First Steps Towards a Source Recommendation Engine: Investigating How Sources Are Used in News Articles

Alexander Spangher, Jonathan May
spangher@usc.edu
Univ. Southern Cal., Information Sciences Institute
Los Angeles, USA

James Youn, Nanyun Peng
Univ. California, Los Angeles
Los Angeles, USA

ABSTRACT

A source-recommendation tool to surface useful sources to journalists could save journalists time and diversify the breadth of sources considered. However, to build an effective service, we must first understand journalists' needs: how and why sources are used today.

We take steps towards this goal by building effective *source attribution* models that can reliably extract a broad variety of sources (i.e. people, documents, databases, etc.) from news articles based on the linguistic patterns associated with their use. We construct a large annotated training dataset and show that models trained on this dataset out-compete previous approaches in the literature. We use these models to audit articles from major news outlets (e.g. *New York Times*, *BBC* and others). We find, for instance, that on average, 50% of sentences in these articles have attributable sources.

Finally we show that there are patterns to the way sources are used in news writing by showing, via two experiments, that we can predict when sources need to be added to a news article. We hope in *future* work to explain these predictions, to study why different *types* of sources are used together, and ultimately how to recommend them for journalists.

1 INTRODUCTION

Journalism informs our worldviews, and articles themselves are informed by various informational sources. Sources tend to be used together in canonical ways: articles covering local crime, for instance, will likely include quotes from both a victim and a police officer [17, 20], and articles covering political debates will include voices from both political parties [4].

Patterns of source-usage beats has not previously been modeled. Previous work, we show, has extracted quotes[1] from news articles [7, 13] with high-precision, low recall. Such work can analyze *aggregate* quote patterns *across* documents [12, 21] but provides little reliable insight into why sources are used together *within* a story. Such insights are crucial for building effective source recommendation engines for journalists.

In this work, we take steps towards a source-recommendation engine by *(1) providing tools to understand which information is attributable to sources in news articles and (2) showing that sources are used in a predictable way*. We build, to-date, the largest annotated dataset, to our knowledge, of sources in news articles, with 1,304 articles. *A source is a person, document or database which provides information* ***directly*** *to a journalist*[2]. We introduce 16 categories of sourcing (some shown in Tables 1 and 2). We use this dataset to train strong models for *source attribution*, achieving an overall attribution accuracy of 83% using GPT3 6.7B[3]. Additionally, we test numerous baselines and show that previous lexical approaches [7], bootstrapping [14], and distant-supervision [21] underperform. Finally, our work is the first to show that sources complement each other in two novel experiments. We hope our work leads to future work analyzing *types* of sources used together and predicting specific sources needed by journalists.

Sentence	
Prime Minister **Laurent Lamothe** announced his resignation.	← *from* Statement
The announcement followed a corruption **commission**'s report.	← *from* Report
"There was no partisan intereference" said the **commission**.	← *from* Quote
However, curfews were imposed in cities in anticipation of protests.	← *from* Order
It remains to be seen whether the opposition will coalesce around a new candidate.	

Table 1: Different informational sources used to compose a single news article. Source attributions shown in bold. Some sources may be implicit (e.g. 4th sent.) or too ambiguous (last sent.). Information types used by journalists are shown on the right. Our central question: *does this article need another source?*

2 PROBLEM FORMULATION

We model a news article as a list of sentences where each sentence can be attributed to zero, one or more sources. We wish to perform two tasks: (1) *source attribution*, where we attribute information in each sentence to a source and (2) *source prediction*, where we predict if a document needs another source.

2.1 Source Attribution

A sentence *is* attributable to a source if there is an *explicit* or *implicit*[4] indication that the facts in it came from that source. A sentence *is not* attributable if the sentence does not convey concrete facts (i.e. it conveys analysis, speculation, or context provided by the journalist), or if it cannot be determined where the facts originated. We allow for

[1] A *quote* is verbatim or paraphrased information from a person or a document. *Sourced information* is broader and includes actions by the journalist to uncover information: first-person observations, analyses or experiments.

