Automatic Attribution of Quoted Speech in Literary Narrative

By:David K. Elson and Kathleen R. McKeown
Columbia University

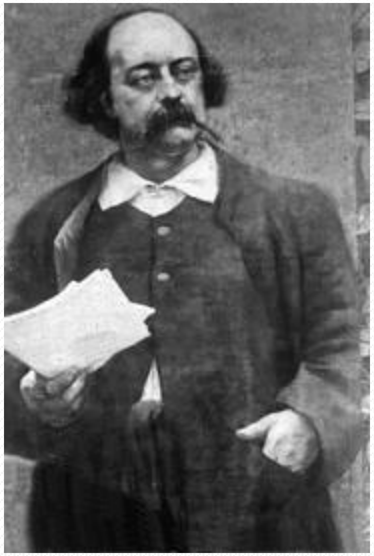
Mehdi Hosseini
Dr. Caroline Sporleder
Saarland University

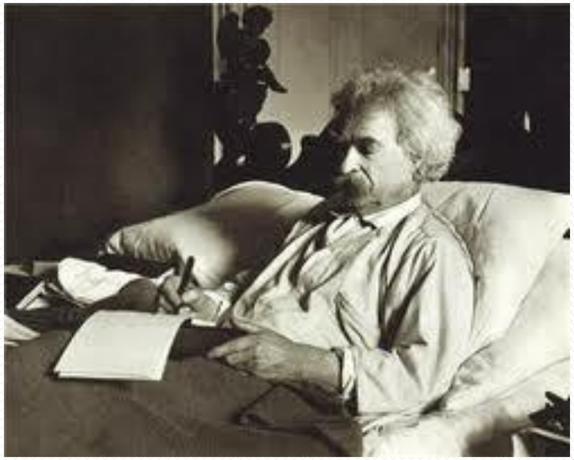
Abstract

- Quoted speech: a block of text within a paragraph falling between quotation marks).

- We will see a method for identifying the speakers of quoted speech in natural-language textual stories

1815 - 1899





Identifying the characters in each scene

- The baseline approach: to find named entities near the quote

Several named entities near the quote

- "Take it," said **Emma**, smiling, and pushing the paper towards **Harriet**– "it is for you. Take your own."

Related Work

- Most Work on the NEWS domain

- Sarmento and Nunes (2009)
- Pouliquen et al. (2007)

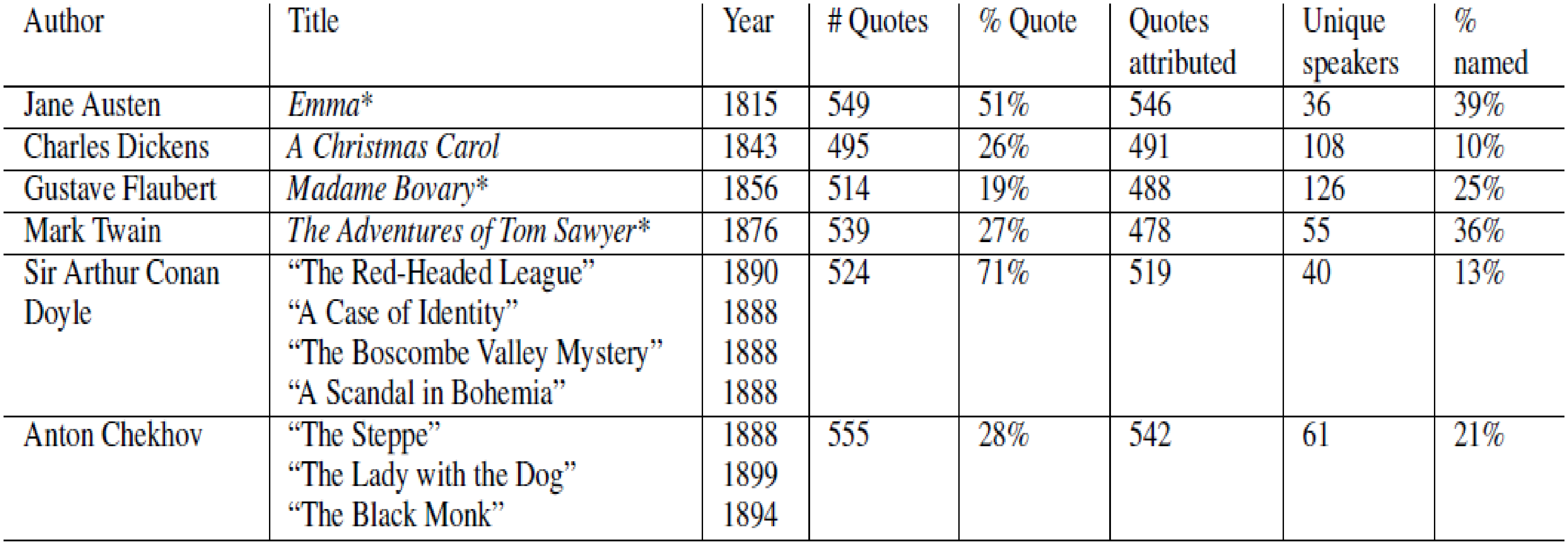
- **Not favorable** for literary narrative, which is less structured than news text in term of attributed quoted speech .

