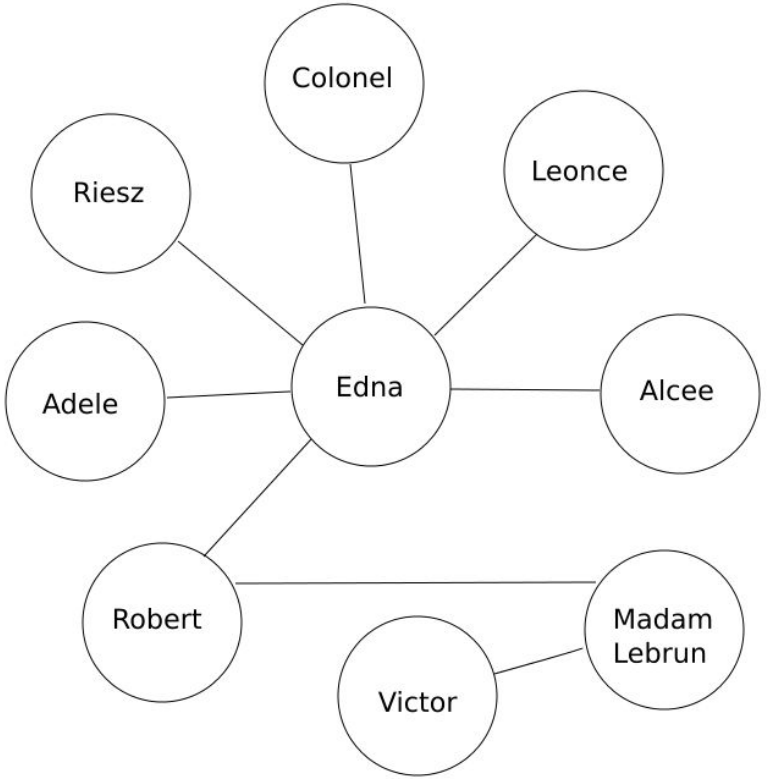
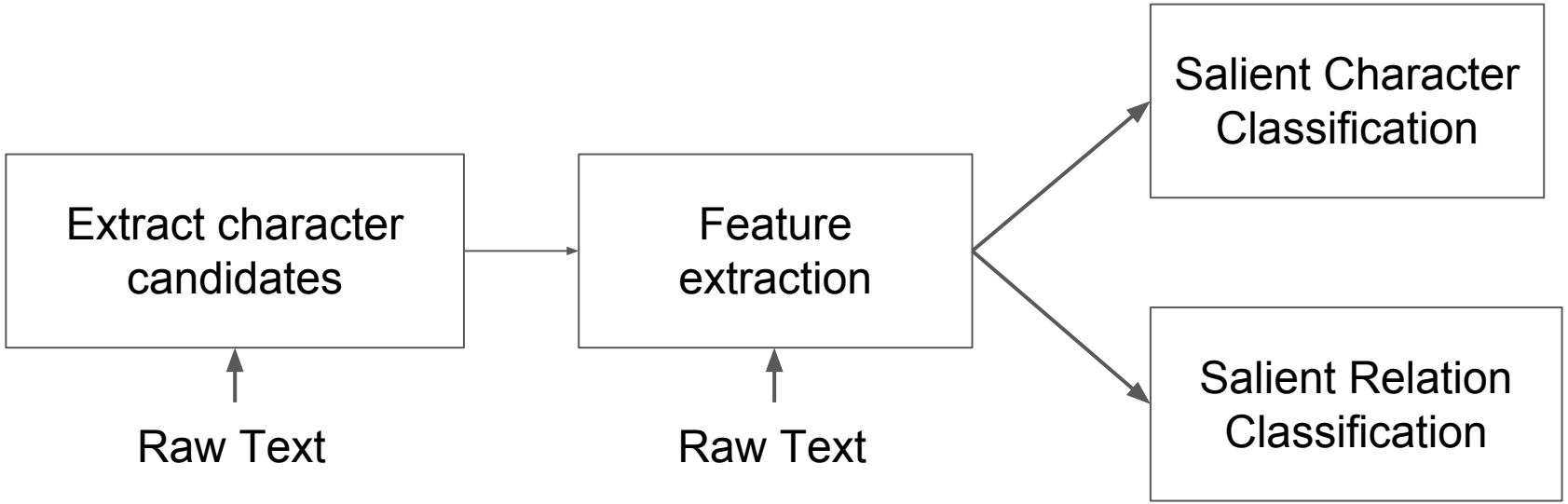
Learning Character Graphs from Literature

Sumit Gogia, Min Zhang, Tommy Zhang

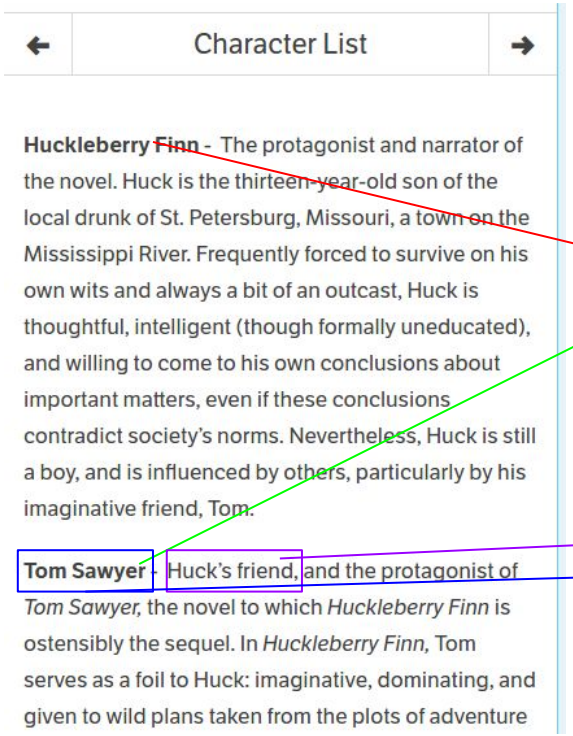
Problem

Can we extract salient characters and whether they are significantly related from raw text?