[2] For example, the source for the following sentence in a news article:"'A perp walk would be great', said Trump, as reported in the *New York Times*." would be the *New York Times*, not *Donald Trump*. They may be named entities (e.g. "Laurent Lamothe," in Table 1), or canonical indicators (e..g "authorities") and they are *not* pronouns.

[3] GPT3 sizes: https://blog.eleuther.ai/gpt3-model-sizes/

[4] In some cases, a sentence's source is not mentioned in the article but can still be determined if (1) the information can only have come from a small number of commonly-used sources[5] (2) the information is based on an eye-witness account by the journalist.

Information Channel	Num. Sentences
No Quote	23614
Direct Quote	7928
Indirect Quote	6564
Background/Narrative	3818
Statement/Public Speech	3280
Published Work/Press Report	2730
Email/Social Media Post	1352
Proposal/Order/Law	896
Court Proceeding	540
Direct Observation	302
Other	610

Table 2: Prevalence of different information channels.

an expansive set of information channels to be considered (see Table 2 for some of the top channels) and design a set of 16 canonical informational categories that journalists rely on.[6]

2.2 Source Prediction

Our original goal: *can we effectively recommend sources to journalists?* would be a hopeless task if sources are combined randomly in news writing. One way to explore whether this is true or not is to ask: can we predict if an article is *missing* sources?

We create two binary classification tasks to probe this question:

(1) *Ablation*: Choose one source in an article. To generate $y = 1$ examples, remove all sentences attributable to that source. To generate $y = 0$ examples, remove an equal number of sentences attributable to no source.
(2) *NewsEdits*: Sample article-versions from the [19] *NewsEdits* corpus, which is a corpus of news articles along with their updates. Identify articles at time t where the update at time $t + 1$ either adds a source ($y = 1$) or does not ($y = 0$).

Ablation assumes that the composition of sources in an article is cohesively balanced, and induces reasoning about this balance. *NewsEdits* relaxes this assumption and probes if this composition might change, either due to the article's completeness, changing world events that necessitate new sources, or some other factor[7].

3 CORPUS CREATION AND ANNOTATION

We select 1,304 articles from the *NewsEdits* corpus [19] and deduplicate across versions. We recruit two annotators to annotate sources. One annotator is a trained journalist with over 4 years of experience working in a major newsroom, and the other is a undergraduate assistant. The senior annotator checks and mentors the junior annotator until they have a high agreement rate. Then, they collectively annotate 1,304 articles including 50 articles jointly. From these 50, we calculate an agreement rate of more than $\kappa = .82$ for source attribution and $\kappa = .45$ for quote-type categorization. Categories shown in Table 2 are developed early in the annotation process and expanded until a reasonable set captures all further observations. Categories are also refined and adjusted following conversations with experienced journalists and professors. For a full list of categories, see appendix. ***Note:*** *we do not perform modeling on these categories in the present work, but use them for illustration and evaluation.*

4 SOURCE ATTRIBUTION RESULTS

We split Source Attribution into two steps: *detection* (*is* the sentence attributable?) and *retrieval* (what is that attribution?) because using different models for each step is more effective than modeling both jointly. Additionally, we test a number of baselines. *Each baseline performs detection and retrieval in one step: so, for detection evaluation comparison we simply ask whether they attributed any source to a sentence.* Since *detection* is a binary classification task, F1-score is used to measure. We use accuracy, or retrieval with $k = 1$ for *retrieval*. Shown in Table 3 is a sample of our results.

Baseline Methods

R1, Co-Occurrence: We identify sentences where a source co-occurs with verbs indicated "speaking" (using a list of 538 verbs [15], PERSON Named Entities and specific 301 noun-phrases, e.g. "authorities"[8]). We group entities by last name.

R2, Governance: We expand on R1 and process using syntactic dependency parsing [9] to introduce additional heuristics.[8]

Quootstrap: [14] We use a set of 1,000 lexical patterns provided by the authors and identify all sentences that match these. [9]

QuoteBank: We match articles in our annotation set with articles processed and released by [21]. We find 139 articles.[10].

Detection Methods

Sentence: We adapt a binary sentence classifier where each token in each sentence is embedding using the `BigBird-base` transformer architecture [22]. Tokens are combined via self attention to yield a sentence embedding, which is fed into a binary classification layer. Thus, each sentence is independent of others.

