An Attribution Relations Corpus for Political News

Edward Newell, Drew Margolin, Derek Ruths
edward.newell@mail.mcgill.ca, dm658@corenell.edu, derek.ruths@mcgill.ca
McGill University, Cornell University, McGill University

Abstract

An attribution occurs when an author quotes, paraphrases, or describes the statements and private states of a third party. Journalists use attribution to report statements and attitudes of public figures, organizations, and ordinary individuals. Properly recognizing attributions in context is an essential aspect of natural language understanding and implicated in many NLP tasks, but current resources are limited in size and completeness. We introduce the Political News Attribution Relations Corpus 2016 (PolNeAR)[2]—the largest, most complete attribution relations corpus to date. This dataset greatly increases the volume of high-quality attribution annotations, addresses shortcomings of existing resources, and expands the diversity of publishers sourced. PolNeAR is built on news articles covering the political candidates during the year leading up to US Presidential Election in November of 2016. The dataset will support the creation of sophisticated end-to-end solutions for attribution extraction and invite interdisciplinary collaboration between the NLP, communications, political science, and journalism communities. Along with the dataset we contribute revised guidelines aimed at improving clarity and consistency in the annotation task, and an annotation interface specially adapted to the task, for reproduction or extension of this work[2].

Keywords: attribution, quotation, sourcing, corpus, news, journalism, politics

1. Introduction

Attribution occurs when an author describes a propositional attitude (Russel, 1940) held by some third party: an agent's statements, intentions, beliefs, knowledge, perceptions, decrees, or sentiments about something (see **Table 1**). Quoting or paraphrasing another person is a familiar form of attribution. But self-attribution, and attribution to artifacts like reports, recordings, or databases are included in the definition.

Prior work defines attributions as consisting of three parts: (1) the source, to whom content is attributed; (2) the content that is attributed; and (3) a cue phrase used to signal attribution, such as "said" or "according to" (Pareti, 2012).

While the statements made by public figures are often inherently newsworthy, attribution to any source is a fundamental mechanism for journalists to lend credibility or authority to, or to nuance, an assertion. Accurate and informative attributions provide readers with transparency and accountability by making journalists' sourcing identifiable (Esser and Umbricht, 2014).

Lately, there has been increased skepticism directed at the mainstream media[1], with questions about the legitimacy of reporting often focusing on sourcing. Thus, attribution phenomena are of fundamental interest for the maintenance of journalistic standards, and should be carefully attended to by the critical reader.

Although attribution is a fundamental rhetorical mechanism in media, it has not received much attention from computational researchers, and partly as a result, datasets for its study are limited in number and depth. Recently, however, direct efforts have been made to study attribution, with the creation of PARC3 (Pareti, 2012), the largest attribution-relations dataset prior to PolNeAR. Despite its important contribution, PARC3 suffers from low annotator recall—many attributions have gone unnoticed by annotators. This impedes the creation of attribution models. In this study we take measures to improve recall, and describe an annotated corpus that doubles the rate of recall of PARC3 and improves inter-annotator agreement[2].

2. Existing resources

The phenomenon of attribution has not typically been studied directly, but instead as part of other phenomena such as opinion analysis, discourse analysis, or the analysis of dialogue in narrative. As a result, there are many existing resources that contain annotations relevant to attribution but which fail to fully capture all attributive phenomena or to adequately label all parts of attributions.

In some cases corpora dedicated to event detection and extraction overlap with attribution, where attribution is seen as a type of event. For example, TimeBank (Pustejovsky et al., 2003) annotates various kinds of events that correspond to attribution. However, because the focus is on events, attributions that do not meet the criteria for being events are not annotated.

Corpora dedicated to opinion analysis and extraction often annotate attribution relations, with Evans et al. (2007) providing an example in English, and Li et al. (2012) providing an example in German. One English resource in particular (Wiebe et al., 2005) provides a corpus of 692 news articles annotated with expressions of speech acts (direct and indirect quotes) and internal states, along with annotations that capture the main components of interest in to the study of attribution.

