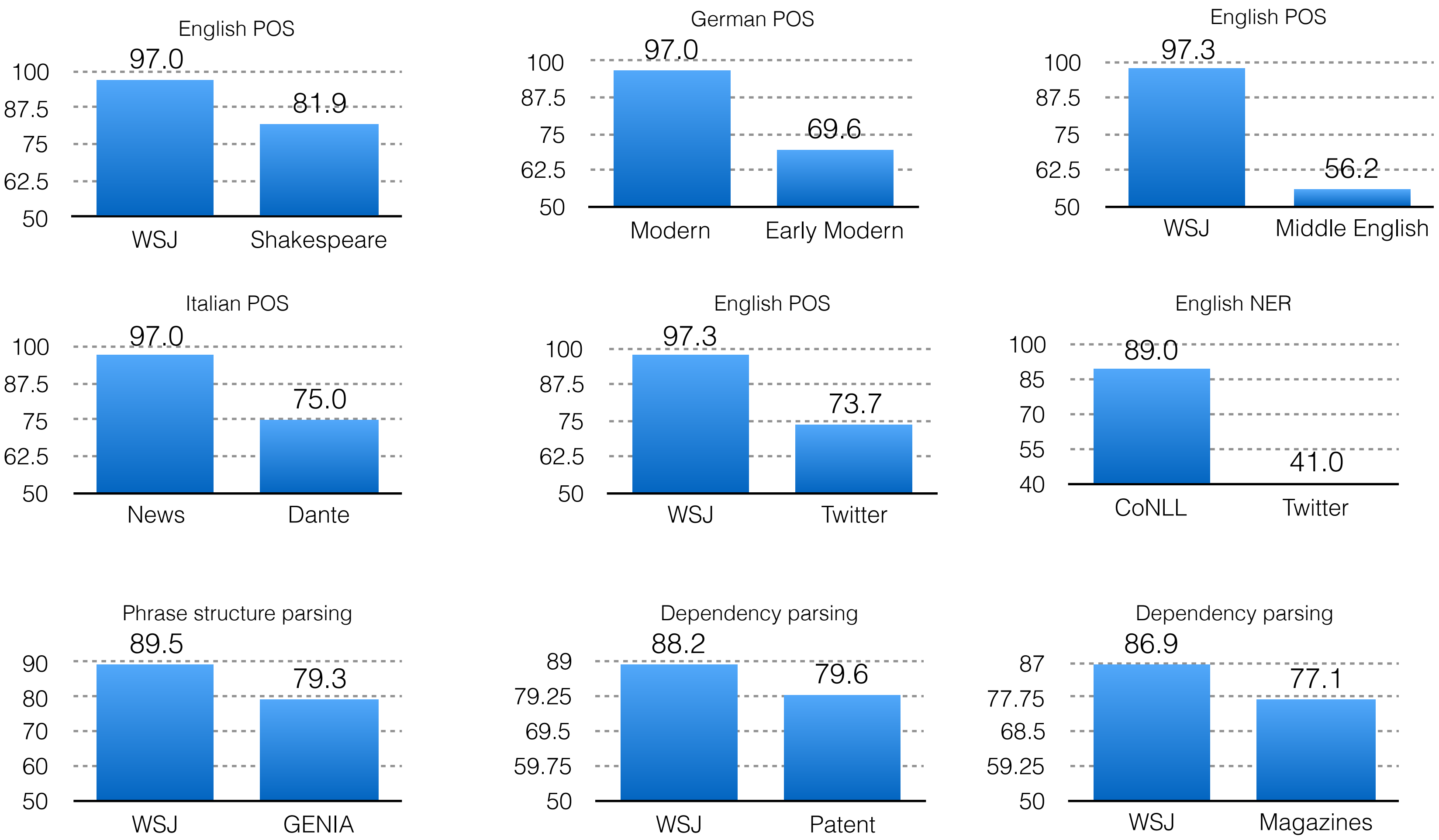
Domain-specific NLP

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

NLP Pipeline

NLP Task	Accuracy (English)
Tokenization	100%
Part-of-speech tagging	98.0% [Bohnet et al. 2018]
Named entity recognition	93.1 [Akbik et al. 2018]
Syntactic parsing	95.1 F [Kitaev and Klein 2018]
Coreference resolution	73.0 F [Lee et al. 2018]

