A Two-stage Sieve Approach for Quote Attribution

Grace Muzny[1], **Michael Fang**[1], **Angel X. Chang**[1,2], **and Dan Jurafsky**[1]
[1]Stanford University, Stanford, CA 94305
[2]Princeton University, Princeton, NJ 08544
{muzny,mjfang,angelx,jurafsky}@cs.stanford.edu

Abstract

We present a deterministic sieve-based system for attributing quotations in literary text and a new dataset: QuoteLi3[1]. Quote attribution, determining who said what in a given text, is important for tasks like creating dialogue systems, and in newer areas like computational literary studies, where it creates opportunities to analyze novels at scale rather than only a few at a time. We release QuoteLi3, which contains more than 6,000 annotations linking quotes to speaker mentions and quotes to speaker entities, and introduce a new algorithm for quote attribution. Our two-stage algorithm first links quotes to mentions, then mentions to entities. Using two stages encapsulates difficult sub-problems and improves system performance. The modular design allows us to tune either for overall performance or for the high precision appropriate for many use cases. Our system achieves an average F-score of 87.5% across three novels, outperforming previous systems, and can be tuned for precision of 90.4% at a recall of 65.1%.

1 Introduction

Dialogue, representing linguistic and social relationships between characters, is an important component of literature. In this paper, we consider the task of quote attribution for literary text: identifying the speaker for each quote. This task is important for developing realistic character-specific conversational models (Vinyals and Le, 2015; Li et al., 2016), analyzing discourse structure, and literary studies (Muzny et al., 2016). But identifying speakers can be difficult; authors often refer to the speaker only indirectly via anaphora, or even omit mention of the speaker entirely (Table 1).

Prior work has produced important datasets labeling quotes in novels, providing training data for supervised methods. But some of these model the quote-attribution task at the mention-level (Elson and McKeown, 2010; O'Keefe et al., 2012), and others at the entity-level (He et al., 2013), leading to labels that are inconsistent across datasets.

We propose entity-level quote attribution as the end goal but with mention-level quote attribution as an important intermediary step. Our first contribution is the QuoteLi3 dataset, a unified combination of data from Elson and McKeown (2010) and He et al. (2013) with the addition of more than 3,000 new labels from expert annotators. This dataset provides both mention and entity labels for *Pride and Prejudice, Emma*, and *The Steppe*.

Next, we describe a new deterministic system that models quote attribution as a two-step process that i) uses textual cues to identify the mention that corresponds to the speaker of a quote, and ii) resolves the mention to an entity. This system improves over previous work by 0.8-2.1 F1 points and its modular design makes it easy to add sieves and incorporate new knowledge.

In summary, our contributions are:

- A unified dataset with both quote-mention and quote-speaker links labeled by expert annotators.
- A new quote attribution strategy that improves on all previous algorithms and allows the incorporation of both rich linguistic features and machine learning components.
- A new annotation tool designed with the specifics of this task in mind.

We freely release the data, system, and annotation tool to the community.[2]

[1] **Quote**s in **Li**terary text from **3** novels.

[2] nlp.stanford.edu/~muzny/quoteli.html

Type	Example	Speaker
Explicit	"Do you really think so?" cried **Elizabeth**, brightening up ...	Elizabeth Bennet
Anaphoric (pronoun)	"You are uniformly charming!" cried **he**, with an air of awkward gallantry;	Mr. Collins
Anaphoric (other)	"I see your design, Bingley," said **his friend**.	Mr. Darcy
Implicit	"Then, my dear, you may have the advantage of your friend, and introduce Mr. Bingley to her."	Mr. Bennet
	"Impossible, Mr. Bennet, impossible, when I am not acquainted with him myself; how can you be so teazing?"	Mrs. Bennet
	"I honour your circumspection. [...] I will take it on myself."	Mr. Bennet
	The girls stared at their father. Mrs. Bennet said only, "Nonsense, nonsense!"	Mrs. Bennet

Table 1: Quotes where speakers are mentioned explicitly, by anaphor, or implicitly (conversationally).

2 Related Work

Early work in quote attribution focused on identifying spans associated with content (quotes), sources (mentions), and cues (speech verbs) in newswire data. This is the approach taken by Pareti et al. (2012; 2013). More recent work by Almeida et al. (2014) performed entity-level quote attribution and showed that a joint model of coreference and quote attribution can help both tasks.

In the literary domain, Glass and Bangay (2007) did early work modeling both the mention-level and entity-level tasks using a rule-based system. However, their system relied on identifying a main speech verb to then identify the actor (i.e. the mention) and link to the speaker (i.e. the entity) from a character list. This system worked very well but was limited to explicitly cued speakers and did not address implicit speakers at all.

Elson and McKeown (2010) took important first steps towards automatic quote attribution. They formulated the task as one of *mention* identification in which the goal was to link a quote to the mention of its speaker. Their method achieved 83.0% accuracy overall, but used gold-label information at test time. Their corpus, the Columbia Quoted Speech Corpus (CQSC), is the most well-known corpus and was used by follow-up work. However, a result of their Mechanical Turk-based labeling strategy was that this corpus contains many unannotated quotes (see Table 4).

O'Keefe et al. (2012) also treated quote attribution as mention identification, using a sequence labeling approach. Their approach was successful in the news domain but it failed to beat their baseline in the literary domain (53.5% vs 49.8% accuracy). This work quantitatively showed that quote attribution in literature was fundamentally different from the task in newswire.

We compare against He et al. (2013), the previous state-of-the-art system for quote attribution. They re-formulated quote attribution as quote-speaker labeling rather than quote-mention labeling. They used a supervised learner and a generative actor topic model (Celikyilmaz et al., 2010) to achieve accuracies ranging from 82.5% on *Pride & Prejudice* to 74.8% on *Emma*.