Various corpora dedicated to discourse analysis and discourse parsing exist, and include annotations of attribution phenomena, including the RST Discourse TreeBank (Carlson et al., 2002), GraphBank (Wolf and Gibson, 2005), with the largest being the Penn Discourse TreeBank 2 (PDTB2) (Prasad et al., 2007). While these corpora do annotate attribution relations, they are designed to annotate all rhetorical

[1]For example theguardian.com/media/2016/dec/18/what-is-fake-news-pizzagate

[2]Dataset: https://github.com/networkdynamics/PolNeAR
Annotation interface: https://github.com/networkdynamics/brat-attribution-annotation

Type	Example
statement	**Jimroy** said *"Sally can't stay."*
intention	**Sally** plans *to run for president*.
decree	**Judge Thomson** issued an injunction *blocking a bill as unconstitutional*.
knowledge	**I** don't know *why they don't pass the bill*.
belief	**He** still thinks *they're on his side.*
perception	**Jimroy** sees *a trend emerging*.
sentiment	**She** disapproved of *the deal*.

Table 1: Basic examples of various forms of attribution, presented in subject-verb-object form for easy comparison. Typesetting indicates the **sources**, cues, and *content*.

relations. Because these corpora are not specific to attribution, they do not indicate key elements of attributions, such as the source. They also tend not to annotate all attributions, missing the attribution of intentions, attributions with implicit sources, and attributions scoped as hypothetical.

Only a limited number of corpora have been designed to address attribution specifically. Elson and McKeown (2010) contribute a corpus of 3176 direct quotes linked to their sources in narrative text. This effort successfully links sources to content, but does not enable the investigation of more challenging cases, such as the attribution of indirect quotes and private states. O'Keefe et al. (2012) provided a corpus of 965 documents from Sydney Morning Herald which annotates direct quotes and their sources, and was later augmented to include indirect quotation. But again internal states are not included in this corpus.

Starting from PDTB2, Pareti (2012) created the PDTB Attribution Relations Corpus 3 (PARC3). This adds specific labels for not just source and content spans, but also *cues*, and annotates attribution relations originally missing from PDTB2. Before PolNeAR, PARC3 is the largest corpus of attribution relations, and the only one which annotates all types of propositional attitude and all three attribution components.

3. Addressing Limitations of Existing Resources

While PARC3 is a valuable resource, it has important limitations. Referring apparently to an earlier version of the dataset, the creators of PARC3 acknowledge that between 30% and 50% of attributions remain unlabelled (Pareti et al., 2013). In the final version of the dataset, there is only one attribution per 69 words, compared to one attribution every 32 words in PolNeAR (see **Table 4**). To compare annotation procedures, we re-annotated 56 randomly selected articles form PARC3, and found that the majority of attributions are un-annotated. Extrapolating the rate of "missed" attributions to the full dataset suggests more than 20 thousand attributions are missing in PARC3.

The creators of PARC3 attempt to mitigate the impact of missing annotations on attribution extraction models by detecting sentences containing words commonly used as cues, and eliminating those sentences that contain no attribution (Pareti et al., 2013). This complicates training. It also fails to address the fact that attributions with less common cues—which are less easily found by annotators—will remain as false-negatives.

We can expect a good extraction model to learn the same biases as, and miss the kinds of attributions missed by trained coders. Less obviously, models so-trained will be biased to over-annotate near common cue words, having never seen a sentence that contains such a word being used outside of the context of attribution. Also, patterns at the inter-sentential level contribute to successful source prediction (Elson and McKeown, 2010). Eliminating whole sentences will distort these patterns relative to raw text.

Improving annotator recall is thus of critical importance. We apply three tactics to increase it:

1. During pilot studies we collect and characterize "tough" attributions—ones which many annotators miss, and ones which annotators find ambiguous under the guidelines. We use these during training and as a reference during annotation, and we clarify the annotation guidelines with respect to those attributions.