Data in NLP


English POS
97.0
81.9
WSJ
Shakespeare
German POS
97.0
69.6
Modern
Early Modern
English POS
97.3
56.2
WSJ
Middle English
Italian POS
97.0
75.0
News
Dante
English POS
97.3
73.7
WSJ
Twitter
English NER
89.0
41.0
CoNLL
Twitter
Phrase structure parsing
89.5
79.3
WSJ
GENIA
Dependency parsing
88.2
79.6
WSJ
Patent
Dependency parsing
86.9
77.1
WSJ
Magazines

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. 210,532 tokens from 100 different novels, annotated for:

  - Entities (person/place, etc.)
  - Events
  - Conference

Named entity recognition

[tim cook]$_{\text{PER}}$ is the ceo of [apple]$_{\text{ORG}}$

- Identifying spans of text that correspond to typed entities that are proper names.

Named entity recognition

Type	Tag	Sample Categories	Example sentences
People	PER	people, characters	**Turing** is a giant of computer science.
Organization	ORG	companies, sports teams	The **IPCC** warned about the cyclone.
Location	LOC	regions, mountains, seas	The **Mt. Sanitas** loop is in **Sunshine Canyon**.
Geo-Political Entity	GPE	countries, states, provinces	**Palo Alto** is raising the fees for parking.
Facility	FAC	bridges, buildings, airports	Consider the **Golden Gate Bridge**.
Vehicles	VEH	planes, trains, automobiles	It was a classic **Ford Falcon**.

Figure 17.1 A list of generic named entity types with the kinds of entities they refer to.

ACE NER categories

Named entity recognition

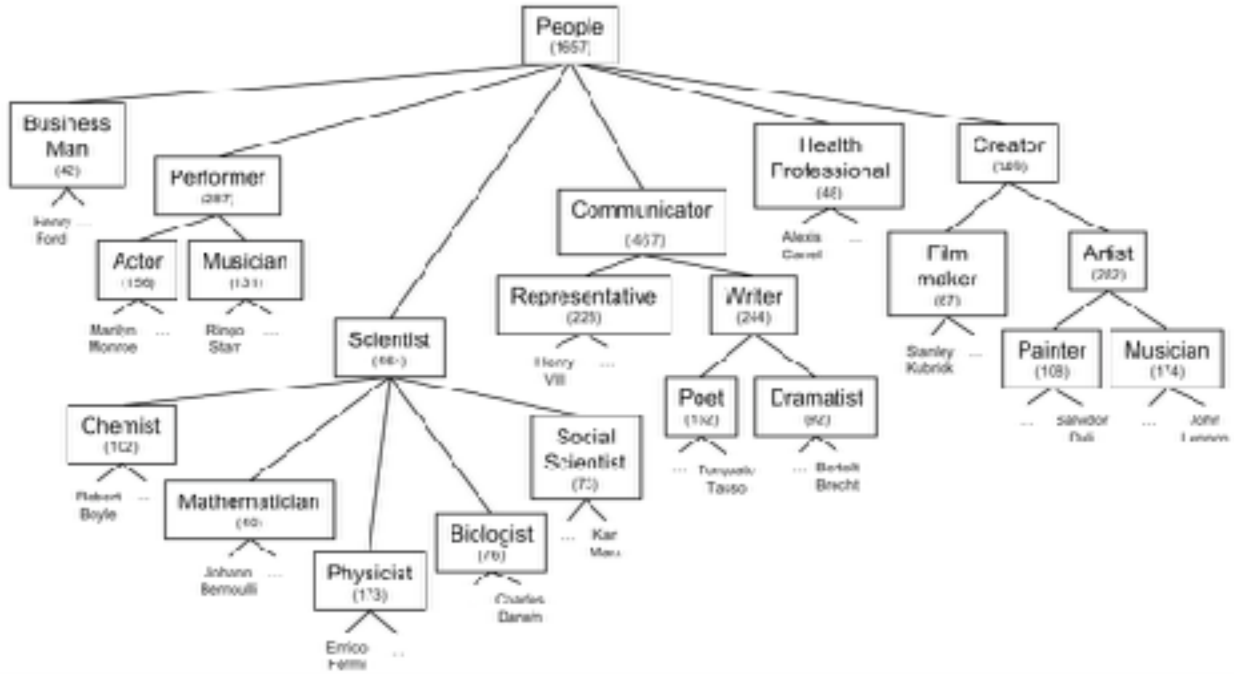
- GENIA corpus of MEDLINE abstracts (biomedical)