3 Data: The QuoteLi3 Corpus

We build upon the datasets of He et al. (2013) and Elson and McKeown (2010) to create a comprehensive new dataset of quoted speech in literature: QuoteLi3. This dataset covers 3 novels and 3103 individual quotes, each linked to speaker and mention for a total of 6206 labels, more than 3000 of which are newly annotated. It is composed of expert-annotated dialogue from Jane Austen's *Pride and Prejudice*, *Emma*, and Anton Chekhov's *The Steppe*.

3.1 Previous Datasets

The datasets described in section 2 are valuable but incomplete and hard to integrate with one another given their different designs.

The Columbia Quoted Speech Corpus is a large dataset that includes both quote-mention and quote-speaker labels (Elson and McKeown, 2010). It suffers from problems often associated with crowdsourced labels and the use of low-accuracy tools. In this corpus, quote-mention labels were gathered from Mechanical Turk, where each quote

Novel	He et al.		CQSC (Elson and McKeown, 2010)		QuoteLi3	
	q-mention	q-speaker	q-mention	q-speaker	q-mention	q-speaker
Pride and Prejudice	○	●			●	●
Emma			◑	◑	●	●
The Steppe			◑	◑	●	●

Table 2: Label coverage per novel: ● is full, ◑ is partial, and ○ is no coverage of annotations.

	QuoteLi3 (uncollapsed)			QuoteLi3 (collapsed)			He et al.		
Quote Type	*P & P*	*Emma*	*The Steppe*	*P & P*	*Emma*	*The Steppe*	*P & P*	*Emma*	*The Steppe*
Explicit (ES)	555	240	278	326	128	184	305	106	112
Anaphoric (AS)	528	132	180	309	73	106	292	55	39
pronoun (AS(p))	405	112	106	241	58	58			
other (AS(o))	123	20	74	68	15	48			
Implicit (IS)	664	362	164	655	357	158	663	236	93
Total	1747	734	622	1290	558	448	1260	397	244
All		3103			2296			1901	

Table 3: Breakdown of our dataset by novel and type of quote (uncollapsed). For comparison with the dataset from He et al. (2013), we provide the collapsed statistics assuming one speaker per paragraph.

Quote Types	*Emma*	*The Steppe*
with mention	546 (74.4%)	371 (59.6%)
with speaker	491 (66.9%)	258 (41.5%)

Table 4: Coverage of the CQSC labels

was linked to a mention by 3 different annotators. Elson and McKeown (2010) report that 65% of the quotes in CQSC had unanimous agreement and that 17.6% of the quotes in this corpus were unlabeled. To generate quote-speaker labels, an off-the-shelf coreference tool[3] was used to link mentions and form coreference chains. We find that 57.8% of the quotes in this corpus either i) have no speaker label (48.1%) or ii) the speaker cannot be linked to a known character entity (9.7%). O'Keefe et al. (2012) find that 8% of quotes with speaker labels are incorrectly labeled. Our analysis of the relevant part of CQSC for this work is shown in Table 4.

The data from He et al. (2013) includes high-quality speaker labels but lacks quote-mention labels. There is no overlap in the data provided by He et al. (2013) and CQSC, but this work did evaluate their system on a subset of CQSC. This dataset assumes that all quoted text within a paragraph should be attributed to the same speaker.[4] While this assumption is correct for *Pride and Prejudice*, it is incorrect for novels like *The Steppe*, which use more complex conversational structures[5]. This assumption leads to a problematic method of system evaluation in which all quotes within a paragraph are considered in the gold labels to be one quote, even if they were in fact uttered by different characters. We refer to this strategy as having "collapsed" quotes in our evaluations and present it for the purpose of providing a faithful comparison to previous work.

In QuoteLi3 we add the annotations that are missing from both datasets and correct the existing ones where necessary. A summary of the annotations included in this dataset and comparison to the previous data that we draw from is described in Table 2. Our final dataset is described in Table 3. It features a complete set of annotations for both quote-mention and quote-speaker labels.

3.2 Annotation

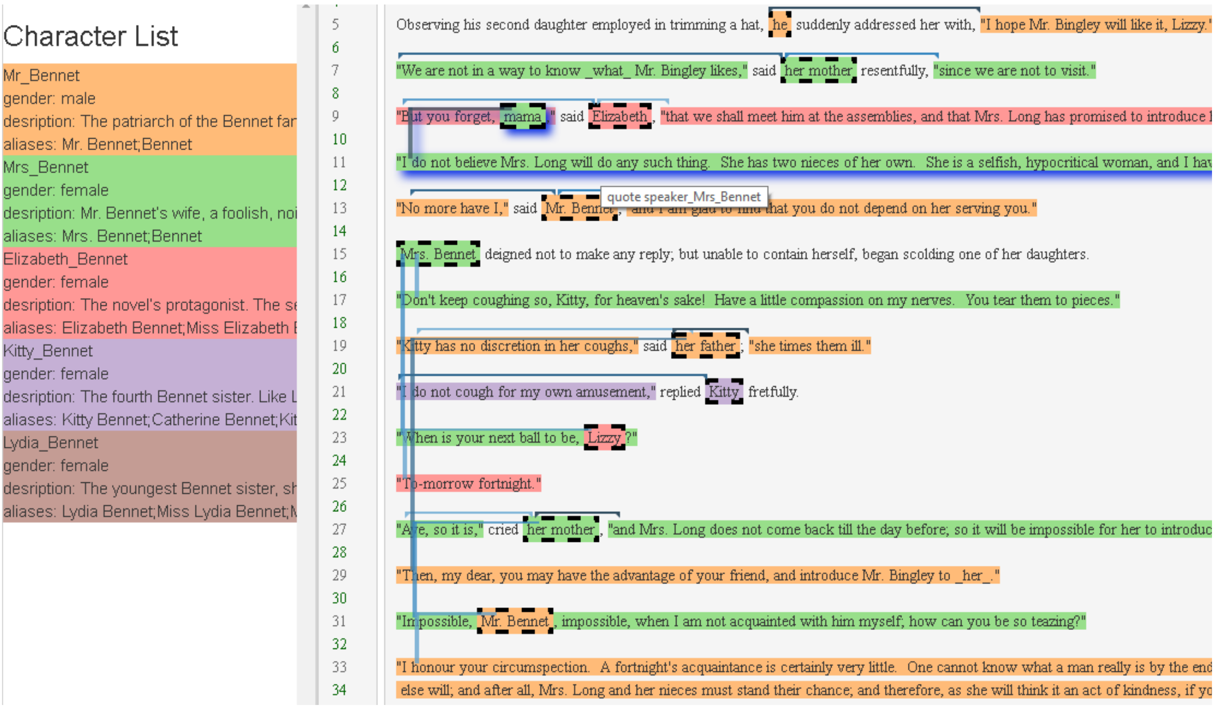
Two of the authors of the paper were the annotators of our dataset. They used annotation guidelines consisting of an example excerpt and a description, which is included in the supplementary materials §A.5. The annotators were instructed to identify the speaker (from a character list) for each quote and to identify the mention that most directly helped them determine the speaker. Unlike Elson and McKeown (2010), mentions can be pronouns and vocatives, not just explicit name referents. Mentions that were closer to the quote and speech verbs were favored over indirect mentions (such as those in conversational chains). Figure 1 shows an example from *Pride and Prejudice*.