Full-Doc: We use a similar architecture, but instead of embedding tokens in each sentence separately, we embed tokens in the whole document, then split into sentences.

Retrieval Methods

Sequence Labeling: predicts whether each token in a document is a source-token or not. We pass each document through a `BigBird-base` to obtain token embeddings and then use a token-level classifier.

Span Detection: predicts start and stop tokens of the sentence's source. We use `BigBird-base`, and separate start/stop-token classifiers [1]. We experiment with inducing decaying reward around start/stop positions to reward near-misses, and expand the objective to induce source salience as in [6], but find no improvement.

Generation: We formulate retrieval as open-ended generation and fine-tune GPT3 models to generate source-names[11].

For *+coref* variations, we evaluate approaches on articles after resolving all coreferences using [10]. For *+Nones* variations, we

[6]These 16 categories are formulated both in conversation with journalists and after extensive annotation and schema expansion.
[7][19] found that many news updates were factual and tied to event changes.
[8]Specifically, we identify sentences where the name is an *nsubj* dependency to a speaking verb governor. *nsubj* is a grammatical part-of-speech, and a governor is a higher node in a syntactic parse tree.
[9]Authors created a bootstrapping algorithm to discover lexical patterns indicative of sourcing. The relatively small size of our dataset compared with theirs prevents us from using this architecture to extract meaningful patterns from our dataset.
[10]We also discard articles where *QuoteBank* reported quotations or context that are not found in our articles, because our corpus was created from *NewsEdits*, so it's possible that the version of the articles that we examined were different from theirs.
[11]We prompts with "`<article>To which source can we attribute the sentence <sentence>?`".

			All	Direct Quote	Indirect Quote	Statement/ Speech	Email/ Social	Published Work/ Press	Other
Detection	f1 score	Rules 1	59.1	64.7	69.3	81.2	76.2	72.7	37.4
		Rules 2	68.8	71.3	79.8	89.8	82.1	79.2	32.5
		Quootstrap	33.4	85.0	81.3	51.3	58.6	33.1	3.0
		Sentence	87.1	91.0	98.7	94.1	**92.7**	85.4	61.4
		Full-Seq	**88.2**	**92.0**	**98.7**	**96.4**	**89.8**	**86.4**	**65.1**
Retrieval	accuracy on gold-labeled sourced sents	Rules 1 (+*coref*)	46.4 (52.8)	47.8 (57.3)	48.4 (54.5)	43.0 (49.8)	51.7 (49.4)	37.8 (38.3)	30.2 (34.9)
		Rules 2 (+*coref*)	22.5 (36.6)	20.7 (31.6)	22.5 (42.0)	30.3 (56.1)	21.3 (30.3)	27.4 (32.3)	30.2 (30.2)
		QuoteBank	5.5	9.9	16.0	16.4	17.7	4.3	0.5
		SeqLabel	38.5	37.2	43.4	40.0	31.2	32.3	17.7
		SpanDetect (+*coref*)	59.5 (53.6)	61.1 (51.2)	59.5 (56.8)	67.6 (60.6)	44.4 (79.0)	51.6 (54.6)	36.5 (42.6)
		GPT3 1.3B (+*coref*)	78.9 (73.2)	80.9 (78.7)	86.9 (82.5)	85.0 (76.3)	71.9 (56.1)	57.9 (54.4)	38.3 (31.2)
		GPT3 6.7B	**91.4**	**94.0**	**95.5**	**91.1**	**91.0**	**81.6**	**57.3**
Both	acc. all sents	GPT3 1.3B (+*Nones*)	70.9 (73.1)	79.5 (82.4)	82.9 (84.8)	82.9 (85.9)	73.4 (73.4)	60.5 (61.0)	53.0 (64.5)
		GPT3 6.7B (+*Nones*)	80.0 (**83.0**)	90.4 (**92.3**)	90.7 (**92.9**)	89.9 (**92.9**)	91.1 (**91.0**)	78.0 (**78.2**)	**68.9** (68.3)

Table 3: *Source Detection* F1 scores (top), measured as correctly identifying source sentences, *Source Retrieval* accuracy (middle), measured as the % of known source-sentenced that are correctly labeled, and *Pipeline* accuracy (bottom), measured as % of all sentenced correctly attributed. In *Both*, we use the top-performing quote-detection module to identify quotes, then perform retrieval. *Takeaway: We can attribute sources with accuracy* > 80%.