- Mamede and Chaleira (2004) work with a set Portuguese children's stories
- Glass and Bangay (2007): focus on finding the link between the quote, its speech verb and the verb's agent.

Corpus and its annotation

- Six authors who published in 19th century
- Four in English, one in French (translated by Constance Garnett) and one in French (translated by Eleanor Marx Aveling)
- Four authors contribute novels, two short stories
- Dickens often wrote in serial form, but *A Christmas Carol* was published as a single novella

- 111,000 words
- 3,176 quoted speech instances



Author	Title	Year	# Quotes	% Quote	Quotes attributed	Unique speakers	% named
Jane Austen	*Emma**	1815	549	51%	546	36	39%
Charles Dickens	*A Christmas Carol*	1843	495	26%	491	108	10%
Gustave Flaubert	*Madame Bovary**	1856	514	19%	488	126	25%
Mark Twain	*The Adventures of Tom Sawyer**	1876	539	27%	478	55	36%
Sir Arthur Conan Doyle	"The Red-Headed League" "A Case of Identity" "The Boscombe Valley Mystery" "A Scandal in Bohemia"	1890 1888 1888 1888	524	71%	519	40	13%
Anton Chekhov	"The Steppe" "The Lady with the Dog" "The Black Monk"	1888 1899 1894	555	28%	542	61	21%

Table 1: Breakdown of the quoted speech usage in six annotated texts. * indicates that excerpts were used.

Methodology

- The method for quoted speech attribution:

1. Preprocessing
   - Identify all names and nominals appear in the passage of text preceding the quote in question.
2. Classification
   - to classify the quote into one of a set of syntactic categories.
3. Learning
   - to extract a feature vector from the passage and send it to a trained model.

Preprocessing: Finding candidate characters

- First step is to identify the candidate speakers by „chunking“ names (Mr. Holmes) and nominals (the clerk)
- Coreferents and proper names link together as the same entity
- Example: Mr. Sherlock Holmes → Mr. Holmes → Sherlock Holmes → Sherlock → Holmes

- Pronouns won‘t be chunked as character candidates!
- 9% of quotes are attributed to pronouns
- Assign gender to as many names and nominals as possible:
  - Gendered titles: Mr.
  - Gendered headwords: nephew
  - First names: Emma

Encoding, cleaning, and normalizing

- Before extracting features for each candidate, the passage is encoded between the candidate and the quote
- The steps include:
  1. Replace the quote and character with symbols
  2. Replace verb indicate verbal expression or thought with a single symbol <EXPRESS_VERB>
  3. Removing extraneous information
  4. Removing paragraphs, sentenses and clauses that have no information to quoted speech attribution

Dialogue chains

- An author often produces a sequence of quotes by the same speaker, but only attribute the first one
- Example: “Bah!” said Scrooge, “Humbug!”