Annotation was done using a browser-based an-

[3] Even current state-of-the-art coreference tools achieve just over 65% average F1 scores (Clark and Manning, 2016).

[4] For first-level quotes, there is typically just one speaker per paragraph. This assumption breaks down in some cases and it is very rarely true for nested quotes.

[5] See supplemental section A.1.



Character List

Mr_Bennet
gender: male
desription: The patriarch of the Bennet far
aliases: Mr. Bennet;Bennet

Mrs_Bennet
gender: female
desription: Mr. Bennet's wife, a foolish, noi
aliases: Mrs. Bennet;Bennet

Elizabeth_Bennet
gender: female
desription: The novel's protagonist. The se
aliases: Elizabeth Bennet;Miss Elizabeth B

Kitty_Bennet
gender: female
desription: The fourth Bennet sister. Like L
aliases: Kitty Bennet;Catherine Bennet;Kit

Lydia_Bennet
gender: female
desription: The youngest Bennet sister, sh
aliases: Lydia Bennet;Miss Lydia Bennet;M

Observing his second daughter employed in trimming a hat, he suddenly addressed her with, "I hope Mr. Bingley will like it, Lizzy."

"We are not in a way to know _what_ Mr. Bingley likes," said her mother resentfully, "since we are not to visit."

"But you forget, mama," said Elizabeth, "that we shall meet him at the assemblies, and that Mrs. Long has promised to introduce h

"I do not believe Mrs. Long will do any such thing. She has two nieces of her own. She is a selfish, hypocritical woman, and I hav

"No more have I," said Mr. Bennet, "and I am glad to find that you do not depend on her serving you."

quote speaker_Mrs_Bennet

Mrs. Bennet deigned not to make any reply; but unable to contain herself, began scolding one of her daughters.

"Don't keep coughing so, Kitty, for heaven's sake! Have a little compassion on my nerves. You tear them to pieces."

"Kitty has no discretion in her coughs," said her father; "she times them ill."

"I do not cough for my own amusement," replied Kitty fretfully.

"When is your next ball to be, Lizzy?"

"To-morrow fortnight."

"Aye, so it is," cried her mother, "and Mrs. Long does not come back till the day before; so it will be impossible for her to introduc

"Then, my dear, you may have the advantage of your friend, and introduce Mr. Bingley to _her_."

"Impossible, Mr. Bennet, impossible, when I am not acquainted with him myself; how can you be so teazing?"

"I honour your circumspection. A fortnight's acquaintance is certainly very little. One cannot know what a man really is by the end
else will; and after all, Mrs. Long and her nieces must stand their chance; and therefore, as she will think it an act of kindness, if yo

Figure 1: Conversation from *Pride and Prejudice* annotated with our annotation tool. Speakers are indicated by color, mentions are marked by dashed outlines, and quote-to-mention links by blue lines.

notation tool developed by the authors. Previously developed tools were either not designed for the task (BRAT (Stenetorp et al., 2012), WebAnno (Yimam et al., 2013), CHARLES (Vala et al., 2016)) or unavailable (He et al., 2013). One problem with the CQSC annotations was that the annotators were shown short snippets that lacked the context to determine the speaker and no character list. We designed our tool to provide context and a character list including name, aliases, gender, and description of the character. Similar to CHARLES, the character list is not static and the annotator can add to the list of characters. Our tool also features automatic data consistency checks such as ensuring that all quotes are linked to a mention.

Our expert annotators achieved high inter-annotator agreement with a Cohen's κ of .97 for quote-speaker labels and a κ of .95 for quote-mention labels.[6] To preseve the QuoteLi3 data for train, dev, and testing sets, we calculated this inter-annotator agreement on excerpts from *Alice in Wonderland* and *The Adventures of Huckleberry Finn* containing 176 quotes spoken by 10 characters, chosen to be similar to the data found in QuoteLi3.

3.3 Statistics

Table 3 shows the statistics of our annotated corpus. Unlike He et al. (2013), we do not assume that all quotes in the same paragraph are spoken by the same speaker. To compare with the dataset used by He et al. (2013), we provide the collapsed statistics as well. As Table 3 shows, we have roughly the same number of annotated quotes for *Pride and Prejudice* as He et al. (2013). For *Emma* and *The Steppe*, which were taken from the CQSC corpus, we have considerably more quotes because of our added annotations (see Table 4).