2. We emphasize the importance of recall at each (weekly) quality control meeting with annotators. To do so (while keeping balanced precision), we visualize text alignments that show how all annotators handled particular attributions on test articles, allowing annotators to see their disagreements.

3. We allow annotators to indicate uncertainty about an attribution using a "discuss" flag. We encourage annotators to flag ambiguous cases as "discuss", so that other annotators can review it. The "discuss" flag allows us to easily inspect the annotations comprising the boundary of the annotation concept.

The success of these measures is reflected in the corpus statistics (estimated from the 32% of PolNeAR annotated at the time of writing[3]), see **Table 4**, which we discuss further in §6.

4. Corpus Curation

In addition to improving on annotation, we also take this opportunity to curate a corpus that is well-suited to studying attribution phenomena from sociological, journalistic, and political science perspectives. We focus on political news, in which attribution plays a particularly central role, and include articles from 7 publishers from across the political spectrum, and which represent traditional print news and newer online-only publishers:

[3]Finalized statistics: https://github.com/networkdynamics/PolNeAR

- New York Times
- Washington Post
- USA Today
- Breitbart
- Politico
- Huffington Post
- Western Journalism.

We focus on articles that cover the two presidential nominees, Hillary Clinton and Donald Trump, during the year of campaigning leading up to the 2016 US Presidential Election. This gives the corpus a coherent focus on a politically, socially, and journalistically important event.

To maximize the usefulness of the corpus in investigating a variety hypotheses, we selected articles for inclusion using stratified sampling. We binned articles according to the publisher, the candidate receiving the most mentions (Trump or Clinton), and divided the dataset into 12 month-long time periods. We then randomly sampled 6 articles from each bin corresponding to a specific publisher, candidate, and time period. Starting from 55,000 eligible articles, we drew 1008 articles using stratified sampling, plus an additional 20 articles for quality control purposes.

4.1. Collection

We began by obtaining all publicly available articles from each publisher having publication dates between 8 Nov 2015 to 8 Nov 2016 (election day), for a total of 404,000 articles covering a variety of topics and genres. The articles were obtained by download from LexisNexis (in the case of New York Times, Washington Post, and USA Today), or were downloaded from the publisher's website (all others).

In addition to capturing the headline and body text for each article, we captured the following metadata: the author, publication date, whether the article was produced from a newswire, and any tags or descriptors provided by the publisher (e.g. indicating that the article is editorial or news).

4.2. Processing

In extracting the headline and body text from the raw HTML obtained from LexisNexis or the respective publishers' websites, we performed several processing steps as follows.

The paragraph structure of articles was preserved using double line breaks to separate paragraphs. Any embedded tweets were normalized into a single XML tag (the only XML tag used in the corpus) which surrounds the tweet text and indicates author and timestamp as attributes (when available). Advertisements and links to related articles were removed. Blockquotes, which were often indicated in the original HTML document using CSS styles, were preserved by adding an opening quote to each block-quoted paragraph (except if already enquoted), and adding closing quotes only to the final paragraph of the blockquote, as is typical for plain-text multi-paragraph quotations. Each article includes its headline as the first paragraph, followed by the body text.

We processed the articles with Stanford's CoreNLP software (Manning et al., 2014) to provide tokenization, sentence splitting, POS tagging, constituency and dependency parsing, named entity recognition, and coreference resolution. These annotations are provided as part of the corpus, in parallel to the articles in plain text and the standoff attribution annotations.

4.3. Screening for Hard News

The collected articles spanned many genres, from hard news covering significant current events, to soft news such as celebrity gossip, as well as editorials, blog posts, and so on (Esser and Umbricht, 2014). We focused on hard news to avoid uncontrolled variations in attribution due to the stylistic differences of genres, and considering hard news to be most important from a journalistic standpoint. To select the hard news articles, we filtered articles using the metadata tags provided by the publisher which signal the topic or section from which the article was drawn. A considerable amount of effort was made to determine how articles were tagged by the publishers, and ensure that we selected hard news without arbitrarily excluding articles from the dataset. Some publishers, particularly Breitbart and Huffington Post, do not exhibit a sharp stylistic distinction between news and opinion, but we used all indicators made available by the publisher in the form of metadata and site structure.

