Assessing the Verifiability of Attributions in News Text

Edward Newell
McGill University
Computer Science

Ariane Schang
McGill University
Computer Science

Drew Margolin
Cornell University
Dept. of Communication

Derek Ruths
McGill University
Computer Science

{edward.newell, ariane.schang}@mail.mcgill.ca,
dm658@cornell.edu, derek.ruths@mcgill.ca

Abstract

When reporting the news, journalists rely on the statements of stakeholders, experts, and officials. The attribution of such a statement is *verifiable* if its fidelity to the source can be confirmed or denied. In this paper, we develop a new NLP task: determining the verifiability of an attribution based on linguistic cues. We operationalize the notion of verifiability as a score between 0 and 1 using human judgments in a comparison-based approach. Using crowdsourcing, we create a dataset of verifiability-scored attributions, and demonstrate a model that achieves an RMSE[1] of 0.057 and Spearman's rank correlation of 0.95 to human-generated scores. We discuss the application of this technique to the analysis of mass media.

1 Introduction

An attribution occurs when an author or speaker represents the discourse, attitude, or inner state of an external source (Piazza, 2009). Attributions are found in virtually every genre of discourse (Fairclough, 1995), but are fundamental to news reporting, where attribution to credible sources is a basic feature of objective, unbiased "hard news" (Esser and Umbricht, 2014). Recently news media have come under increasing scrutiny for spreading biased and even fabricated information[2]. This trend suggests the need for scalable, computational approaches to understanding attribution.

From a natural language processing perspective, attribution is a fundamental phenomenon that touches a broad set of applications, including summarization, question answering, information extraction, and discourse analysis. Once content is scoped under attribution, its contribution to the discourse can change substantially depending on the source of the attribution and their relationship to the statement. For instance, in 2001 U.S. President George W. Bush famously warned that in fighting terrorists, nations were either "with us or against us"[3]. This statement was threatening not only because it was made by the president, or because it was blunt, but because such blunt statements are not normally made by national leaders. If they are to reach human-level performance, systems for automated text understanding must not only accurately segment attribution in the flow of text, but also represent the many ways that attributions can differ rhetorically.

One important way in which attributions can differ, particularly with respect to news reporting, is in their *verifiability*, the ease with which an attribution's fidelity to the source can be checked. Consider the following hypothetical attributions:

> **Lindsay Walls**, CEO of Inovatron, said in a press release yesterday *that the "allegations of intentionally selling sub-par product are completely unfounded."*

> A **source** close to the issue hinted *that quality control standards had been on the decline.*

The former attribution is more verifiable insofar as it attributes a specific statement to a specific person who, in theory, could be asked to corroborate it. The latter is harder to verify. The source is not named, and even if it were known, it is not clear how it could be confirmed that such "hints"

[1] Root mean squared error

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

[3] edition.cnn.com/2001/US/11/06/gen.attack.on.terror

Proceedings of the The 8th International Joint Conference on Natural Language Processing, pages 754–763,
Taipei, Taiwan, November 27 – December 1, 2017

were given. Note that verifiability does not depend on the truth of a statement, nor its fidelity to the source, but rather on the *ability to confirm or deny* its fidelity.

The ability to confirm or deny an attribution's fidelity is not binary, but rather occupies a continuum of difficulty, which can depend on whether the source was precisely identified, and how definite the reported statement or content was. Therefore, we operationalize verifiability as a continuous variable between 0 and 1. Attributions are scored by having humans compare attributions and judge which are more verifiable.

We build a dataset[4] of verifiability-scored attributions on top of the Penn Attribution Relations Corpus, version 3 (PARC3) (Pareti, 2012, 2015). PARC3, derived from the Penn Discourse TreeBank, consists of Wall Street Journal news articles in which attributions have been manually annotated.