Syntactic categories

- The quotes and their passgaes are classified to leverage two aspects:
  1. Dialogue chains
  2. The frequent use of expressions

  Pattern matching algorithm assigns to each quote one of five syntactic categories:
    1. Added Quote
    2. Quote Alone
    3. Character trigram: **Quote-Said-Person**: „Bah!“ said Scrooge.
    4. Anaphora trigram
    5. Back Off

- Two categories automatically imply a speaker:
  1. Added Quote
  2. Character Trigram

  The rest are divided to three datasets:
  1. No Apparent Pattern
  2. Quote Alone
  3. Anaphora Trigram

Feature extraction and learning

- To build the mentioned three predictive models, the feature vector ʄ for each candidate-vector pair is used. That include:
  - Distance between candidate and quote (in words)
  - The presence and type of punktuations between the candidate and quote
  - Ordinal position of candidate from the quote among the characters
  - Proportion of the recent quotes, were spoken by the candidate
  - Number of names, quotes, and words in each paragraph
  - Number of apprearance of the candidate
  - For each word near the candidate and quote, whether the word is an expression verb, a punctuation mark, or another person
  - Features of the quote itself: length, position in paragraph, the presence or absence of character names within, ...

$ʄ_{\text{mean}}$: The average value of each feature across the set

Replace the absolute value for each candidate (ʄ) with $ʄ-ʄ_{\text{mean}}$

$ʄ-ʄ_{\text{median}}$

$ʄ-ʄ_{\text{product}}$

$ʄ-ʄ_{\text{max}}$

$ʄ-ʄ_{\text{min}}$

And sending them to the three learners: J48, Jrip, and a two-class logistic regression model