We have shown that [interleukin-1]PROTEIN ([IL-1]PROTEIN) and [IL-2]PROTEIN control [IL-2 receptor alpha (IL-2R alpha) gene]DNA transcription in [CD4-CD8- murine T lymphocyte precursors]CELL LINE

protein
cell line
cell type
DNA
RNA

http://www.aclweb.org/anthology/W04-1213

Fine-grained NER



Giuliano and Gliozzo (2008)

Entity recognition

Person	… named after [the daughter of a Mattel co-founder] …
Organization	[The Russian navy] said the submarine was equipped with 24 missiles
Location	Fresh snow across [the upper Midwest] on Monday, closing schools
GPE	The [Russian] navy said the submarine was equipped with 24 missiles
Facility	Fresh snow across the upper Midwest on Monday, closing [schools]
Vehicle	The Russian navy said [the submarine] was equipped with 24 missiles
Weapon	The Russian navy said the submarine was equipped with [24 missiles]

ACE entity categories
https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-entities-guidelines-v6.6.pdf

Named entity recognition

- Most named entity recognition datasets have flat structure (i.e., non-hierarchical labels).

✔ [The University of California]$_{ORG}$

✖ [The University of [California]$_{GPE}$]$_{ORG}$

- Mostly fine for named entities, but more problematic for general entities:

[[John]$_{PER}$'s mother]$_{PER}$ said …

Nested NER

named	after	the	daughter	of	a	Mattel	co-founder
						B-ORG	
					B-PER	I-PER	I-PER
		B-PER	I-PER	I-PER	I-PER	I-PER	I-PER

Sequence labeling

$$x = \{x_1, \ldots, x_n\}$$

$$y = \{y_1, \ldots, y_n\}$$

- Training data: for a set of inputs x with n sequential time steps, one corresponding label y_i for each x_i
- Model correlations in the labels y.

NER sequence labeling

identity of w_i, identity of neighboring words
embeddings for w_i, embeddings for neighboring words
part of speech of w_i, part of speech of neighboring words
base-phrase syntactic chunk label of w_i and neighboring words
presence of w_i in a **gazetteer**
w_i contains a particular prefix (from all prefixes of length ≤ 4)
w_i contains a particular suffix (from all suffixes of length ≤ 4)
w_i is all upper case
word shape of w_i, word shape of neighboring words
short word shape of w_i, short word shape of neighboring words
presence of hyphen