Prior work investigating sourcing typically relies on a binary concept of named versus anonymous sources (Wulfemeyer and McFadden, 1986). It is also common to distinguish between verbatim quotes, mixed quotes, and paraphrases (also known as reported speech) (Sundar, 1998). Therefore, we create a baseline model of verifiabilty based on these features. In reality, the sources of attributions span the range from completely obscured to named with credentials and affiliations, along with intermediate examples such as "Whitehouse official" or "company spokesperson". We compare the baseline model to a more sophisticated model that considers a large number of syntactic and semantic cues based on the source and other parts of the attribution[5]. Using ablation testing, we assess the contribution that these various features make to the regression of verifiability.

Before concluding the paper, we discuss the applications of automated verifiability scoring to the study of attribution in mass media. We explore the feasibility of an end-to-end system that extracts attributions from raw text and then scores the attribution's verifiability, by implementing an existing attribution extraction pipeline from prior work (Pareti et al., 2013). We analyze how errors cascade through the coupled extraction-regression pipeline, which illuminates the challenges to the end-to-end version of the task for future work.

[4] cs.mcgill.ca/~enewel3/publications/verifiability-IJCNLP-2017-09

[5] github.com/networkdynamics/Verifiability-IJCNLP-2017

2 Related work

2.1 Computational approaches to attribution

The extraction of attributions from text requires (1) the detection of attributed content (e.g. a quotation), and (2) linking of that content to a source entity. Although detection of enquoted text is trivial, a great deal of attributed content is not found within quotes.

The earliest systems attempt to attribute quotations in children's stories to the correct speaker (Zhang et al., 2003; Mamede and Chaleira, 2004). These systems used rule-based approaches, and although they achieved high accuracy on extracting quotations, their accuracy in attributing them to the correct speaker was quite low. The performance of these systems was also highly dependant on the genre of material.

In news text, a substantial fraction of attributions are much harder to extract, being signaled by the discursive structure of the text instead of by explicit quotation marks (O'Keefe et al., 2012). Early systems for extracting and linking attributions in this domain assumed low recall to achieve higher precision (Pouliquen et al., 2007; de Morais et al., 2009). To perform well in such domains, a machine learning approach was needed. Elson and McKeown provided the first contribution in this direction (2010). However, their approach relied on gold-standard labels for attributions occurring earlier in the text as features to extract later attributions. In 2012, the first practical machine learning-based approach capable of extracting attributions from non-annotated news text was developed using a sequence labelling approach (O'Keefe et al., 2012).

Further efforts were spurred by the development of corpora with annotated attributions, including PARC3 (Pareti, 2012, 2015). In PARC3 an attribution consists of: (1) a source to whom content is being attributed, (2) the content being attributed, and (3) the cue phrase referring to the act of attribution (e.g. "said", "according to"). PARC3 enabled the development of an attribution extraction system that we replicate to investigate end-to-end attribution extraction and verifiability regression (Pareti et al., 2013). This multi-step extraction pipeline first identifies candidate reporting words (e.g. *said*, *lamented*), and uses these as features to extract the attribution content. Using the ex-

tracted content, candidate source entities are identified and correct links are found using a classifier. The source entity and cue word are then expanded deterministically into a source *span* and cue *span*, which collects informative modifiers such as the source's affiliation.

2.2 Attribution and verifiability in news

Attribution plays a fundamental yet complex role in news reporting. It is the "bread and butter" of hard news journalism (Sundar, 1998). However, attribution affords the author the opportunity to frame or interpret information by proxy. Attributions tend to have evaluative content, suggesting that "external voices are allowed to speak their minds much more loudly than journalists" (Jullian, 2011), yet the rhetorical use of attributions is often subtle (Fairclough, 1995).