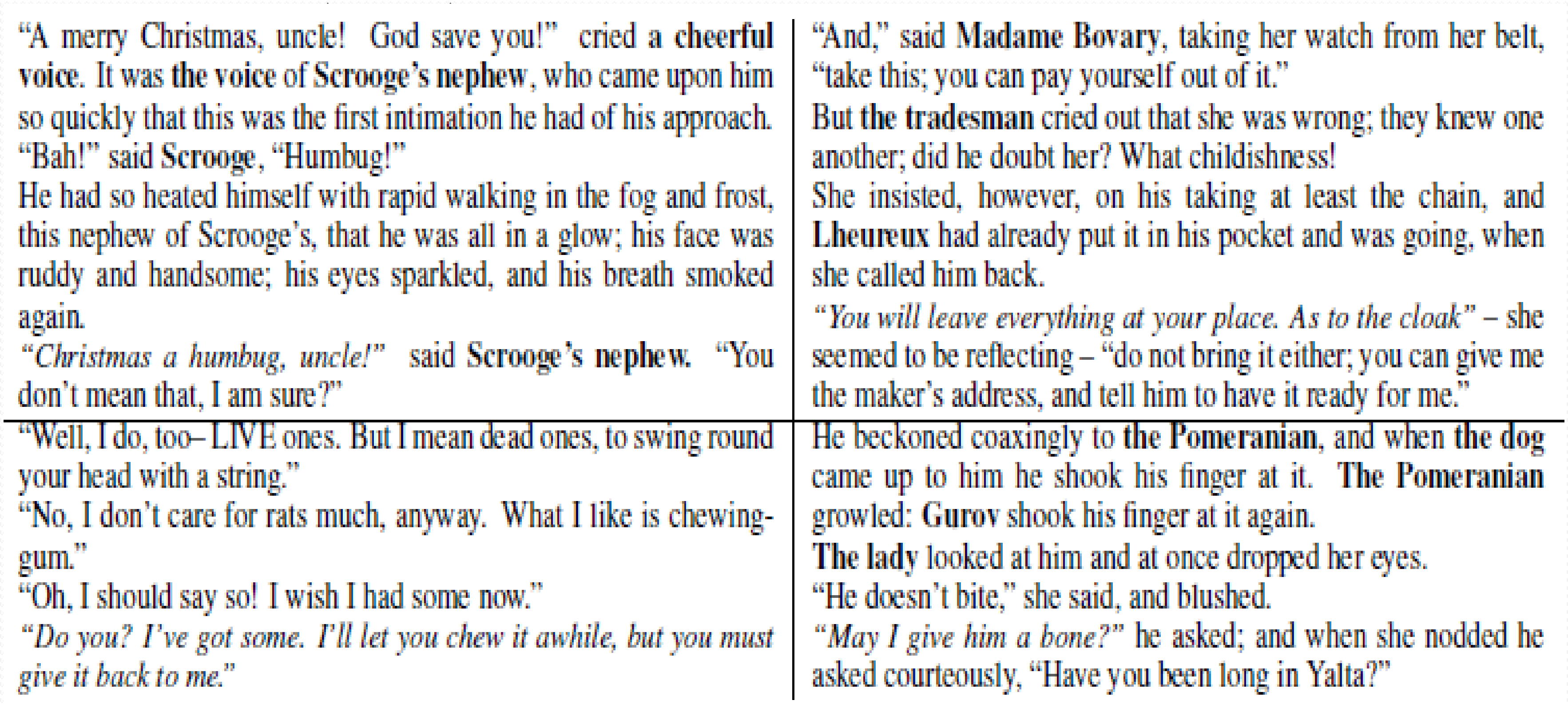
Final Step

- to reconcile the binary results into a single decision for each quote, using one of the four methods:

1. **Label:** Ambiguous, Non-dialogue,
   - Missattributions: (Errors): Overattribution, Underattribution
2. **Single Probability:** threshold
3. **Hybrid:** like Label, if more than one candidat → S.P
4. **Combined Probability:** like S.P, but probability of each candidate is derived from two or three probabilities provided by the classifier: mean, median, product and maximum

Results and discussion

- High recall of the names and nominals chunker



“A merry Christmas, uncle! God save you!” cried **a cheerful voice**. It was **the voice** of **Scrooge’s nephew**, who came upon him so quickly that this was the first intimation he had of his approach. “Bah!” said **Scrooge**, “Humbug!” He had so heated himself with rapid walking in the fog and frost, this nephew of Scrooge’s, that he was all in a glow; his face was ruddy and handsome; his eyes sparkled, and his breath smoked again. *“Christmas a humbug, uncle!”* said **Scrooge’s nephew**. “You don’t mean that, I am sure?”	“And,” said **Madame Bovary**, taking her watch from her belt, “take this; you can pay yourself out of it.” But **the tradesman** cried out that she was wrong; they knew one another; did he doubt her? What childishness! She insisted, however, on his taking at least the chain, and **Lheureux** had already put it in his pocket and was going, when she called him back. *“You will leave everything at your place. As to the cloak”* – she seemed to be reflecting – “do not bring it either; you can give me the maker’s address, and tell him to have it ready for me.”
“Well, I do, too– LIVE ones. But I mean dead ones, to swing round your head with a string.” “No, I don’t care for rats much, anyway. What I like is chewing-gum.” “Oh, I should say so! I wish I had some now.” *“Do you? I’ve got some. I’ll let you chew it awhile, but you must give it back to me.”*	He beckoned coaxingly to **the Pomeranian**, and when **the dog** came up to him he shook his finger at it. **The Pomeranian** growled: **Gurov** shook his finger at it again. **The lady** looked at him and at once dropped her eyes. “He doesn’t bite,” she said, and blushed. *“May I give him a bone?”* he asked; and when she nodded he asked courteously, “Have you been long in Yalta?”

Table 3: Four samples of output that show the extracted character names and nominals (in bold).

- High learning results (83% in average)

Thanks For Your Attention ☺ Any Question?