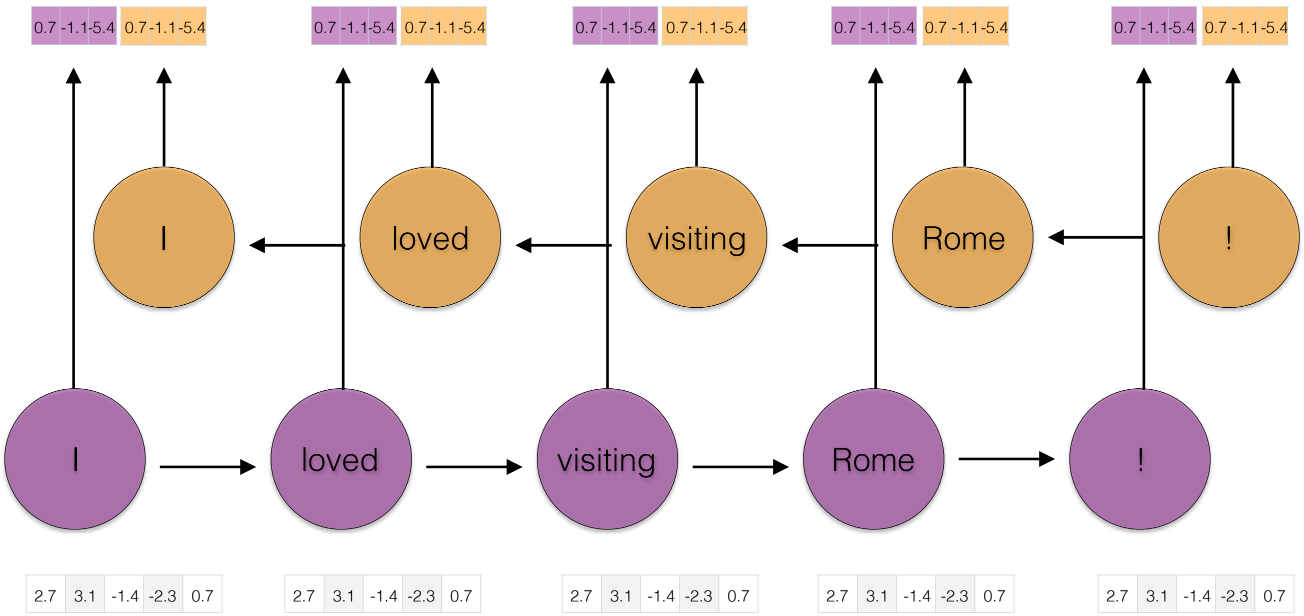
Figure 17.5 Typical features for a feature-based NER system.