4.4. Screening for Mentions of Political Candidates

Articles were scanned for mentions of the two presidential candidates using a combination of regular expressions and logical rules devised to provide high precision and recall in disambiguating candidate mentions. This screening procedure is described further in the supplementary material, but, for instance, it explicitly avoids high-profile false positives such as "Bill Clinton" or "Donald Trump Jr." The number of disambiguated mentions of both candidates was tallied for each article, and only those articles with at least one mention of one of the candidates were considered eligible for inclusion in the final corpus. Following this screening, we performed stratified sampling as described above.

5. Corpus Annotation

In constructing PolNeAR, we tried to adhere as closely as possible to the annotation scheme devised for PARC3, while addressing inconsistencies and ambiguities[4]. Here we describe the approach we took to refining the guidelines, and some key deviations. Other minor deviations from PARC3's guidelines are listed in the Appendix. Overall, the revisions were meant improve consistency in annotation without changing what should be considered an attribution.

5.1. Annotation Guidelines

We made significant efforts to clarify the annotation concept in pilot studies. We started by training annotators with the guidelines from PARC3, and then investigated cases of disagreement to uncover gaps and inconsistencies.

In revising the guidelines, we favored the following outcomes, by order of priority: reducing inconsistency, reducing ambiguity, and leaving rules and rulings on examples unchanged from the PARC3 guidelines.

Based on the pilot studies, we developed a set of "templates" consisting of a catalogue of examples and abstract attribution types serving to define the annotation concept

[4]Guidelines: https://github.com/networkdynamics/PolNeAR

of *attribution*. The examples were accumulated from cases generating disagreement during pilot testing. Based on the inspection of thousands of attributions, we created the "types"—abstractions that categorize the examples and illustrate commonality among them. The following types were included:

A source makes statements about,
decrees,
has knowledge of,
believes,
understands,
contemplates,
perceives,
desires,
likes or dislikes,
has an attitude about,
supports, or
has a feeling about *something*.

Many of the types relate to internal states as such examples tended to be a source of ambiguity. The types are not meant to be a mutually exclusive taxonomy. Instead, they serve as an abstract template against which annotators can match specific cases they face during annotation. Culled examples from piloting are organized under each type, so that annotators can easily compare cases they come across with several similar cases from the templates. The interested reader find the templates in the PolNeAR dataset[5].

5.2. Preference for Agentive Sources

All cases of attribution can be seen as belonging to one of two "supertypes":

1. **A communicative agent** expresses, issues an artifact expressing, or holds an internal state representing *something* (the agent may be implicit).
2. **An artifact** expresses or represents *something*.

The main difference between the two is whether content is attributed to an *agent* or an *artifact*. The following examples respectively show annotation according to the above two supertypes (using typesetting to indicate the **source**, cue, and *content*):

1. **Shell** boasted *exceptional earnings*.
2. **The press release** boasted *exceptional earnings*.

Many examples can be read as belonging to both supertypes:

1. **Shell's** press release boasted *exceptional earnings*.
2. **Shell's press release** boasted *exceptional earnings*.

In cases where both readings are possible, there is ambiguity under PARC3's guidelines. It seems more useful to take the attribution's source to be the agent that composed the content, rather than an intermediate message-bearing artifact. Therefore, we resolve the ambiguity by instructing annotators to favor annotation according to supertype (1), and use (2) only when a reading according (1) is not available.