	Gold (Train)	Gold (Test)	Silver
# docs	1032	272	9051
# sent / doc	30	67.5	27
doc len (chars)	3952	7885	3984
# sources / doc	6.8	12.1	8.2
% sents sourced	47.7%	46.9%	57.4%
% source-sents, top	37.5%	28.1%	31.8%
% source-sents, bot	5.9%	2.4%	6.7%
source entropy	1.6	2.1	1.8
∇ sources / version	n/a	n/a	+2
most sourced sent	96th p	92th p	0th p

Table 4: Corpus-level statistics for our training, test, and silver-standard datasets. Shown are averages across the entire corpus. Documents in the test set are longer than the training, but the model seems to generalize well to the silver-standard corpus, as statistics match. "Source-sents, top" and "Source-sents, bot" refer to the % of *sourced* sentences attributed to the most and least used sources in a story. "most sourced sent" refers to the sentence with the most likelihood of being sourced in the document (as a percentile of doc. length)

additionally train our models to detect when sentences do *not* contain sources. We use this as a further corrective to eliminate false positives introduced during detection.

4.1 Results and Discussion

As shown in Table 3, we find that the GPT3 6.7B retrieval model paired with the Full-Seq detection module in a pipeline performed best, achieving an overall attribution accuracy of 83%. In the +*None* setting, both GPT3 1.3B and 6.7B are used to identify false positives introduced by the detection stage and outperform their counterparts. Overall, we find that resolving coreference does not improve performance.The poor performance of both rules-based approaches and QuoteBank, which also uses heuristics,[12] indicates that this task is more complex than simple lexical cues.

5 INSIGHTS FROM SOURCE ANALYSIS

Having built an attribution pipeline that performs reasonably well, we wish to derive insights into how sources are used in news articles. We further sample 9051 unlabeled documents from *NewsEdits* and use our best-performing attribution model to extract all sources. *We ask two questions: how much an article is sourced? When do sources get used in the reporting and writing process?* We report our statistics in Table 4 (more detailed analysis in the appendix.)

Insight #1: ∼ 50% *of sentences are sourced, and sources are used unevenly.* Most articles, we find, attribute roughly half the information in their sentences to sources. This the percentage of sources used is fairly consistent between longer and shorter documents. So, as a document grows, it adds roughly an equal amount of sourced and unsourced content (e.g. explanations, analysis, predictions, etc.).[13] We also find that sources are used unevenly. The most-used source in each article usually contributes ∼ 35% of sourced sentences, whereas the least-used source contributes ∼ 5%. This shows a hierarchy between major and minor sources used in reporting and suggests future work analysing the differences between these sources.

Insight #2: Sources begin and end documents, and are added while reporting. Next we examine when sources are used in the reporting process. We find that articles early in their publication cycle tend to have fewer sources, and add on average two sources per subsequent version. This indicates an avenue of future work: understanding which kinds of sources get added in later versions can help us recommend sources as the journalist is writing. Finally, we

[12] Quotebank's algorithm condenses input data to a BERT span-classifier by (1) looking for double-quotes (2) identifying candidate speakers through a lookup table.
[13] For more details, see the appendix.

also find, in terms of narrative structure, that journalists tend to lead their stories with sourced information: the most likely position for a source is the first sentence, the least likely position is the second. The second-most likely position is the end of the document.[14]

6 SOURCE PREDICTION

We wish to examine whether there is a pattern to the way sources are used together in news reporting, so we design a task called *source prediction*. We outlined two approaches to this task in Section 2.2.