The credibility of attributions is fundamental to trust in the media. As Burriss (1988) states, "one of the basic tenets of journalism is that news reports are supposed to deal with verifiable facts... Unfortunately the public who receive the news generally has no way to independently verify the accuracy of a news story and must thus depend upon (1) the reputation of the news organization, (2) the reputation of the reporter, or (3) information within the story itself, in order to determine the accuracy of a news report." Journalists do not always provide the information necessary to verify an attribution (Adams, 1962; Wulfemeyer, 1985; Wulfemeyer and McFadden, 1986). Anonymous sourcing has recently been criticized for distorting coverage of the 2016 presidential campaign (Silver, 2017), but the practice has been recognized and cautiously accepted by media researchers and practitioners for decades (Wulfemeyer, 1985; Boeyink, 1990; Duffy and Freeman, 2011). Anonymizing sources does tend to undermine the credibility of a story (Sternadori and Thorson, 2009; Pjesivac and Rui, 2014; Mackay and Bailey, 2012), though not in all cases (Sundar, 1998), and there is variation in the kinds of unnamed sources who are found credible (Adams, 1962; Riffe, 1980). A typical set of "code-words" are often applied to veil source identity (e.g. "official" or "spokesman") (Burriss, 1988), with the choice of terms having significant impact on the credibility of the report (Adams, 1962). Direct quotes also appear to enhance credibility compared with paraphrases (Sundar, 1998). Thus, researchers have recently tracked the use of anonymous sources over time and across cultures (Esser and Umbricht, 2014; Lee and Wang, 2016).

2.3 Conceptualizing verifiability

Source anonymity is just one aspect of the problem of how attributions might be used to influence reader interpretations. Popper 2003 argued that any knowledge claim possesses a *degree of verifiability* [6]: the extent to which it is possible to test for evidence that could corroborate or contradict it. Claims with higher verifiability are more credible even prior to testing because authors have less incentive to be accurate when making unverifiable claims (Margolin and Monge, 2013). For example, since no one can know whether an anonymous source really made the statement attributed to them in a news article, the reporter could distort or even fabricate the attributed statement.

Source identification aside, we note that Popper emphasized the form of the proposition, for example, claims made with qualifiers or weak quantifiers. Additionally, physical and technical barriers also apply (Deutsch, 1997). Sources who are difficult to access or that lack a platform from which to correct mis-attribution are less verifiable. The language of attributed content may also matter: verbal categories can be vague or sharp (Hampton, 2007), modifying the extent to which claims made with them are verifiable.

3 Operationalizing verifiability

3.1 Task definition

As mentioned, the ability to confirm or deny the fidelity of an attribution occupies a sliding scale of difficulty, so we operationalize verifiability as a quantity between 0 and 1. Similar to credibility or relevancy, verifiability fundamentally reflects the perceptions on the part of readers. Although it is possible, at least in principle, to directly test the difficulty of verifying a given attribution, the psycholinguistic notion of verifiability is more relevant to characterizing mass media production and consumption.

To approximate the perceptions of the general public, we use crowdsourced human judgments in creating the ground truth verifiability-scored dataset. Crowdworkers were shown pairs of attributions, and asked to decide which is more ver-

[6] Popper technically refers to "falsifiability" which is the inverse of verifiability.

ifiable (we describe the details of the annotation setup below). Various methods exist to convert pairwise comparisons into a set of scores (e.g. (Kiritchenko and Mohammad, 2016)). We use the Bradley Terry model (Hunter, 2004) to assign verifiability scores, and then shift and scale the scores to fall into the $[0, 1]$ interval.

As discussed above, methods for extracting attributions from raw text have been developed in prior work. Therefore, this task focuses on the regression of perceived verifiability from text that has already been annotated with attributions in PARC3 annotation style.

3.2 Dataset annotation

Annotation was carried out using the CrowdFlower platform[7]. Crowdworkers were shown pairs of attributions, and asked to consider the effort required to confirm or deny the fidelity of each. They were asked to select the attribution that was easiest to verify from each pair.

