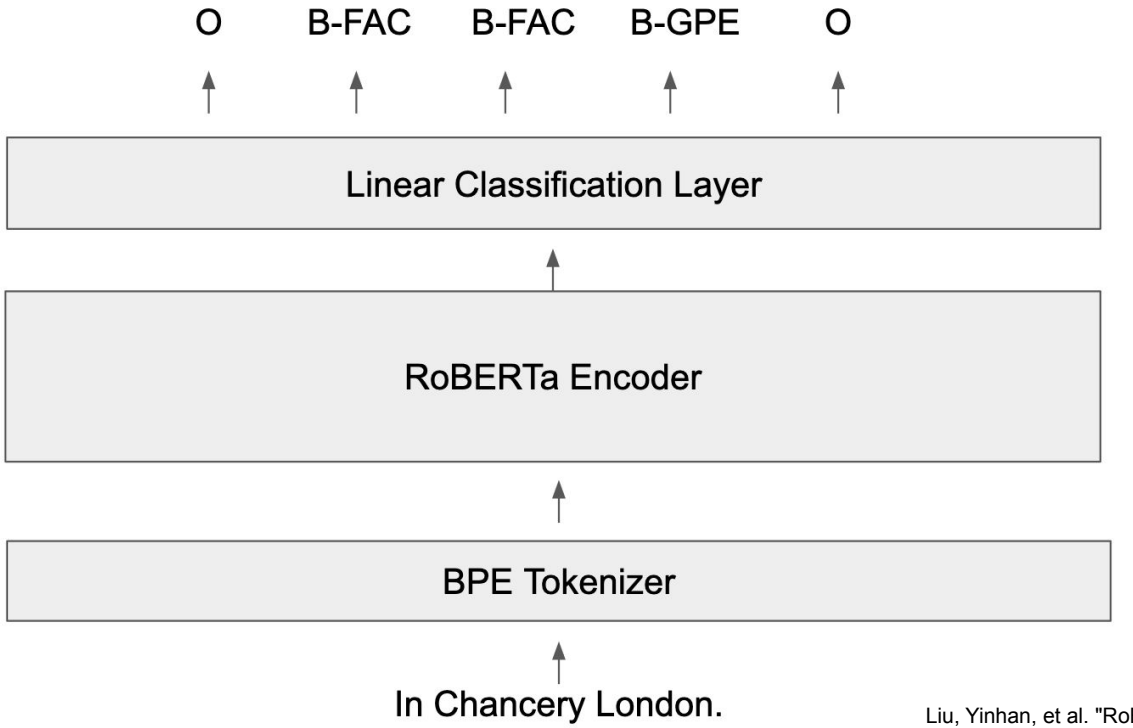
Motivation/Next Steps:

- RoBERTA = Robust BERT. Supposedly a more carefully pre-trained BERT
- Look into some of the errors the model is making

Differences Between BERT and RoBERTa

- Dynamic Masking
  - Generate mask pattern each time a sequence is fed in instead of statically for all sequences
- Full training sentences without Next Sentence Prediction
- Large mini-batches
- Byte Pair Encoding Tokenizer instead of WordPiece

Phase 3: Fine-tuned RoBERTa + Error Analysis



Liu, Yinhan, et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." *arXiv preprint arXiv:1907.11692* (2019).

Phase 3 vs Phase 2

	Precision		Recall		F-Score	
	LER-bert	LER-roberta	LER-bert	LER-roberta	LER-bert	LER-roberta
PER	0.74	0.80	0.66	0.71	0.69	0.75
FAC	0.57	0.69	0.48	0.57	0.52	0.62
GPE	0.69	0.77	0.60	0.69	0.64	0.73
LOC	0.57	0.59	0.46	0.48	0.51	0.53
VEH	0.69	0.77	0.81	0.83	0.74	0.80
ORG	0.20	0.22	0.10	0.07	0.14	0.11
micro avg	0.68	0.75	0.59	0.65	0.63	0.70
macro avg	0.52	0.56	0.57	0.64	0.54	0.59

Table 1: System results on the test set containing entities from ALL nesting levels. `LER-bert` is the bert-based token classifier from Phase2 and `LER-roberta` is the model introduced in Phase3.

Phase 3 vs Phase 2 (non-nested entities)

	Precision		Recall		F-Score	
	LER-bert	LER-roberta	LER-bert	LER-roberta	LER-bert	LER-roberta
PER	0.73	0.79	0.78	0.83	0.75	0.81
FAC	0.56	0.68	0.59	0.71	0.57	0.70
GPE	0.68	0.76	0.61	0.71	0.65	0.73
LOC	0.56	0.59	0.59	0.63	0.57	0.61
VEH	0.69	0.77	0.83	0.86	0.75	0.81
ORG	0.20	0.22	0.14	0.10	0.17	0.13
micro avg	0.67	0.74	0.70	0.77	0.69	0.76
macro avg	0.59	0.64	0.57	0.64	0.58	0.63

Table 2: System results on the test set containing entities from 0^{th} nesting level. `LER-bert` is the bert-based token classifier from Phase2 and `LER-roberta` is the model introduced in Phase3.

Example Errors

Entity Missed

PER

- “... to be surgeon to the Swallow, Captain Abraham Pannel
- “… bibliophile…”
- “... not a portrait of a man, but a distinct and dynamic personality…”

FAC

- “... the narrow street in Soho”
- “... the British crown had given to one of theses forest-fastnesses the name of William Henry.”

GPE

- “A VOYAGE TO LILLIPUT…”
- “He generally arrived…from the Continent …”

Example Errors

Entity Missed

LOC

- “... the latitude of 30 degrees 2 minutes south …”
- “ … its banks …”

VEH

- “… a Greenland whaler …”
- “... the prows of racing shells…”

ORG

- “... the remnants of the British army …”
- “... my business began to fail…”
- “... college …”

Lessons Learned

- Even though we can get high scores for NER
  - there is a long tail of rare entities that are hard to capture
- Still struggle with categories that have a small amount of entity annotations (ORG)
- Identifying nested entities can make the task even more challenging.
- Different domains can have both different entity types AND different distribution of those types.
- RoBERTa actually worked better than BERT (I was skeptical)
- Do Error Analysis!
  - Helps you know how your model is actually doing (wish I would have done more of this along the way)
  - Helps you understand the data better
  - Can find bugs in your evaluation script (like I did)

External Resources

NTLK - POS tagging and tokenization for phase 1 model

scikit-learn - logistic regression model training

Hugging Face - fine-tuned roberta-base and bert-base-cased models with token classification head

Demo

https://huggingface.co/nates/LER-roberta