6.1 Task Dataset Creation

Ablation. We take the 9051 silver-standard documents explored in the previous section and design three variations of this task. As shown in Table 4, articles tend to use sources lopsidedly: one source is usually primary. Thus, we design Easy (Top Source, in Table 1), Medium (Secondary) and Hard (Any Source) variations of our task. For Easy, we choose the source with the most sentences attributed to it. For Medium, we randomly choose among the top 3 sources. And for Hard, we randomly choose any of the sources. Then, we create a $y = 1$ example by removing all sentences attributed to the chosen source, and a $y = 0$ example from the same document by removing an equal number of sentences not attributed to any sources.

NewsEdits. We sample an additional $40,000$ articles from the *NewsEdits* corpora and perform *attribution* on them. We sample versions pairs that have roughly the same number of added, deleted and edited sentences in between versions in order to reduce possible confounders, as [19] showed that these edit-operations were predictable. We identify article-version pairs where 2 or more sources were added between version t and $t + 1$ and label these as $y = 1$, and 0 or 1 sources added as $y = 0$.

6.2 Modeling

We use three models: (1) FastText [5] for sentence classification, (2) A BigBird-based model: we use BigBird with self-attention for document classification, similar to [19].[15] Finally, (3) we fine-tune GPT3 1.3 to perform prompt-completion for binary classification.

For each model, we test two setups. First, we train on the vanilla text of the document. Then, in the *+source* variants, we train by appending each sentence's source *attribution* to the end of it.[16] The source annotations are obtained from our attribution pipeline.

To further make sense of our results, we train a classifier to identify four reporting topics plus one general topic[17]. We identify articles in the *New York Times Annotated Corpus* [16] with keyword sets corresponding to each topic (or "all" for Other News). Using these as distant supervision, we train a FastText classifier to output one of these 5 categories.

[14]The sources might be used in for different purposes: [18] performed an analysis on news articles' narrative structure, and found that sentences conveying the *Main Idea* lead the article while sentences conveying *Evaluations* or *Predictions*.

[15]Concretely, we obtain token embeddings of the entire document, which we combine for each sentence using self-attention. We contextualize each sentence embedding using a shallow transformer architecture. We finally combine these sentence embeddings using another self-attention layer to obtain a document embedding for classification. We utilize curriculum learning based on document length, a linear loss-decay schedule.

[16]Like so: `<sent 1>. SOURCE: <source 1>. <sent 2> SOURCE: <source 2>... <sent n> SOURCE: <source n>.`

[17]These four have been identified as especially socially valuable topics, or "beats," due to their impact on government responsiveness [3]

6.3 Results and Discussion

Our results are shown in Table 5. Overall, we find that our experiments are statistically significant with t-test $p < .01$. However, statistical significance does not preclude confounding, and both the *Ablation* and the *NewsEdits* setups contain possible confounders.

In the *Ablation* set up, we might be inadvertently learning stylistic differences rather than source-based differences. To address this, we investigate (1) whether lexical confounders, such as speaking verbs, might be artificially removed in the ablated documents: they are not[18] (2) whether statistically significant differences between counts of named entities or source signifiers (defined in Section 4) exist: they do not and (3) we create secondary test sets where $y = 0$ is *non-ablated* documents. This changes the nature of the stylistic differences between $y = 1$ and $y = 0$ while not affecting sourcing differences[19]. We find that under this setting, the accuracy of our classifiers differs by within 3 points.

In the *NewsEdits* setup, we have taken care to balance our dataset along axes where prior work have found predictability.[20] We balance for length, version number and edit operations.

Having attempted to address confounding in various ways in both experiments, we have more confidence in our conclusions that sources are chosen to complement each other. To illustrate, consider Table 5, where Election coverage is the most easily predictable across all tasks. This might be because of efforts to include both left-wing and right-wing voices. It also might be because the cast of characters (e.g. campaign strategists, volunteers, voters) stays relatively consistent across stories.

Two additional findings are that (1) harder tasks yield lower accuracies and, (2) larger GPT3-based language models generally perform better. Although not especially surprising, this further confirms our intuitions about what these tasks are probing. We were surprised to find that, in general, adding additional information in both stages of this project (i.e. coreference in the *attribution* stage or source information in the *prediction* stage), did not improve the models' performance [21]. We had hypothesized that the signal introduced would not harm the GPT3-based models, but this was untrue. It could be that the larger models are already incorporating a notion of coreference and attribution, and our method of adding this information changed English grammar in a way that harmed performance.