In judging verifiability, it is reasonable to expect that the precision with which a source is identified would be the major determinant in most cases. This creates a risk that crowdworkers will begin to rely only on source definiteness, rather than judging attribution verifiability holistically. Thus, we took steps to ensure that crowdworkers were vigilent to verifiability cues of various kinds. As part of general quality control, crowdworkers had to complete 7 out of 10 training / test examples correctly to ensure they understood the task, and then maintain this proportion of correct responses on test examples randomly dispersed throughout the annotation tasks. To address the specific concern that crowdworkers may become overly reliant on source definiteness, we selected training / test examples to which the correct answers depended on a variety of cues. Test examples were collected by performing a pilot round of annotation with 8 expert annotators, and selecting from the high-agreement examples.

Attributions were presented within the full sentence(s) that contained them. Limiting the context to the containing sentence(s) did not appear to interfere with annotation during the pilot round. Nevertheless, we took steps to mitigate effects from the loss of context. In the majority of cases where the definiteness of the source plays an important role in determining verifiability, the most useful context is likely to be how the source was first introduced in the article, e.g. whether the source's name and affiliation were given. To bring that context into the attribution, we used the CoreNLP coreference resolution software (Manning et al., 2014) to augment the source. Whenever a source was mentioned using a personal pronoun, we interpolated the pronoun using the representative coreferent mention, except where that mention already occured in the sentence. Thus, for example,

> "I don't know," she said,

might become

> "I don't know," Lindsay Walls, CEO of Inovatron said.

Spot checks of 100 pronoun-containing attributions in PARC3 showed that this produced reliable, grammatical interpolations. However, similarly interpolating non-personal pronouns and references such as "the company" was not reliable. We instructed workers to consider, when faced with such references, whether it appeared that the reference was to a specific named individual / entity. Thus, the worker should treat "a company" differently from "the company". We included many test examples in which workers had to act on that instruction to get the correct answer.

Attributions were presented with the source, cue, and content highlighted, to ensure that workers knew what specific attribution they they were annotating.

We solicited comparisons involving 2100 attributions, presented as 39930 unique pairs, with each pairing annotated by at least 3 workers (increased to 5 when the first judgments were non-unanimous), resulting in 140277 total pairwise comparisons. The data comprise annotations from 337 workers. These figures exclude the discarded data from 70 crowdworkers that had poor performance on training / test examples. Every attribution in the dataset was compared to at least 20 other attributions. After 20 pairings, we found that the verifiability scores and the model regression error (to be discussed below), had become relatively stable, as can be seen in **Fig. 1**.

Annotation proceeded in two phases (not including the pilot round). In the first phase, we randomly sampled 100 attributions from PARC3 and solicited an exhaustive set of annotations on all 4950 pairs. This gave highly accurate scores for a small subset of annotations. In the second

[7] crowdflower.com

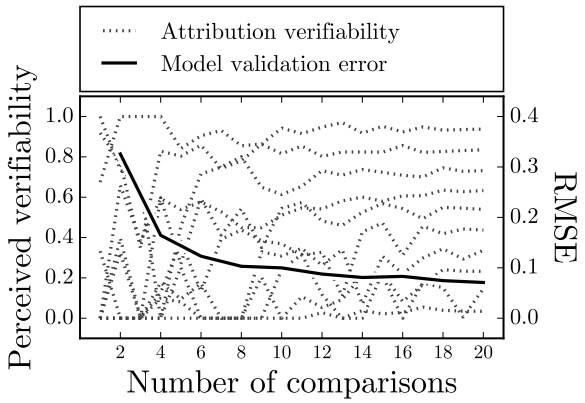

Figure 1: Estimated verifiability scores converged as a function of the number of comparisons per attribution increased, reducing error in the model.