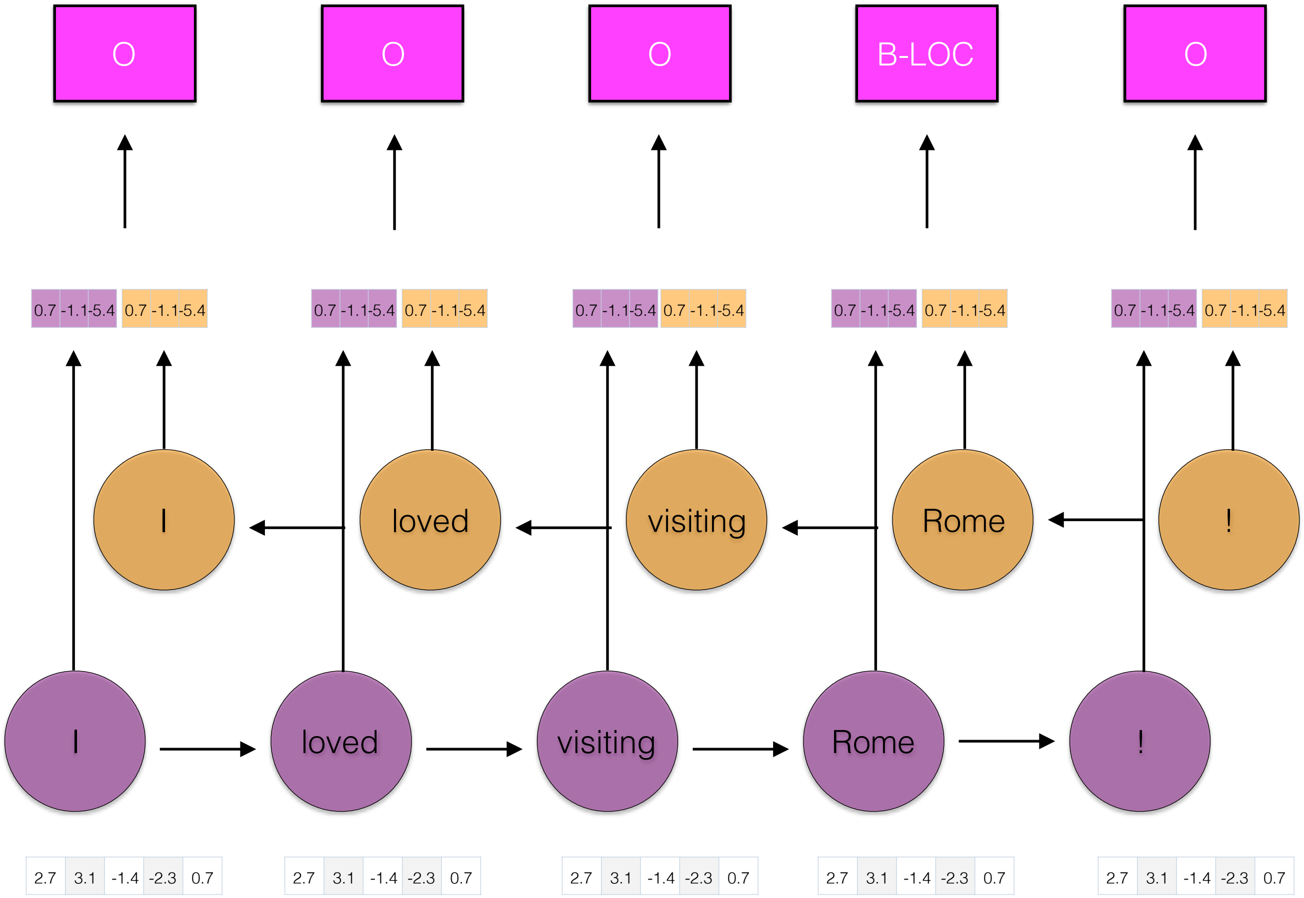
Gazetteers

- List of place names; more generally, list of names of some typed category
- GeoNames (GEO), US SSN (PER), Getty Thesaurus of Geographic Placenames, Getty Thesaurus of Art and Architecture

Cliff
Bun Cranncha
Dromore West
Dromore
Youghal Harbour
Youghal Bay
Youghal
Eochaill
Yellow River
Yellow Furze
Woodville
Wood View
Woodtown House
Woodstown
Woodstock House
Woodsgift House
Woodrooff House
Woodpark
Woodmount
Wood Lodge
Woodlawn Station
Woodlawn
Woodlands Station
Woodhouse
Wood Hill
Woodfort
Woodford River
Woodford
Woodfield House
Woodenbridge Junction Station
Woodenbridge
Woodbrook House
Woodbrook
Woodbine Hill
Wingfield House
Windy Harbour
Windy Gap

BiLSTM for sequence tagging





O
O
O
B-LOC
O
0.7 -1.1 -5.4 0.7 -1.1 -5.4
0.7 -1.1 -5.4 0.7 -1.1 -5.4
0.7 -1.1 -5.4 0.7 -1.1 -5.4
0.7 -1.1 -5.4 0.7 -1.1 -5.4
0.7 -1.1 -5.4 0.7 -1.1 -5.4
I
loved
visiting
Rome
!
I
loved
visiting
Rome
!
2.7 3.1 -1.4 -2.3 0.7
2.7 3.1 -1.4 -2.3 0.7
2.7 3.1 -1.4 -2.3 0.7
2.7 3.1 -1.4 -2.3 0.7
2.7 3.1 -1.4 -2.3 0.7

BiLSTM for sequence tagging

	POS		NER					
	Dev	Test	Dev			Test		
Model	Acc.	Acc.	Prec.	Recall	F1	Prec.	Recall	F1
BRNN	96.56	96.76	92.04	89.13	90.56	87.05	83.88	85.44
BLSTM	96.88	96.93	92.31	90.85	91.57	87.77	86.23	87.00
BLSTM-CNN	97.34	97.33	92.52	93.64	93.07	88.53	90.21	89.36
BRNN-CNN-CRF	97.46	97.55	94.85	94.63	94.74	91.35	91.06	91.21

Ma and Hovy (2016), "End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF"

Literary entities

Most work in NLP focuses on *named* entity recognition — mentions of specific categories (person, place, organization) that are explicitly named.