7 RELATED WORK

Prior work in quote attribution has been aimed at identifying direct and indirect quotes in news articles. Early work explored rules-based methods [2, 11] and statistical classifiers [13] to attribute sources to quotes. More recent work has extended these ideas by using bootstrapping to discover new patterns [14] and perturbations on these patterns to generalize to larger language models [21]. Such works focuses on a narrow set of sources: namely, quotes given by

[18]We use lexicons defined in our rules-based methods to measure the number of speaking verbs in our dataset. We find a mean of $n = [34, 32]$ speaking verbs per document in $y = [0, 1]$ classes in the Top case, $n = [35, 34]$ in the Medium, and $n = [35, 37]$ in Hard. None of these differences are statistically significant.

[19]We do not want to *train* on such datasets, because there are statistically significant length differences and other stylistic concerns ablated and non-ablated articles.

[20]For instance, [19] found that whether a sentence would be added or removed between versions could be predicted.

[21]In contrast, adding source information to smaller language model, `BigBird`, helped with harder tasks like the Medium, Hard and *NewsEdits*).

		Other News	Disaster	Elections	Labor	Safety
Top Sour. Ablated	FastText (+*source*)	66.1 (66.0)	65.8 (64.5)	69.8 (69.8)	68.8 (68.2)	68.0 (68.0)
	BigBird (+*source*)	74.2 (73.9)	68.4 (69.7)	78.3 (74.9)	**74.0** (73.4)	78.1 (73.4)
	GPT3 1.3B (+*source*)	**78.3** (74.9)	**75.5** (69.5)	**81.5** (78.0)	72.7 (70.9)	**80.0** (65.1)
Secondary Source	FastText (+*source*)	57.6 (57.8)	63.2 (63.2)	60.8 (61.1)	61.0 (62.3)	63.3 (64.1)
	BigBird (+*source*)	63.8 (65.1)	61.8 (**69.7**)	63.1 (65.7)	64.3 (64.9)	61.7 (62.5)
	GPT3 1.3B (+*source*)	**67.1** (65.4)	67.9 (65.1)	**72.9** (68.0)	58.8 (**65.9**)	65.6 (**66.7**)
Any Source	FastText (+*source*)	54.5 (54.8)	60.5 (**59.2**)	57.1 (57.6)	57.8 (56.5)	56.2 (56.2)
	BigBird (+*source*)	57.5 (**59.4**)	53.9 (55.3)	55.5 (60.6)	55.8 (60.4)	**57.8** (56.2)
	GPT3 1.3B (+*source*)	55.0 (59.0)	53.9 (56.1)	**63.6** (61.3)	**63.4** (39.3)	49.0 (51.7)
News Edits	FastText (+*source*)	58.1 (56.8)	48.9 (55.8)	62.1 (61.9)	58.6 (61.2)	48.8 (49.6)
	BigBird (+*source*)	63.5 (**69.4**)	63.9 (**65.3**)	64.5 (62.6)	**64.8** (60.4)	64.8 (**64.2**)
	GPT3 1.3B (+*source*)	65.0 (64.0)	63.9 (56.1)	**64.6** (61.3)	62.4 (39.3)	51.0 (51.7)

Table 5: Results for *source prediction*, broken into four canonical news topics and 'other.' The "Top Source Ablated" category (top grouping) is our prediction task run on articles ablated by removing the source that has the most sentences, the "Secondary Source Ablated" category (second grouping) is where a source contributing more than 10% of sentences is removed, and the "Any Source Ablated" category (third grouping) is where any source is randomly removed. The *NewsEdits* task (bottom grouping) is to predict whether the article at time t will be added sources at time $t+1$. *Takeaway: On all of these tasks, our models were able to significantly outperform random, or 50% accuracy. In general, our expectations are confirmed that: (a) harder tasks yield lower-accuracy results and (b) more powerful models improve performance. This indicates that there is a pattern to the way sources are included in news writing.*

people, rather than our more expansive set of informational sources. Surprisingly, we observe low performance from QuoteBank, even in categories it is trained to detect.