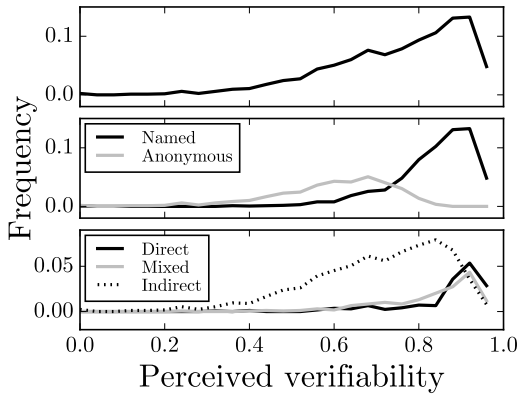

Figure 2: Distribution of crowdsourced verifiability scores for 2100 attributions (top), conditioned on the presence of ORG and PERS named entities (middle) and attribution directness (bottom).

phase, 2000 randomly sampled attributions were systematically compared to the attributions scored in phase 1. Each phase-2 attribution was paired with two phase-1 attributions randomly drawn from each decile of verifiability, giving twenty pairings each.

After the pairwise comparison data were collected we estimated the maximum-likelihood latent verifiability scores in a Bradley-Terry model using a majorization-minorization algorithm (Hunter, 2004), then linearly transformed the verifiability scores to fall in the range $[0, 1]$.

3.3 Annotation results

The resulting scores for the 2100 attributions produced a smooth unimodal distribution across a range of perceived verifiability (**Fig. 2**). In a plot of the latent scores inferred for a Bradley-Terry model, the distance between two scores is a direct representation of the probability that the higher-scoring item would "win" in a comparison with the other[8]. The distribution has most of its mass near verifiability score 0.9, showing that most attributions adhere to a high standard of verifiability. In addition, we see heavy tail of low perceived verifiability. This distribution is, prima facie, consistent with the competing journalistic norms of transparent attribution and source protection.

It would be reasonable to have expected the distribution to be somewhat more clustered, with peaks in the data corresponding to discrete features such as named vs. anonymous sources (Wulfemeyer and McFadden, 1986) or direct vs. indirect quotes (Sundar, 1998). These features would have clear bearing on efforts to verify an attribution and should be fairly universally recognizable. The distribution instead suggests there is a diversity of factors that contribute to perceived verifiability and that focusing solely on one or two discrete, a priori obvious indicators to analyze attribution behavior conceals a great deal of variation. As shown in **Fig. 2** source anonymity and quote directness do have explanatory power, however they appear unable to explain the continuum of perceived verifiabilities.

3.4 Inter-annotator agreement

Inter-annotator agreement serves as a standard check to ensure that workers understood and reliably performed the task. In this case some care is needed in interpreting the inter-annotator agreement, however: we expect a certain amount of inbuilt disagreement due to comparisons made between attributions having very close verifiability scores. In fact, the Bradley-Terry model (and other methods for deriving scores from pairwise comparisons) is predicated on the notion that two items have similar latent score precisely when workers are unable to reliably decide which is to be preferred (i.e. which has higher verifiability).

For this reason, the agreement level for attributions having sufficiently similar scores will necessarily be low. To properly assess reliability for this kind of annotation, it is necessary to disaggregate the comparisons based on how far apart the compared attributions are in verifiability score.

To ensure that disaggregated agreement scores

[8]The probability of a worker judging a_1 to be more verifiable than a_2, given they have verifiability scores v_1 and v_2, is modeled as $\Pr(a_1 \succ a_2) = (1 + e^{\beta(v_2 - v_1)})^{-1}$, where β is the scale parameter used to transform scores to $[0, 1]$.

Separation	0	1	2	3	4
Agreement	.174	.363	.651	.825	.904

Table 1: Krippendorff's α for comparisons between attributions from quintiles of given separation. A separation of 0 means comparisons between attributions in the same quintile.

are unbiased, we use half of the pairings in the dataset to assign scores, and then assess agreement on the other half. Using the scores derived from half the data, we divide the attributions into quintiles. We then assess agreement on comparisons that are made between attributions from quintiles having a given separation.