Extract character candidates
Raw Text
Feature extraction
Raw Text
Salient Character Classification
Salient Relation Classification

Data Collection



Characters
Huckleberry Finn
Tom Sawyer
...

Relations
Huckleberry Finn
|
Tom Sawyer
...

Heuristically Extracted
(person-like noun phrases)

```
CANDIDATES
==================
('Well',)
('Tom',)
('the', 'king')
('the', 'duke')
('Huck',)
('Mary',)
('the', 'old', 'man')
('the', 'nigger')
('Buck',)
('the', 'widow')
('Tom', 'Sawyer')
...
```

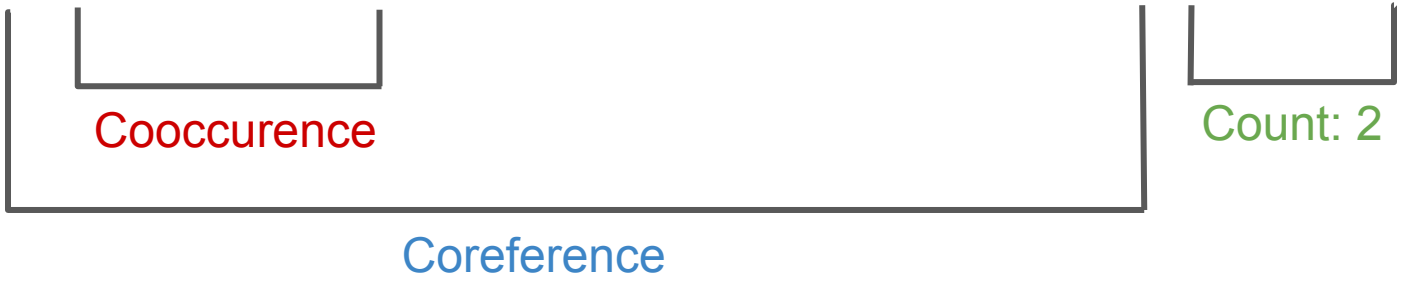
Feature Extraction

Tag Features: Captalization and NER

the widow Bartley

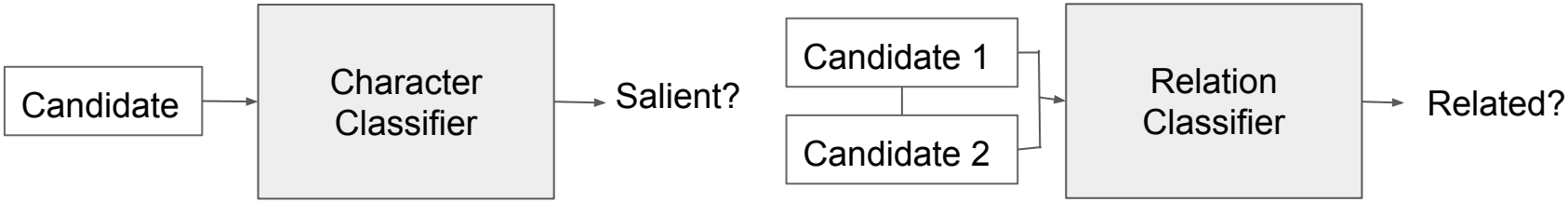
Coreference Features **Count Features** **Cooccurence Features**

Huckleberry was **Tom**'s friend. But everyone called him **Huck**, **Huck** the incredible!



Modeling and Inference

- Simple binary classification methods



- SVM, Random Forests
- Main adjustment is in large weights for positive data (lots of negative examples)

Results (Character)

Precision, Recall (Direct)

('Huck', 'Finn') → 1
('Huck') → 1
('Tom, Sawyer') → 1
('Aunt', 'Sally') → 1
('George', 'Jackson') → 0

Precision, Recall (Disambiguated)

('Huck', 'Finn') → 1
('Huck')
('Tom, Sawyer') → 1
('Aunt', 'Sally') → 1
('George', 'Jackson') → 0

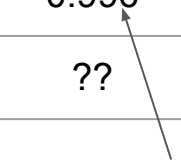
	P1	R1	P2	R2
Baseline 1 (Top n-grams [20, 10, 5])	0.347	0.486	0.351	0.508
Baseline 2 (SVM, RBF, Count Features)	0.630	0.584	0.682	0.593
Full Feature Set (SVM, RBF)	0.638	0.702	0.684	0.593

??

Results (Relations)

Metrics: Similar to those for characters!

	P1	**R1**
Baseline 1 (Top co-occurring)	0.303	0.198
Baseline 2 (SVM, linear, cooc features)	0.188	0.996
Full Feature Set (SVM, linear)	??	??



Very high bias (10)!