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*

Entity recognition

- Mr. Knightley
- a sensible man about seven or eight-and-thirty
- a very old and intimate friend of the family
- the family
- Isabella
- Isabella’s husband
- 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*

Entity recognition

- Mr. Knightley
- a sensible man about seven or eight-and-thirty
- a very old and intimate friend of the family
- the elder brother of Isabella's husband

- the family

- Isabella

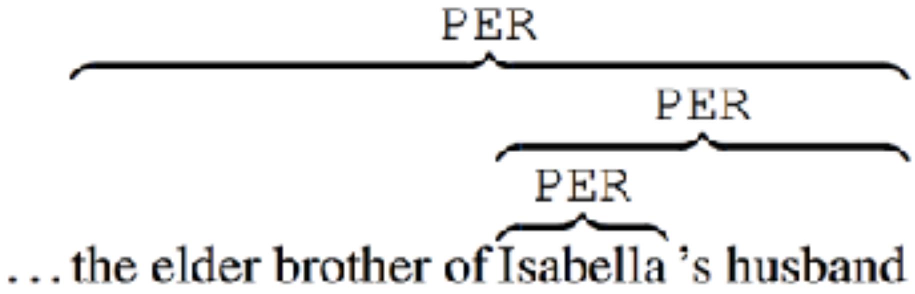
- Isabella's husband

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*

Nested entity recognition

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



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.

PER PER

John is a doctor

PER PER ???

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*

Data

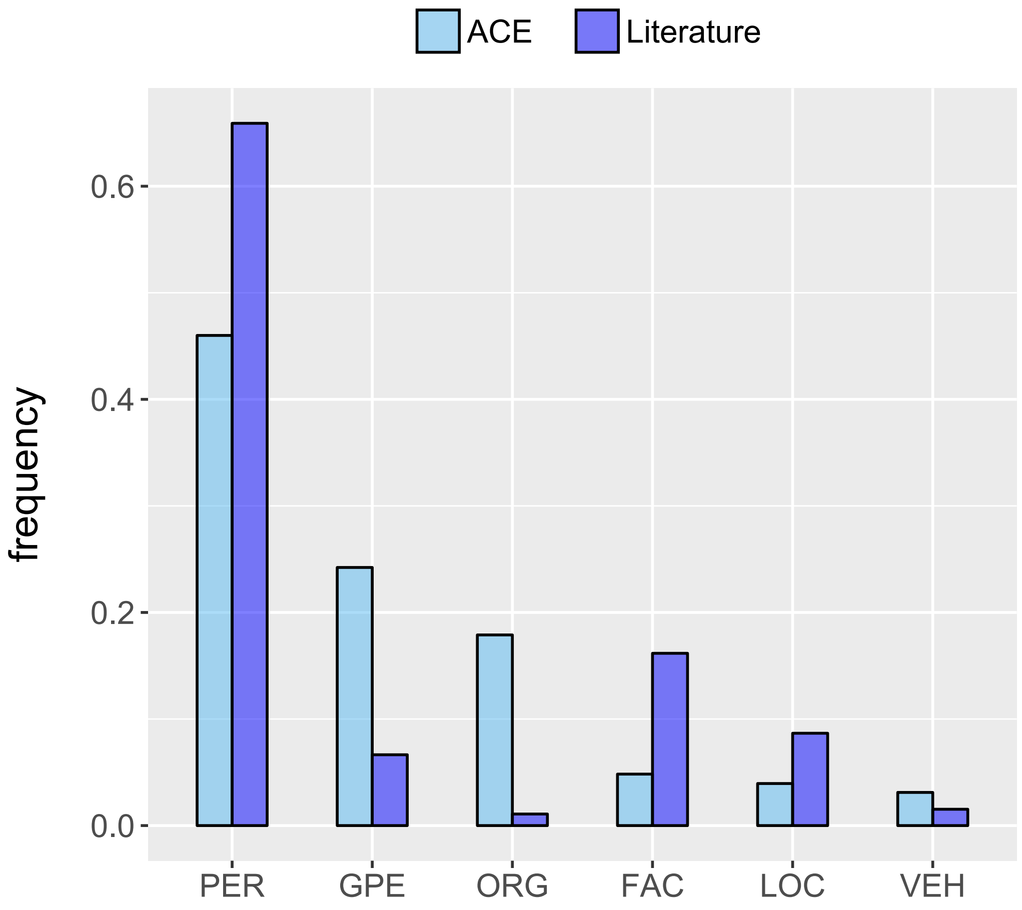
Cat	Count	Examples
PER	9,383	my mother, Jarndyce, the doctor, a fool, his companion
FAC	2,154	the house, the room, the gardne, the drawing-room, the library
LOC	1,170	the sea, the river, the country, the woods, the forest
GPE	878	London, England, the town, New York, the village
VEH	197	the ship, the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