[5]https://github.com/networkdynamics/PolNeAR

5.3. Qualified Scopes

An early question that arises in annotation is whether attributions scoped as conditional, hypothetical, uncertain, or negated, should be annotated. In keeping with PARC3, we consider a candidate attribution to be valid regardless of being under such a qualified scope[6]. For example, the following are valid attributions:

- **She** says *"Hi!"*
- If it were raining, **she** would not have hastily said *"Hi!"*

While this is done mainly for consistency with PARC3, it has the advantage of decoupling the attribution-extraction task from the task of detecting qualified scopes.

5.4. Univocality

One technically important deviation from the PARC3 annotation guidelines relates to vocality—the number of roles that a token can play in the annotation. PARC3 annotations are not univocal—within the same attribution, one token can be both part of both the source and cue.

Allowing multivocality means that the annotations can no longer be modelled directly as a sequence. This increases the space of unique annotations, and concomittantly the hypothesis space needed to fully model it.

Rather than capturing essential structure, this seems to be a quirk of the annotation guidelines. To avoid using degrees of freedom to model an annotation quirk, we stipulate using one label per token. For instance, our guidelines would provide the following annotation:

Sally's advice: *get out before it's too late.*

By means of a partial example, the PARC3 guidelines explicitly indicate that in such cases, "advice" should be labeled as both source *and* cue.

5.5. Exclusion of Nested Attributions

Attributions can themselves contain attributions. For example:

She said *he plans to appeal.*
→ **he** plans *to appeal.*

We do not include nested attributions in PolNeAR for two reasons. First, non-nested attributions are those which are reported by the journalist, whereas nested attributions are reported by the people and organizations described by the journalist. As professional and ostensibly unbiased reporters, journalists can be held to a high standard of accuracy, but this does not apply to the agents they report on. Nested attributions are not relevant to questions about journalistic standards and practice. Second, any attempt to model nested attributions directly severely violates univocality, and represents a qualitatively more difficult endeavor. After annotating non-nested attributions, the creators of PARC3 augmented it with nested attributions, but these have not been used to train or test models of attribution. From both the phenomenological and modelling

[6] This is one of a set of *invariance principles*, illustrated in the templates, and used to drive greater consistency in annotation.

perspectives, we consider nested attributions to be of secondary interest, and elect to focus annotation effort on the high-quality annotation of non-nested attributions.

6. Corpus Validation and Statistics

We now assess the the quality of annotations using multiple agreement-based metrics. The statistics we report in this section are based on a random 32% subset of the corpus[7] and a random selection of 56 PARC3 articles that were re-annotated by PolNeAR annotators.

6.1. Attribution-level agreement: *agr*

The first metric, called *agr*, is one adopted by PARC3, originally sourced from Wiebe et al. (2005). It measures the extent to which annotators agree on the existence of an attribution at a given location in the text, without concern for whether the boundaries of the composite spans are exactly aligned. Under this metric, two annotators agree on an attribution if the respective source, cue, and content from their annotations each have some overlap. Formally, *agr* averages the fraction of attributions by one annotator recalled by the other, and vice versa, for all annotator pairs.

This metric is relevant to questions about the distribution of attributions and journalistic practice, where the precise boundaries of spans are not important. Though annotators may disagree on the boundaries of a source, the spans will generally concur on the grammatical head of the source. For example, one annotator might label "Donald Trump" while another labels "Donald Trump, Republican nominee". For questions pertaining to the distribution of attributions throughout articles, and in terms of their focus and their sourcing, small differences in span boundaries are unlikely to systematically bias statistics.

PolNeAR achieves a high *agr* of 92.3% (see **Table 4**). This is an improvement over PARC3, whose *agr* was 83%[8] (although 87% *agr* was achieved in a pilot study by the PARC3 creators).

The form of *agr* is equivalent to the expected arithmetic mean of the precision and recall when one annotator is randomly selected as the ground truth, and the others are compared to it. From a modelling perspective, we could expect this to be a rough upper bound for the level of precision and recall of a model that learns to annotate like human annotators.