4 The Quote Attribution Task

The task of quote attribution can be summarized as "who said that?" Given a text as input, the final output is a speaker for each uttered quote in the text. We assume that all quotes have been previously identified. O'Keefe et al. (2012) find that regular-expression approaches to quote detection yield over 99% accuracy for clean English-language data. A number of other approaches to quote detection have been studied in recent years for more complex data (Pouliquen et al., 2007; Pareti et al., 2013; Muzny et al., 2016; Scheible et al., 2016). Following He et al. (2013), we assume that there is a predefined list of characters available, with the name, aliases, and gender of each

[6] The reported agreement is the average of the Cohens kappas from these passages.

character.[7]

Some key challenges in quote attribution are resolving anaphora (i.e., coreference) and following conversational threads. Literature often follows specific patterns that make some quotes easier to attribute than others. Therefore, an approach that anchors conversations on easily identifiable quotes can outperform approaches that do not.

Figure 1 shows an example of a complex conversation at the beginning of *Pride and Prejudice*. This example illustrates the spectrum of easy to difficult cases found in the task: simple explicit named mention (lines 9, 13, 21), nominal mentions (lines 7, 19, 27), and pronoun mentions (line 5). Sometimes explicitly named mentions embedded in more complex sentences can still be challenging as they require good dependency parses. This example also illustrates a conversational chain with alternating speakers between Mrs. Bennet and Elizabeth Bennet (lines 7 to 11), and between Mr. Bennet and Mrs. Bennet (lines 27 to 34). In this case, vocatives (expressions that indicate the party being addressed) are cues for who the other speaker is (lines 9, 23, 31). When the simple alternation pattern is broken, explicit speech verbs with the speaking character are specified. To summarize, there are several explicit cues and some easy cases in a conversation that can be leveraged to make the hard cases easier to address.

First, consider the quote→mention linking subtask. This is an inherently ambiguous task (i.e. any mention from the same coreference chain is valid,) but we know that if the target quote is linked to the annotated mention that this is one correct option. This means that the evaluation of the quote→mention stage is a lower-bound. In other words, since a given quote may have multiple mentions that could be considered correct, our system may choose a "wrong" mention for a quote but link it to the correct speaker in the end. Thus, if our mention→speaker system could perfectly resolve every mention to its correct speaker, our overall quote attribution system would be guaranteed to get at minimum the same results as the quote→mention stage.

The quote→speaker task can be tackled directly without addressing quote→mention, but identifying a mention associated with the speaker allows us to incorporate key outside information. Another advantage of this approach is that we can then separately analyze and improve the performance of the two stages.

Therefore we evaluate both subtasks to give a more complete picture of when the system fails and succeeds. We use precision, recall, and F1 so that we can tune the system for different needs.

5 Approach

Our model is a two-stage deterministic pipeline. The first stage links quotes to specific mentions in the text and the second stage matches mentions to the entity that they refer to.

By doing both quote→mention and mention→entity linking, our system is able to leverage additional contextual information, resulting in a richer, labeled output. Its modular design means that it can be easily updated to account for improvements in various sub-areas such as coreference resolution. We use a sieve-based architecture because having accurate labels for the easy cases allows us to first find anchors that help resolve harder, often conversational, cases. Sieve-based systems have been shown to work well for tasks like coreference resolution (Raghunathan et al., 2010; Lee et al., 2013), entity linking (Hajishirzi et al., 2013), and event temporal ordering (Chambers et al., 2014).

5.1 Quote→Mention

The quote→mention stage is a series of deterministic sieves. We describe each in detail in the following sections and show examples in Table 5.

Trigram Matching This sieve is similar to patterns used in Elson and McKeown (2010). It uses patterns like Quote-Mention-Verb (e.g `''...'' she said`) where the mention is either a character name or pronoun to isolate the mention. Other patterns include Quote-Verb-Mention, Mention-Verb-Quote, and Verb-Mention-Quote.

Dependency Parses The next sieve in our pipeline inspects the dependency parses of the sentences surrounding the target quote. We use the enhanced dependency parses (Schuster and Manning, 2016) produced by Stanford CoreNLP (Chen and Manning, 2014) to extract all verbs and their dependent `nsubj` nodes. If the verb is a common speech verb[8] and its `nsubj` relation points to a

[7] Character lists are available on sites like sparknotes.com. The automatic extraction of characters from a novel has been identified as a separate problem (Vala et al., 2015).

[8] This list of verbs as well as the family relation nouns list are available in supplemental section A.4.

Sieve	Example
Trigram Matching	“They have none of them much to recommend them,” replied **he**.
Dependency Parses	**Mrs. Bennet** said only, “Nonsense, nonsense!”
Single Mention Detection	...**Elizabeth** impatiently. “There has been many a one, I fancy, overcome in the same way. I wonder who first discovered the efficacy of poetry in driving away love!”
Vocative Detection	“My dear **Mr. Bennet**,...” “Is that his design in settling here?”
Paragraph Final Mention Linking	After a silence of several minutes, **he** came towards her in an agitated manner, and thus began, “In vain have I struggled...”
Supervised Sieve	–
Conversation Detection	“Aye, so it is,” cried **her mother** ... “Then, my dear, you may have the advantage of your friend, and introduce Mr. Bingley to her.” “Impossible, Mr. Bennet, impossible, when I am not acquainted with him myself; how can you be so teazing?”
Loose Conversation Detection	“I will not trust myself on the subject,” replied **Wickham**; ”I can hardly be just to him.” Elizabeth was again deep in thought, and after a time exclaimed, “To treat in ... the favourite of his father!” She could have added, “A young man, too,... being amiable”– but she contented herself with, “and one, too, ... in the closest manner!” “We were born in the same parish, within the same park; the greatest part of our youth was passed together;...”

Table 5: Quote→Mention sieves and example quotes that they apply to.