For example, all comparisons made between attributions from the same quintile have separation 0, while comparisons made between attributions from adjacent quintiles have separation 1. The agreements associated to each level of separation are shown in **Table 1**. As the separation between attributions is increased, the level of agreement monotonically increases, becoming very high (Krippendorff's α 0.904) for comparisons between attributions from the highest and lowest quintiles. This indicates that once we factor out disagreement due to perceived similarity of attributions, workers were able to understand and perform the task with high reliability.

4 Modeling verifiability

In the PARC3 annotation style, an attribution consists of a *source*, the attributed *content*, and a *cue*, such as "said" or "according to", signalling the existence of an attribution. A priori, any of the three parts of an attribution (source, cue, content) could contribute to the perceived verifiability of an attribution. In addition to the baseline model based on source anonymity and whether the attribution is a direct, indirect, or mixed quote, we also test a "feature-rich" model based on a large number of features extracted from attributions' source, cue, and content, listed in **Table 2** (the baseline model is based on features S2 and C4). We test multiple regression algorithms for both the baseline and rich feature set, and we do feature ablation to optimize the feature-rich model. In the next subsection, we discuss the selection of features and ablation results, then in the following section, we describe the learning algorithms and the best results achieved for the baseline and feature-rich models.

		Feature set	RMSE_i	$\Delta_{-i}\text{RMSE} \times 10^3$
source	S	All source features	.093	14.09
	S1	Length	.124	1.70
	S2	Anonymity	.110	1.34
	S3	Each of 7 NE types	.102	.79
	S4	Head's determiner	.164	.52
	S5	Head lemma	.115	.46
	S6	Head plural	.152	.15
	S7	Fuzzy quantifiers	.159	.04
	S8	Pronoun[9]	.163	.03
	S9	Date or numeric NEs	.164	-.04
	S10	Head's `amod`	.164	-.09
cue	Q	All cue features	.134	8.11
	Q1	LIWC dictionary counts	.137	.88
	Q2	Lemmatized BOW	.136	.39
	Q3	Length	.161	-.02
	Q4	Cue class	.142	-.33
content	C	All content features	.139	1.90
	C1	Date or numeric NEs	.163	.13
	C2	Fract. enquoted tokens	.149	.01
	C3	PERS or ORG NEs	.164	-.05
	C4	Direct, indirect, mixed	.149	-.11
	C5	Each of 7 NE types	.161	-.22
	C6	Length	.153	-.59

Table 2: Features for verifiability regression. RSME when using the feature on its own (RMSE_i), and drop in RMSE occurring when the feature is removed from a model built from all features ($\Delta_{-i}\text{RMSE}$). Entries sorted in descending order of $\Delta_{-i}\text{RMSE}$.

4.1 Feature design and selection

For the sake of continuity, as we describe features, we will also discuss the results of ablating them. Ablation results are based on the training-set performance of a Support Vector Regressor (SVR), optimizing for minimum root mean squared error (RMSE) between predicted verifiability scores and those derived from human annotations. The full set of features is listed in **Table 2**. For ablation testing, we assessed each feature in two contexts: (1) as the only feature used, and (2) as the only feature left out. While the first measures the straightforward predictiveness of the feature, the second measures the marginal improvement in the context of other features and is used for final feature selection in the feature-rich model.

Source features. Based on the CoreNLP named entity recognision (NER) software (Manning et al., 2014), we created features indicating whether any of the 7 types[10] of named entities (NEs) were present in the source (feature S3, **Ta-**

[9] Aside from "he", "she", and "they", which are interpolated.

[10] CoreNLP recognizes seven types of named entities: PERS, ORG, DATE, MONEY, DURATION, PERCENT, NUMBER.

ble 2). This feature was beneficial both alone and in the context of other features. We also included a feature encoding the anonymity of the source, based on whether either a PERS or ORG NE was present (this was one of the baseline features, S2). Although this may seem redundant with S3, encoding the source anonymity in this way boosted performance even in the context of S3. Several of the other NEs are number-like: we also created a feature indicating whether any number-like NE was present (S9), but it hindered performance in the context of other features.