6.2. Attribution Pseudo-Recall

Given our concern over false negatives, we consider a metric that proxies for recall during annotation. We cannot, in principle, measure recall exactly, since we lack ground truth on what is and is not an attribution. Nevertheless, we can calculate a kind of pseudo-recall by comparing the performance of PolNeAR and PARC3 annotators on the articles from PARC3 that were annotated by both groups. In other words, we ask: how many of the PARC3 annotations are recalled by the PolNeAR annotators, and vice versa? We use the same notion of what counts as a matching attribution as used in *agr*.

Applying this to the 56 randomly selected PARC3 articles that were re-annotated by PolNeAR annotators, we find that the PARC3's pseudo-recall is only 31.3%, whereas PolNeAR's is 94.2%. In fairness, there are three factors that may explain the difference in pseudo-recall:

1. PARC3 annotators erroneously under-annotated,
2. PolNeAR annotators erroneously over-annotated, or
3. PolNeAR's annotation concept represents an expansion of PARC3's.

To tease apart these factors, it is necessary to inspect the specific attributions annotated by PolNeAR but missed by PARC3 annotators. We have randomly selected 6 such cases for display in **Table 2**.

We obtained these by first randomly choosing 4 of the 56 articles, and collecting all PolNeAR-annotated attributions not recalled by PARC3 annotators, of which there were 51. Inspecting these attributions manually, at least three represent over-zealous annotation on by PolNeAR annotators—we isolate these for display in **Table 3**. In another 5 cases, a matching PARC3 annotation *did* exist, but technically failed to match due to sufficient disagreement on one of the spans. Forgiving these 8 cases, we adjust the pseudo-recall accordingly, arriving at 41.8% quoted in **Table 4**. (We have not made such adjustments for PolNeAR's recall.)

It is from the remaining 43 attributions that we randomly sampled 6 shown in **Table 2**. To show that these are not merely over-zealous annotations, nor reflect an expanded annotation concept, we have collected annotations from elsewhere in PARC3 that bear resemblance to each missed attribution. Given our close adherence to PARC3's guidelines (with alterations only for resolving inconsistency, ambiguity, and multivocality), and given that similar attributions to the ones missed appear elsewhere in PARC3, we submit that the difference in pseudo-recall reflects true low recall in PARC3, and substantially improved recall in PolNeAR.

6.3. Token-level agreement: Krippendorff's α

Finally, we include an agreement metric that focuses on the extent to which the annotators agree on a detailed token-by-token basis, taking into account both discrepancies in whether an attribution exists, and where exactly the boundaries for the component spans lie. Treating each token as an independent labelling decision, we obtain a Krippendorff's α of 75.4%.

By most standards, this is an acceptable (but not high) level of agreement. To understand the sources of disagreement, we selected attributions where PolNeAR annotators had agreed on the existence of attributions, but which had poor overlap in the component spans.

In the majority of cases, the source span was to blame. The following example is characteristic:

1. *"How are you today?"* **Miller**, a retired worker from a nuclear plant, said pleasantly.
2. *"How are you today?"* **Miller, a retired worker from a nuclear plant,** said pleasantly.

[7] The portion of the corpus annotated at the time of writing. This reflects a random sample balanced across strata.

[8] This value reflects agreement on attributions not already present in the PDTB2 annotations, since PDTB2 annotated some attribution phenomena as discourse structure.

Missed by PARC3	Similar attribution annotated by PARC3 elsewhere
That may have pleased **the secretary**, but. . . .	**Maidenform Inc.** loves *to be intimate with its customers, but not with the rest of the public*.
. . . . but **he** has left no doubt *that he still likes the ideas the commission advanced nearly two years ago*.	There is doubt *that the change would accomplish much* but at least Congress, as in 1935, would . . . .
By January it should be fairly clear *what's hot—and what's not*.	It's understood *that MGM/UA recently contacted Rupert Murdoch's News Corp.*, which made two failed bids for the movie studio, *to see if the company was still interested*.
When traders see the Fed is in the exchange market it may make them tread a little carefully, for fear of *what the central bank may do*.	*The stock market's precipitous drop* frightened **foreign investors**, who quickly bid the dollar lower.
Searle, a unit of Monsanto Co., said *the "beta-blocker" high-blood-pressure drug Kerlone is the first product to reach the market through Lorex Pharmaceuticals*. . . .	*The Maidenform name "is part of American pop culture,"* says **Joan Sinopoli**. . . .
Owner Al Brownstein originally planned *to sell it for $60 a bottle*, but	**PaineWebber** considered *an even harder sell, recommending specific stocks*.