8 CONCLUSIONS

We have offered a more expansive definition sourcing in journalism and introduced the largest attribution dataset capturing this notion. We have developed strong models to identify and attribute information in news articles. We have used these attribution models to create a large silver standard dataset that we used to probe whether source inclusion in news writing follows predictable patterns.

We have future work planned to identifying groups of source types and to ultimately use these insights to build a source recommendation engine.

REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[2] David K Elson and Kathleen R McKeown. 2010. Automatic attribution of quoted speech in literary narrative. In *Twenty-fourth AAAI conference on artificial intelligence*.

[3] James T Hamilton. 2011. All the news that's fit to sell. In *All the News That's Fit to Sell*. Princeton University Press.

[4] Tiancheng Hu, Manoel Horta Ribeiro, Robert West, and Andreas Spitz. 2022. Quotatives Indicate Decline in Objectivity in US Political News. *arXiv preprint arXiv:2210.15476* (2022).

[5] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of Tricks for Efficient Text Classification. *arXiv preprint arXiv:1607.01759* (2016).

[6] Yuval Kirstain, Ori Ram, and Omer Levy. 2021. Coreference Resolution without Span Representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. 14–19.

[7] Grace Muzny, Michael Fang, Angel Chang, and Dan Jurafsky. 2017. A two-stage sieve approach for quote attribution. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. 460–470.

[8] Edward Newell, Drew Margolin, and Derek Ruths. 2018. An attribution relations corpus for political news. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

[9] Joakim Nivre. 2010. Dependency parsing. *Language and Linguistics Compass* 4, 3 (2010), 138–152.

[10] Shon Otmazgin, Arie Cattan, and Yoav Goldberg. 2022. LingMess: Linguistically Informed Multi Expert Scorers for Coreference Resolution. *arXiv preprint arXiv:2205.12644* (2022).

[11] Tim O'Keefe, Silvia Pareti, James R Curran, Irena Koprinska, and Matthew Honnibal. 2012. A sequence labelling approach to quote attribution. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. 790–799.

[12] Sebastian Padó, André Blessing, Nico Blokker, Erenay Dayanık, Sebastian Haunss, and Jonas Kuhn. 2019. Who sides with whom? towards computational construction of discourse networks for political debates. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2841–2847.

[13] Silvia Pareti, Tim O'keefe, Ioannis Konstas, James R Curran, and Irena Koprinska. 2013. Automatically detecting and attributing indirect quotations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 989–999.

[14] Dario Pavllo, Tiziano Piccardi, and Robert West. 2018. Quootstrap: Scalable unsupervised extraction of quotation-speaker pairs from large news corpora via bootstrapping. In *Twelfth International AAAI Conference on Web and Social Media*.

[15] Jeroen Peperkamp and Bettina Berendt. 2018. Diversity Checker: Toward recommendations for improving journalism with respect to diversity. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization*. 35–41.

[16] Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia* 6, 12 (2008), e26752.

[17] Alexander Spangher and Divya Choudhary. 2022. If it Bleeds, it Leads: A Computational Approach to Covering Crime in Los Angeles. *arXiv preprint arXiv:2206.07115* (2022).

[18] Alexander Spangher, Xinyu Hua, Yao Ming, and Nanyun Peng. 2023. Sequentially Controlled Text Generation. *arXiv preprint arXiv:2301.02299* (2023).

[19] Alexander Spangher, Xiang Ren, Jonathan May, and Nanyun Peng. 2022. NewsEdits: A News Article Revision Dataset and a Novel Document-Level Reasoning Challenge. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 127–157.

[20] Kobie Van Krieken. 2022. Story character, news source or vox pop? Representations and roles of citizen perspectives in crime news narratives. *Journalism* 23, 9 (2022), 1975–1994.

[21] Timote Vaucher, Andreas Spitz, Michele Catasta, and Robert West. 2021. Quotebank: a corpus of quotations from a decade of news. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 328–336.

[22] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems* 33 (2020), 17283–17297.