Sieve	Example
Exact Name Match	*“Do you really think so?”* cried **Elizabeth**, brightening up for a moment.
Coreference Disambiguation	*“You are uniformly charming!”* cried **he**, with an air of awkward gallantry;
Conversational Pattern	*“Impossible, Mr. Bennet, impossible ...”* (Mrs. Bennet) “I honour your circumspection...I will take it on myself.” (Mr. Bennet) The girls stared at their father. Mrs. Bennet said only, “Nonsense, nonsense!” (Mrs. Bennet)
Family Noun Vocative Disambiguation	“...You know, **sister**, we agreed long ago never to mention a word about it. And so, is it quite certain he is coming?” *“You may depend on it,”* replied the other ...
Majority Speaker	–

Table 6: Mention→Speaker Sieves and example quotes that they apply to. Bold text indicates where the speaker information comes from while italic text indicates the target quote being labeled.

character name, a pronoun, or an animate noun,[9] we assign the quote to the target mention.

Single Mention Detection If there is only a single mention in the non-quote text in the paragraph of the target quote, link the quote to that mention.

Vocative Detection If the preceding quote contains a vocative pattern (see supplemental section A.2), link the target quote to that mention. Vocative detection only matches character names and animate nouns.

Paragraph Final Mention Linking If the target quote occurs at the end of a paragraph, link it to the final mention occurring in the preceding sentence.

Conversational Pattern If a quote in paragraph n has been linked to mention m_n, then this sieve links an unattributed quote two paragraphs ahead, $n + 2$, to mention m_n if they appear to be in conversation. We consider two quotes “in conversation” if the paragraph between is also a quote, and the quote in paragraph $n + 2$ appears without additional (non-quote) text.

Loose Conversational Pattern We include a looser form of the previous sieve as a final, high-recall, step. If a quote in paragraph n has been linked to mention m_n, then this sieve links quotes in paragraph $n + 2$ to m_n without restriction.

5.2 Mention→Speaker

The second stage of our system involves linking the mentions identified in the first stage to a speaker entity. We again use several simple, deterministic sieves to determine the entity that each mention and quote should be linked to. A description of these sieves and example mentions and quotes that they are applied to appears in Table 6.

For the following sieves, we construct an ordered list of *top_speakers* by counting proper name and pronoun mentions around the target quote. If gender for the target quote’s speaker can be determined either by the gender of a pronoun mention or the gender of an animate noun (Bergsma and

[9] The list of animate nouns is from Ji and Lin (2009).

Lin, 2006), this information is used to filter the candidate speakers in the *top_speakers* list.

We use a window size from 2000 tokens before the target quote to 500 tokens after the target quote. If no speakers matching in gender can be found in this window, it is expanded by 2000 tokens on both sides.

Exact Name Match If the mention that a quote is linked to matches a character name or alias in our character list, label the quote with that speaker.

Coreference Disambiguation If the mention is a pronoun, we attempt to disambiguate it to a specific character using the coreference labels provided by BookNLP (Bamman et al., 2014).

Conversational Pattern Similarly as in the quote→mention section, we match a target quote to the same speaker as a quote in paragraph $n + 2$, if they are in the same conversation and it is labeled. Next, we match it to the quote in paragraph $n - 2$ if they are in the same conversation and it is labeled. This sieve receives gender information from the mention that the target quote is linked to.

Family Noun Vocative Disambiguation If the target quote is linked to a vocative in the list of family relations (e.g. "papa"), pick the first speaker in *top_speakers* that matches the last name of the speaker of the quote containing the vocative.

Majority Speaker If none of the previous sieves identified a speaker for the quote, label the quote with the first speaker in the *top_speakers* list.

6 Experiments

In all experiments, we divide the data as follows: *Pride and Prejudice* is split as in He et al. (2013) with chapters 19-26 as the test set, 27-33 as the development set, and all others as training. *Emma* and *The Steppe* are not used for training.

6.1 Baseline

As a baseline, for the quote→mention stage we choose the mention that is closest to the quote in terms of token distance. This is similar to the approach taken in BookNLP (Bamman et al., 2014), in which quotes are attributed to a mention by first looking for the closest mention in the same sentence to the left and right of the quote, then before a hard stop or another quote to the left and right of the target quote. For the mention→speaker stage, we use the Exact Name Match and Coreference Disambiguation sieves.

6.2 Comparison to Previous Work

Table 7 shows a direct comparison of our work versus the previous systems. We replicate the test conditions used by He et al. (2013) as closely as possible in this comparison.

In this comparison, the evaluations based on CQSC are of non-contiguous subsets of the quotes that are also not necessarily the same between our work and the previous work. As discussed in section 3, CQSC provides an incomplete set of quote-speaker labels. In this work we follow the same methodology as He et al. (2013) to extract a test set of unambiguously labeled quotes by using a list of character names to identify those that are unambiguously labeled. In section 7, we analyze *The Steppe* and *Emma* more thoroughly, showing that this method results in an easier subset of the quotes in these novels.

Our preferred evaluation, shown in Table 8, differs from previous evaluations in four important ways. We hope that this work can establish consistent guidelines for attributing quotes and evaluating system performance to encourage future work.

- Each quote is attributed separately.[10]
- The test sets are composed of every quote from the test portion of each novel, no subsets are used because of incomplete annotations.[11]
- No gold data is used at test time.[12]
- Precision and recall are reported in preference to accuracy for a more fine-grained understanding of the underlying system.

6.3 Adding a Supervised Component

To test how orthogonal our two-stage approach is to previous systems, we experiment by adding a supervised sieve to the quote→ mention stage. We train a binary classifier, using a maxent model to distinguish between the correct and incorrect candidate mentions.

[10] This is in contrast to the work of He et al. (2013)

[11] This is in contrast to the work of Elson and McKeown (2010) and He et al. (2013). The work of O'Keefe et al. (2012) is the only previous work to augment the unlabeled portions of CQSC. They achieved 53.3% accuracy on CQSC from a rule-based system similar to our baseline. This data is not available.