Prediction

How well can find these entity mentions in text as a function of the training domain?

Data

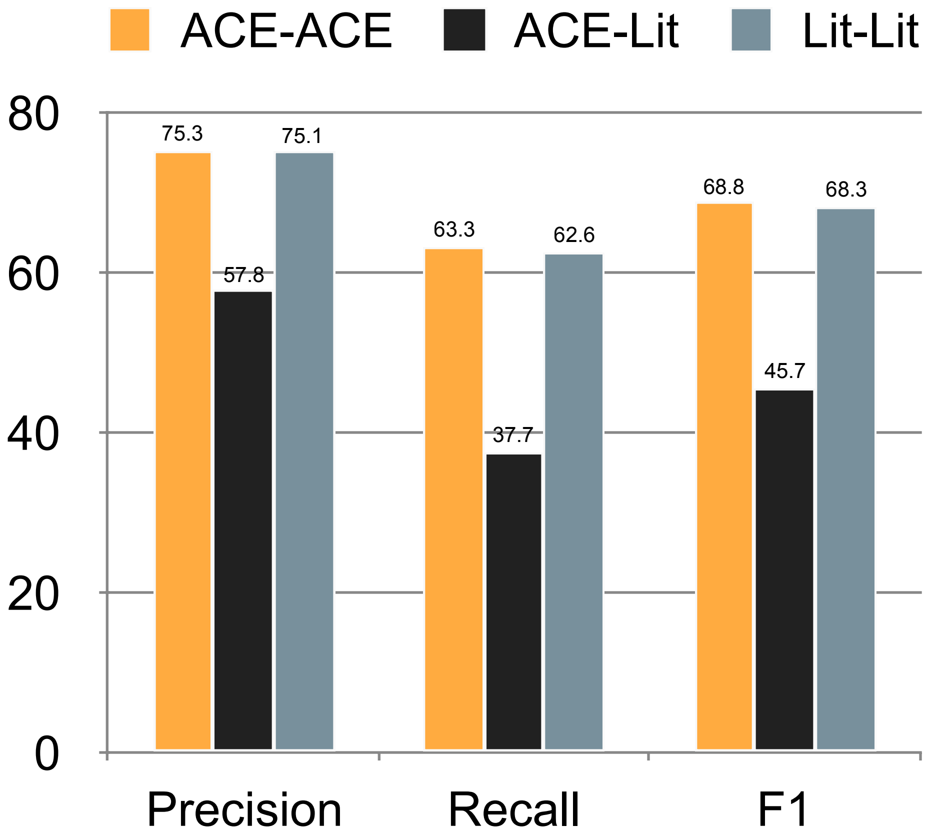
- ACE (2005) data from newswire, broadcast news, broadcast conversation, weblogs