Table 2: Random sample of the 45 PolNeAR-annotated attributions not annotated in PARC3. Each such attribution is paired with an attribution that was annotated in PARC3 elsewhere, by way of showing that these attributions do fall under PARC3's annotation concept.

Take *Lake Vineyard Cabernet from Diamond Creek.*
This recommendation might have encouraged **a turf hungry bureaucrat** to try to expand his power ...
Earlier this month the **St. Louis Fed** held a conference to assess *the system's first 75 years*.

Table 3: The 3 PolNeAR annotations in four randomly-selected PARC3 articles, not annotated in PARC3, which seem to violate the PARC3 and PolNeAR annotation concepts.

	PARC3	PolNeAR
Articles	2294	1028
Publishers	1	7
Words (millions)	1.11	0.76
Attributions (thousands)	16.5	23.9
Words-to-attributions ratio	69	32
Token-wise Krippendorff's α (%)	—	75.4
Attibution-wise agreement (%)	83.0 - 87.0	92.3
Pseudo-recall (%)	41.8	94.2
Est. false negatives (thousands)†	22.9	1.06

Table 4: Overall statistics for PARC3 and PolNeAR, based on the 32% currently-annotated fraction of PolNeAR.
† Calculated from pseudo-recall.

As seen in this example, journalists often include background information on their sources. To varying degrees, this information may be needed to identify the source, or may be added to make a more interesting or compelling narrative. According to both PARC3's and PolNeAR's guidelines, such text should be included in the source span when it is needed to identify the source. But in pilot studies, as well as throughout the annotation to date, this notion has proven fraught and attempts to clarify it have failed so far.
Another important source of disagreement comes from ambiguity in the content span. The following example illustrates a common problem:

1. **Friends of Mr. Clinton.'s. . . .** say *that the ebullient energy he is known for*—whether addressing a crowd or spending an hour on a rope line with voters—*has matured into an elder statesman 's self-assurance*.
2. **Friends of Mr. Clinton.'s. . . .** say *that the ebullient energy he is known for—whether addressing a crowd or spending an hour on a rope line with voters—has matured into an elder statesman 's self-assurance*.

In the first annotation, the portion of text between dashes is attributed to "Friends of Mr. Clinton's", while in the second, it is considered an insertion by the author. In these cases, it is the author's intended reading that is unclear, as opposed to the annotation concept, so such cases seem insoluble without a proliferation of arbitrary rules.
For the most part, such ambiguities in source and content span boundaries will probably not have an impact on

substantive sociological, journalistic, or political scientifc questions. But, from the standpoint of model building, this represents noise in the target label sequence, which should be kept in mind in assessing model performance.

7. Conclusion

PolNeAR[9] is, to date, the largest corpus of attribution relations, in terms of number of attributions, and the most complete, in terms of annotator recall. These features address key limitations in existing resources needed to advance more sophisticated models of attribution such as those based on recurrent neural architectures.

PolNeAR is built on a corpus sampled from the coverage of an event of great sociological, political, and journalistic import—campaign coverage during the year leading up to the 2016 US Presidential Election. This period in recent history is of particular significance to attributive phenomena, due to widely held suspicions of bias and improper sourcing practices. Care has been taken to provide equal representation of candidates, publishers, and time periods in the corpus, to maximize statistical power and interpretability in substantive investigations. As automatic attribution extraction and analysis techniques continue to mature, we hope this will spur interdisciplinary work in which automated tools can be used in service of questions about mass media, communications, and political campaign coverage.