[12] Gold data was used at test time by Elson and McKeown (2010) who achieved 83.0% accuracy on the CQSC.

Test	He et al.	Baseline	This work	+ supervised
Pride and Prejudice	82.5	45.3	83.6	**85.2**
Emma	74.8*	40.7*	75.3*	**76.1***
The Steppe	80.3*	66.7*	81.8*	**83.8***

Table 7: Comparison with previous work. This table reports accuracy and comes with some caveats: ∗ indicates that a non-contiguous subset of the quotations were used (not all subsets are guaranteed to be the same as described in section 6.2), and all quotes within the same paragraph were collapsed. *Emma* and *The Steppe* come from CQSC. All systems are trained on *Pride and Prejudice*.

System	Test	Quote→Mention			Mention→Speaker			
		P	R	F1	P	R	F1	Accuracy
+supervised	*Pride and Prejudice*	86.7	93.5	89.9	85.1	100	92.0	85.1
+supervised	*Emma*	75.2	85.2	79.9	75.9	100	86.3	75.9
+supervised	*The Steppe*	81.7	88.6	85.0	72.7	100	84.2	72.7
	Average	81.2	89.1	84.9	77.9	100	87.5	
+precision	*Pride and Prejudice*	90.2	80.1	84.9	92.1	70.9	80.1	
+precision	*Emma*	84.6	68.3	75.6	85.7	59.0	69.9	
+precision	*The Steppe*	92.5	75.3	83.0	93.3	65.5	77.0	
	Average	89.1	74.6	81.2	90.4	65.1	75.7	

Table 8: Precision, recall, and F-Score of our systems on un-collapsed quotations and the fully annotated test sets from the QuoteLi3 dataset.

Test	ES	AS(p)	AS(o)	IS	All
P & P	98.4	77.3	42.9	82.3	85.1
Emma	92.1	62.5	35.0	71.5	75.9
The Steppe	97.5	67.0	14.9	60.4	72.7

Table 9: Breakdown of the accuracy of our system per type of quote (see Table 3) in each test set.

Candidate Mentions We take as candidate mentions all token spans corresponding to names, pronouns, and animate nouns in a one-paragraph range on either side of the quote. Names are determined by scanning for matches to the character list. We restrict pronouns to singular gendered pronouns, i.e. ‘he’ or ‘she’.

Features We featurize each (quote, mention) pair based on attributes of the quote, mention, and how far apart they are from one another. These features largely align with previous work and can be found in supplemental section A.3 (Elson and McKeown, 2010; He et al., 2013).

Prediction At test time our model predicts for each quote whether each candidate mention is or is not the correct mention to pair with that quote. If the model predicts more than one mention to be correct, we take the most confident result.

This sieve goes just before the conversation pattern detection sieves in the quote→mention stage (see Table 5). This forms our +*supervised* system.

6.4 Creating a High-Precision System

One advantage of our sieve design is that we can easily add and remove sieves from our pipeline. This means that we can determine the combination of sieves that result in the system that achieves the highest precision with respect to the final speaker label. We use an ablation test to find the combination of sieves with the highest precision (95.6%) for speaker labels on the development set from *Pride and Prejudice*. These results are achieved by removing the Loose Conversation Detection sieve for the quote→mention stage and keeping only the Exact Name Match and Coreference Disambiguation sieves for the mention→speaker stage.

Together, these sieves create a system that we call +*precision* that emphasizes overall precision rather than F-score or accuracy.

7 Results

We show that a simple deterministic system can achieve state-of-the-art results. Adding a lightweight supervised component improves the system across all test sets. The sieve design allows us to create a high precision system that might be more appropriate for real-world applications that value precision over recall.

The results in Table 8 confirm that the subset of test quotes from *Emma* and *The Steppe* used in previous work were an easier subset of the whole set of quotations. When evaluating based off of the whole set of quotations, we lose 0.2 and 11.1 points of accuracy for *Emma* and *The Steppe*, respectively. As we show in Table 4, *The Steppe* is missing a significant portion (50.9%) of the annotations whereas *Emma* is missing 28.6%. Our error analysis shows us that *The Steppe* features more complicated conversation patterns than the novels of Jane Austen, which makes the task of quote attribution more difficult.

One type of error analysis we performed was inspecting the accuracy of our system by quote type. As seen in Table 9, the main challenge lies in identifying anaphoric and implicit speakers. We find that resolving non-pronoun anaphora is much more challenging for our system than pronouns. This is because the only mechanism for dealing with these mentions is the Family Noun Vocative Disambiguation sieve; otherwise, the only information we gather from them is gender information. This indicates that adding information about the social network of a novel and attributes of each character (such as job and relationships to other characters) would further increase system performance.

8 Conclusion

In this paper, we provided an improved, consistently annotated dataset for quote attribution with both quote-mention and quote-speaker annotations. We described a two-stage quote attribution system that first links quotes to mentions and then mentions to speakers, and showed that it outperforms the existing state-of-the-art. We established a thorough evaluation and showed how our system can be tweaked for higher precision or refined with a supervised sieve for better overall performance.

A clear direction for future work is to expand the dataset to a more diverse set of novels by leveraging our annotation tool on Mechanical Turk or other crowdsourcing platforms. This work has also provided the background to see the pitfalls that a dataset produced in such a way might encounter. For example, annotators could label mentions and speakers separately, and examples with high uncertainty could be transferred to expert annotators. An expanded dataset would allow us to evaluate how well our system generalizes to other novels and also allow us to train better models. Another interesting direction is to eliminate the use of predefined character lists by automatically extracting the list of characters (Vala et al., 2015).