Prediction

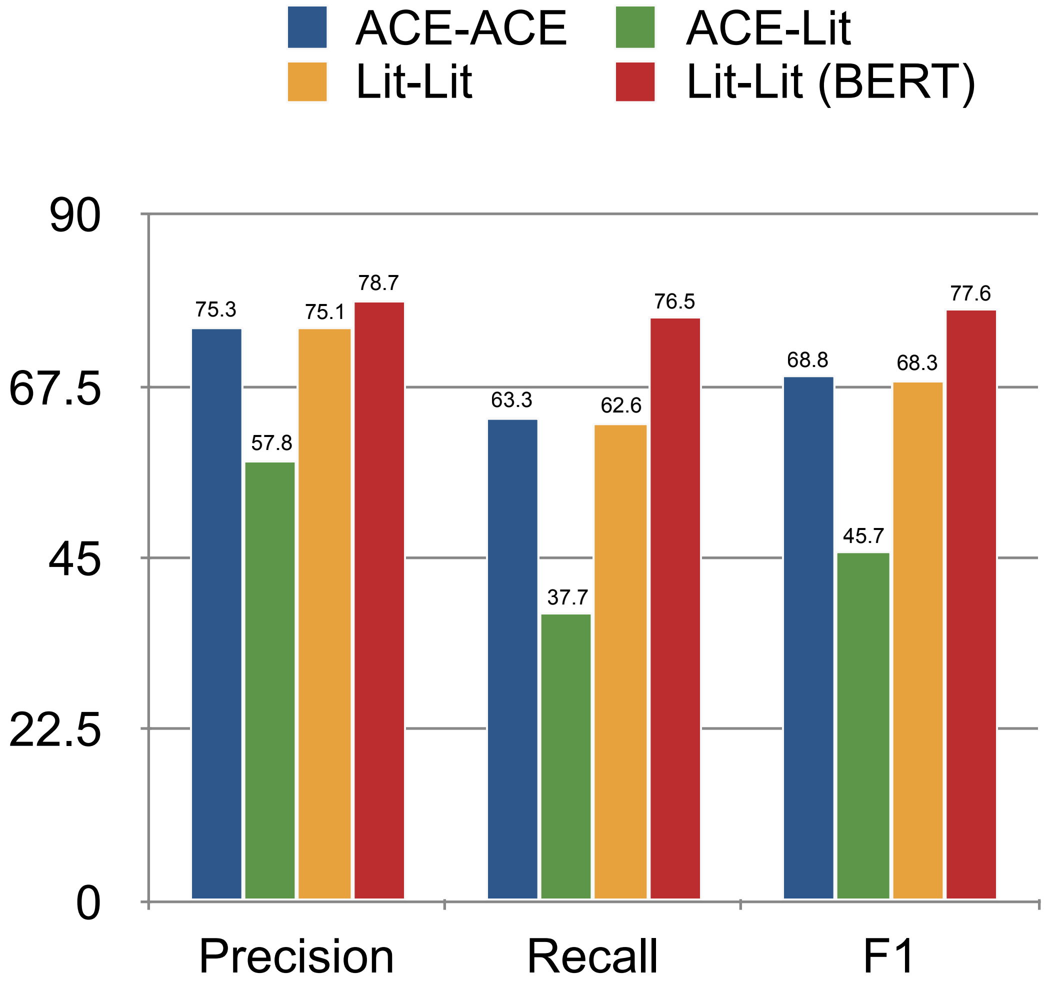
- Ju et al. (2018): layered BiLSTM-CRF; state-of-the-art on ACE 2005.

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



Prediction

- Ju et al. (2018): layered BiLSTM-CRF; state-of-the-art on ACE 2005.
- Evaluate performance difference when altering the training/test domain.
- Adding BERT contextual embeddings (Devlin et al. 2019) yields +9.3 F1 score



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.

Mrs.
Miss
Lady
Aunt

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.

AFP_ENG_20030417.0307

Chapter I: The Bertolini

“The Signora had no business to do it,” said Miss Bartlett, “no business at all. She promised us south rooms with a view close together, instead of which here are north rooms, looking into a courtyard, and a long way apart. Oh, Lucy!”

“And a Cockney, besides!” said Lucy, who had been further saddened by the Signora’s unexpected accent. “It might be London.”

Forster, *A Room with a View*

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.

Training	Women	Men	Diff
ACE	38.0	49.6	-11.6
Literary	69.3	68.2	1.1

Thanks!

David Bamman
dbamman@berkeley.edu