8. Appendix—Deviations from PARC3's Annotation Guidelines

Punctuation as cue. PARC3's guidelines instruct annotators to consider punctuation used to introduce attributions as part of the cue, but only when no other cue words are present. We instead consider such punctuation to *always* be part of the cue to provide greater consistency.

Univocality. Near the boundaries of the spans, there is sometimes ambiguity as to which of two spans a token belongs. We address gaps in the guildelines that permit the ambiguity, and then require that tokens can have at most one label. This means sequence-to-sequence models and transition-based parsers are straightforward to apply.

Crediting of work. We specifically address the issue of attributing works, such as plays, books, paintings, etc. The PARC3 guidelines are ambiguous on this issue. We do not consider the attribution of a work to its creator to be sufficient for attribution. But, the attribution of some *content*, to either the work or its author *is* annotated. This distinction is illustrated by examples in the templates (see §13).

Possessive edge clitic ('s). In attributions where the source has an edge clitic (*'s*), as in "**the author's** view is *that...*", we consider the *'s* to be part of the source, not the cue. This is for consistency with the treatment of attributions in which a possessive pronoun is the source (e.g. "**her** view is *that...*"), wherein the possessive pronoun is labelled as the source, and the possessed noun is the cue.

[9]Obtain the dataset: https://github.com/networkdynamics/PolNeAR

Empty content. PARC3 and PolNeAR both exclude "empty content", such that this would not be a valid attribution: "**John** said *these three words*". This seems reasonable because "these three words" refers not to the content, but to the medium of expression. However, PARC3's guidelines make an exception just in case it is anaphoric to content elsewhere in the document, as in "*"I am sorry."* **John** said *these three words*". For greater consistency, we never annotate tokens referent to the medium of communication as content. (This should not be confused with the annotation of non-personal pronouns which *are* coreferent with content, such as "**she** denies *it*". See the templates, §13, for examples of this distinction.)

9. Bibliographical References

Carlson, L., Okurowski, M. E., and Marcu, D. (2002). *RST discourse treebank*. Linguistic Data Consortium, University of Pennsylvania.

Elson, D. K. and McKeown, K. (2010). Automatic attribution of quoted speech in literary narrative. In *AAAI*.

Esser, F. and Umbricht, A. (2014). The evolution of objective and interpretative journalism in the Western press: Comparing six news systems since the 1960s. *Journalism & Mass Communication Quarterly*, 91(2):229–249.

Evans, D., Ku, L.-W., Seki, Y., Chen, H.-H., and Kando, N. (2007). Opinion analysis across languages: An overview of and observations from the ntcir6 opinion analysis pilot task. *Applications of Fuzzy Sets Theory*, pages 456–463.

Li, H., Cheng, X., Adson, K., Kirshboim, T., and Xu, F. (2012). Annotating opinions in german political news. In *LREC*, pages 1183–1188.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

O'Keefe, T., Pareti, S., Curran, J. R., Koprinska, I., and Honnibal, M. (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*, pages 790–799. Association for Computational Linguistics.

Pareti, S., O'Keefe, T., Konstas, I., Curran, J. R., and Koprinska, I. (2013). Automatically detecting and attributing indirect quotations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 989–999.

Pareti, S. (2012). A database of attribution relations. In *LREC*, pages 3213–3217.

Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A., Robaldo, L., and Webber, B. L. (2007). The penn discourse treebank 2.0 annotation manual.

Pustejovsky, J., Hanks, P., Sauri, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L., et al. (2003). The timebank corpus. In *Corpus linguistics*, volume 2003, page 40. Lancaster, UK.

Russel, B. (1940). *An Inquiry into Meaning and Truth*. W. W. Norton Company.

Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2):165–210.

Wolf, F. and Gibson, E. (2005). Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31(2):249–287.