Acknowledgments

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-114747, and by the NSF via IIS IIS-1514268. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors thank their anonymous reviewers and the Stanford NLP group for their helpful feedback.

References

Mariana SC Almeida, Miguel B Almeida, and André FT Martins. 2014. A joint model for quotation attribution and coreference resolution. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

David Bamman, Ted Underwood, and Noah A Smith. 2014. A Bayesian mixed effects model of literary character. In *Proceedings of Association for Computational Linguistics (ACL)*.

Shane Bergsma and Dekang Lin. 2006. Bootstrapping path-based pronoun resolution. In *Proceedings of Association for Computational Linguistics (ACL)*.

Asli Celikyilmaz, Dilek Hakkani-Tur, Hua He, Greg Kondrak, and Denilson Barbosa. 2010. The actor-topic model for extracting social networks in literary narrative. In *Proceedings of NIPS Workshop: Machine Learning for Social Computing*.

Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. 2014. Dense event ordering with a multi-pass architecture. *Transactions of the Association for Computational Linguistics*, 2:273–284.

Danqi Chen and Christopher D Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of Empirical Methods on Natural Language Processing (EMNLP)*.

Kevin Clark and Christopher D. Manning. 2016. Improving coreference resolution by learning entity-level distributed representations. In *Proceedings of Association for Computational Linguistics (ACL)*.

David K Elson and Kathleen McKeown. 2010. Automatic attribution of quoted speech in literary narrative. In *Proceedings of Association for the Advancement of Artificial Intelligence (AAAI)*.

Kevin Glass and Shaun Bangay. 2007. A naïve, salience-based method for speaker identification in fiction books. In *Proceedings of the 18th International Symposium of the Pattern Recognition Association of South Africa (PRASA)*.

Hannaneh Hajishirzi, Leila Zilles, Daniel S. Weld, and Luke S. Zettlemoyer. 2013. Joint coreference resolution and named-entity linking with multi-pass sieves. In *Proceedings of EMNLP*, pages 289–299.

Hua He, Denilson Barbosa, and Grzegorz Kondrak. 2013. Identification of speakers in novels. In *Proceedings of Association for Computational Linguistics (ACL)*.

Heng Ji and Dekang Lin. 2009. Gender and animacy knowledge discovery from web-scale n-grams for unsupervised person mention detection. In *Proceedings of Pacific Asia Conference on Language, Information and Computation (PACLIC)*.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4):885–916.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of Association for Computational Linguistics (ACL)*.

Grace Muzny, Mark Algee-Hewitt, and Dan Jurafsky. 2016. The dialogic turn and the performance of gender: the English canon 1782-2011. In *Proceedings of Digital Humanities*, pages 296–299.

Tim O'Keefe, Silvia Pareti, James R Curran, Irena Koprinska, and Matthew Honnibal. 2012. A sequence labelling approach to quote attribution. In *Proceedings of Empirical Methods on Natural Language Processing (EMNLP)*.

Silvia Pareti, Timothy O'Keefe, Ioannis Konstas, James R Curran, and Irena Koprinska. 2013. Automatically detecting and attributing indirect quotations. In *Proceedings of Empirical Methods on Natural Language Processing (EMNLP)*.

Silvia Pareti. 2012. A database of attribution relations. In *Proceedings of International Conference on Language Resources and Evaluation (LREC)*.

Bruno Pouliquen, Ralf Steinberger, and Clive Best. 2007. Automatic detection of quotations in multilingual news. In *Proceedings of Recent Advances in Natural Language Processing*.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A multipass sieve for coreference resolution. In *Proceedings of Empirical Methods on Natural Language Processing (EMNLP)*.

Christian Scheible, Roman Klinger, and Sebastian Padó. 2016. Model architectures for quotation detection. In *Proceedings of Association for Computational Linguistics (ACL)*.

Sebastian Schuster and Christopher D. Manning. 2016. Enhanced English Universal Dependencies: An improved representation for natural language understanding tasks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Hardik Vala, David Jurgens, Andrew Piper, and Derek Ruths. 2015. Mr. Bennet, his coachman, and the archbishop walk into a bar but only one of them gets recognized: On the difficulty of detecting characters in literary texts. In *Proceedings of Empirical Methods on Natural Language Processing (EMNLP)*.

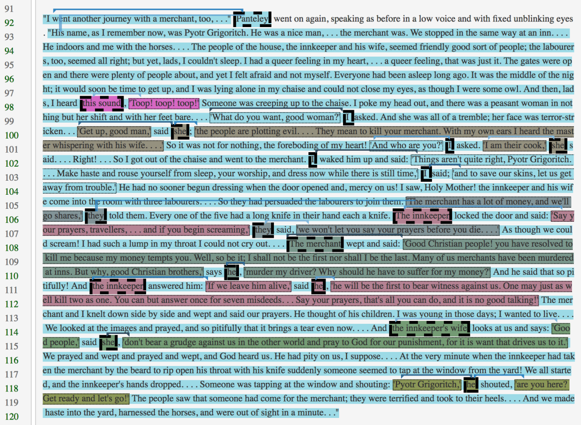
Hardik Vala, Stefan Dimitrov, David Jurgens, Andrew Piper, and Derek Ruths. 2016. Annotating characters in literary corpora: A scheme, the CHARLES tool, and an annotated novel. In *Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC)*.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. In *Proceedings of ICML Deep Learning Workshop*.

Seid Muhie Yimam, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. 2013. WebAnno: A flexible, web-based and visually supported system for distributed annotations. In *Proceedings of Association for Computational Linguistics (ACL) System Demonstrations*.

