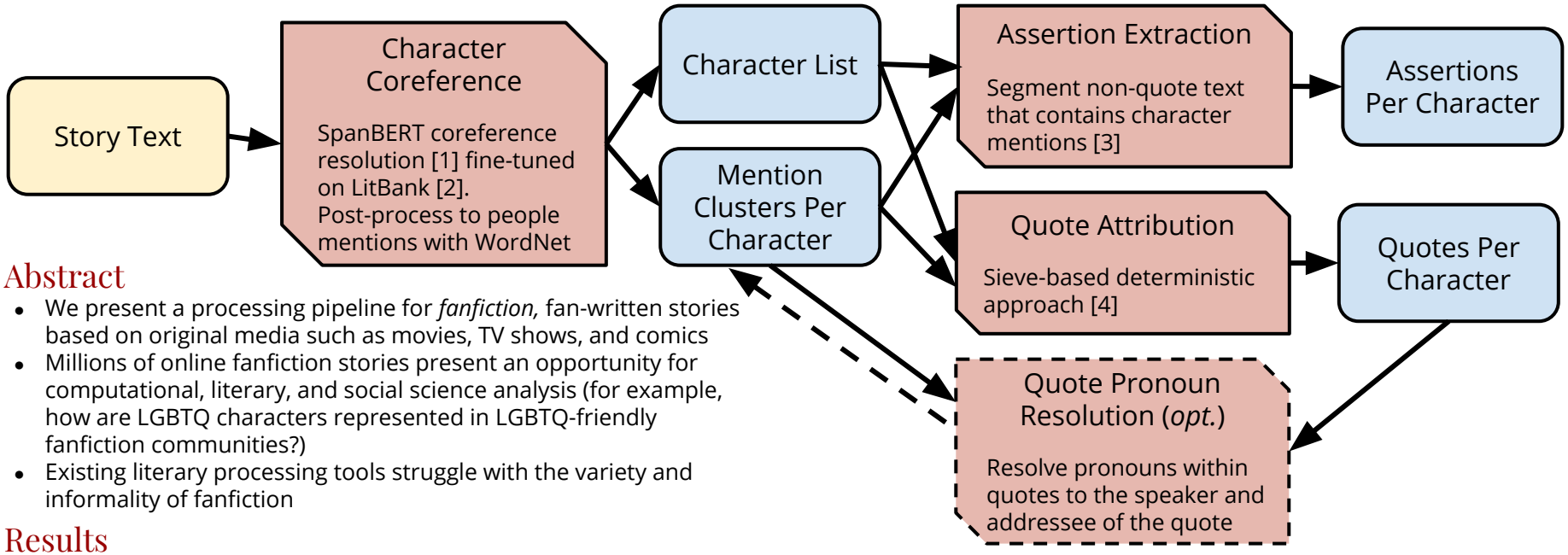
CARNEGIE MELLON UNIVERSITY / LANGUAGE TECHNOLOGIES INSTITUTE

FanfictionNLP: A text processing pipeline for fanfiction

Michael Miller Yoder*, Sopan Khosla*, Qinlan Shen, Aakanksha Naik, Huiming Jin, Hariharan Muralidharan, Carolyn P. Rosé

* denotes equal contribution



Abstract

- We present a processing pipeline for *fanfiction,* fan-written stories based on original media such as movies, TV shows, and comics
- Millions of online fanfiction stories present an opportunity for computational, literary, and social science analysis (for example, how are LGBTQ characters represented in LGBTQ-friendly fanfiction communities?)
- Existing literary processing tools struggle with the variety and informality of fanfiction

Results

- On an annotated dataset of 10 fanfiction stories

Coreference resolution system	CoNLL F1
BookNLP [5]	38.5
BERT-base (LitBank fine-tune)	58.4
SpanBERT-base (LitBank fine-tune)	64.8
FanfictionNLP	**71.4**

Quote attribution system	F1
BookNLP [5]	34.7
He+2013 [6]	53.6
FanfictionNLP [4]	**67.8**

References

1. Joshi, Chen, Liu, Weld, Zettlemoyer, and Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *TACL* 8:64–77.
2. Bamman, Lewke, and Mansoor. 2020. An Annotated Dataset of Coreference in English Literature. *LREC*, 44–54.
3. Hearst. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.
4. Muzny, Fang, Chang, and Jurafsky. 2017. A two-stage sieve approach for quote attribution. *EACL*, 460–470.
5. Bamman, Underwood, and Smith. 2014. A Bayesian Mixed Effects Model of Literary Character. *ACL*, 370–379.
6. He, Barbosa, and Kondrak. 2013. Identification of speakers in novels. *ACL*, 1312–1320.