A Supplemental Material

A.1 Nested Conversation Example



"I went another journey with a merchant, too, . . ." Panteley went on again, speaking as before in a low voice and with fixed unblinking eyes . "His name, as I remember now, was Pyotr Grigoritch. He was a nice man, . . . the merchant was. We stopped in the same way at an inn. . . . He indoors and me with the horses. . . . The people of the house, the innkeeper and his wife, seemed friendly good sort of people; the labourers, too, seemed all right; but yet, lads, I couldn't sleep. I had a queer feeling in my heart, . . . a queer feeling, that was just it. The gates were open and there were plenty of people about, and yet I felt afraid and not myself. Everyone had been asleep long ago. It was the middle of the night; it would soon be time to get up, and I was lying alone in my chaise and could not close my eyes, as though I were some owl. And then, lads, I heard this sound, 'Toop! toop! toop!' Someone was creeping up to the chaise. I poke my head out, and there was a peasant woman in nothing but her shift and with her feet bare. . . . 'What do you want, good woman?' I asked. And she was all of a tremble; her face was terror-stricken. . . 'Get up, good man,' said she; 'the people are plotting evil. . . . They mean to kill your merchant. With my own ears I heard the master whispering with his wife. . . .' So it was not for nothing, the foreboding of my heart! 'And who are you?' I asked. 'I am their cook,' she said. . . . Right! . . . So I got out of the chaise and went to the merchant. I waked him up and said: 'Things aren't quite right, Pyotr Grigoritch. . . . Make haste and rouse yourself from sleep, your worship, and dress now while there is still time,' I said; 'and to save our skins, let us get away from trouble.' He had no sooner begun dressing when the door opened and, mercy on us! I saw, Holy Mother! the innkeeper and his wife come into the room with three labourers. . . . So they had persuaded the labourers to join them. 'The merchant has a lot of money, and we'll go shares,' they told them. Every one of the five had a long knife in their hand each a knife. The innkeeper locked the door and said: 'Say your prayers, travellers, . . . and if you begin screaming,' they said, 'we won't let you say your prayers before you die. . . .' As though we could scream! I had such a lump in my throat I could not cry out. . . . The merchant wept and said: 'Good Christian people! you have resolved to kill me because my money tempts you. Well, so be it; I shall not be the first nor shall I be the last. Many of us merchants have been murdered at inns. But why, good Christian brothers,' says he, 'murder my driver? Why should he have to suffer for my money?' And he said that so pitifully! And the innkeeper answered him: 'If we leave him alive,' said he, 'he will be the first to bear witness against us. One may just as well kill two as one. You can but answer once for seven misdeeds. . . Say your prayers, that's all you can do, and it is no good talking!' The merchant and I knelt down side by side and wept and said our prayers. He thought of his children. I was young in those days; I wanted to live. . . . We looked at the images and prayed, and so pitifully that it brings a tear even now. . . . And the innkeeper's wife looks at us and says: 'Good people,' said she, 'don't bear a grudge against us in the other world and pray to God for our punishment, for it is want that drives us to it.' We prayed and wept and prayed and wept, and God heard us. He had pity on us, I suppose. . . . At the very minute when the innkeeper had taken the merchant by the beard to rip open his throat with his knife suddenly someone seemed to tap at the window from the yard! We all started, and the innkeeper's hands dropped. . . . Someone was tapping at the window and shouting: 'Pyotr Grigoritch,' he shouted, 'are you here? Get ready and let's go!' The people saw that someone had come for the merchant; they were terrified and took to their heels. . . . And we made haste into the yard, harnessed the horses, and were out of sight in a minute. . ."

Figure 2: An example paragraph that contains multiple speakers from *The Steppe*

Figure 2 shows a screen shot of our annotation tool displaying a paragraph with a complex conversational structure from *The Steppe*.

A.2 Vocative Patterns

Pattern	Example
between , and !	, Nastasya!
between , and ?	, Mr. Bennet?
between , and .	, Yegorushka.
between , and ;	, papa;
between , and ,	, Emma,
between " and ,	"Father Christopher,
between , and "	, mother"
after the word "dear"	Dear Lydia
between "oh" and !	Oh Henry!

Table 10: Vocative patterns for extracting mentions.

A.3 Supervised Classifier Features

We used the following features in our supervised classifier:

- *Distance*: token distance, ranked distance (relative to mentions), paragraph distance (left paragraph and right paragraph separate)
- *Mention*: Number of quotes in the mention paragraph, number of words in mention paragraph, the order of the mention within the paragraph (compared to other mentions), whether the mention is within conversation (i.e. no non-quote text in the same paragraph), whether the mention is within a quote, POS of the previous and next words.
- *Quote*: the length of the quote, the order of the quote (i.e. whether it is the first or second quote in a paragraph), the number of words in the paragraph, number of names in the paragraph, whether the quote contains text in it, whether the present quote contains the name of the mention (if mention is a name).

A.4 Words Lists

Common Speech Verbs Similar to He et al. (2013), we use *say, cry, reply, add, think, observe, call,* and *answer*, present in the training data from *Pride and Prejudice*.

Family Relation Nouns ancestor aunt bride bridegroom brother brother-in-law child children dad daddy daughter daughter-in-law father father-in-law fiancee grampa gramps grandchild grandchildren granddaughter grandfather grandma grandmother grandpa grandparent grandson granny great-granddaughter great-grandfather great-grandmother great-grandparent great-grandson great-aunt great-uncle groom half-brother half-sister heir heiress husband ma mama mom mommy mother mother-in-law nana nephew niece pa papa parent pop second cousin sister sister-in-law son son-in-law stepbrother stepchild stepchildren stepdad stepdaughter stepfather stepmom stepmother stepsister stepson uncle wife

A.5 Annotation Guidelines

- Each quote should be annotated with the character that is that quote's speaker.
- Each quote should be linked to a mention that is the most obvious indication of that quote's speaker.
  - Quotes can be linked to mentions inside other quotes.
  - Multiple quotes may be linked to the same mention.
- Mentions should also be annotated with the character that they refer to.
  - If a character's name is meaningfully associated with an article (e.g. "...," said *the Bear*), include that article in the mention.