When not a NE, the head of the source span is often a title ("director"), occupation ("lawyer"), or collective designation ("homeowners"). Intuitively, pluralized designations seem more nebulous, so we introduced a feature indicating whether the head of the source span is pluralized (S6). Although less predictive on its own, this feature did provide a benefit to the model in the context of other features.

Similarly, the determiner of the head of the source can influence definiteness: consider "a lawyer" versus "the lawyer". We added feature S4 indicating the kind of determiner used, if any, which also made an important contribution to the overall model.

Quantifiers such as "most", "some", "many", and "several", can also render a source imprecise, so we included a feature indicating the presence of such quantifiers or the word "source(s)" (S7). In ablation testing this feature marginally improved the model's accuracy in the context of other features.

We included the lemma of the head (S5), which made a notable contribution. This feature subset likely suffered from sparsity, so its contribution might be more important given a larger training set.

Modifiers to the source could also influence verifiability, so we included features for the lemma of tokens under the `amod` dependency tree relation to the head (S10)[11]. This feature did not benefit the model, but again may perform better in larger datasets.

The source feature providing the greatest contribution in the context of other features was the length of the source (S1) (although on its own it is less predictive than source anonymity, S2). Intuitively, the longer the specification of the source, the more definite it is, and the more verifiable. The strong performance of this feature suggests additional features might account more specifically for language not accounted for by other features which contribute to verifiability.

Cue features. We derived four features from the cue, which were lemmatized bag-of-words (Q2), length (Q3), counts of words belonging to each of the LIWC dictionaries (Q1) (Tausczik and Pennebaker, 2010), and the presence of specific sets of reporting verbs (Q4). This last (Q4) was based on our observation that reporting verbs either indicate the statement neutrally ('said', 'reported'), qualify the statement as true ('confirmed', 'showed'), indicate an intention ('will', 'plans'), or call to question whether the statement is true ('believes', 'claimed'). The liwc dictionary counts (Q1) and lemmatized bag-of-words (Q2) both made substantial contributions to model accuracy.

Content features. A verbatim attribution seems inherently easier to verify than a paraphrase, so attribution directness was included (C4). Ablation testing showed that attribution directness was actually detrimental to the regression overall and was a poor predictor alone. However, we created a more nuanced representation of quote directness based on the fraction of enquoted words (C2), which was marginally beneficial.

Given that verifiability depends not on the truth of a statement, but on the ability to check the fidelity of the attribution to the source, one might expect that (aside from attribution directness) the content would have little effect on verifiability. However, it is inherently harder to verify the attribution of a vague paraphrase, which could be consistent with a wider range of original statements. Conversely, numerical quantities and the naming of people and organizations should increase the verifiability. As we did for the source, we included features representing the 7 types of named entities (C5), a feature indicating either PERS or ORG (C3), and a feature indicating numerical entities (C1). The numeric entities feature did indeed improved model accuracy although the other features derived from NEs in the content did not.

Finally, we included the length of the content (C6), but this feature was detrimental to the model.

Ablating feature blocks. Considering the role that they play in attribution, one would expect

[11] Based on the CoreNLP dependency parse (Manning et al., 2014)

Model	RMSE	ρ
baseline	0.102	0.833
feature-rich	0.057	0.951

Table 3: Test-set RMSE and Spearman's rank correlation (ρ) for each model.

that, overall, the source would be most informative due to its importance in being able to trace the attribution to a specific person, group, or artifact, followed by the cue, due to the fact that the cue describes the act of attribution and can indicate the certainty or degree of interpretive licence exercised by the author (e.g. if the cue is "hinted").

To test the importance of the source, cue, and content, we ablate each set of features as a whole, the results of which are indicated in **Table 2** by the rows that have feature symbols containing only an 'S', a 'Q', or a 'C' (with no number). These results confirm that the source is most informative, followed by the cue.

4.2 Models training and testing

We randomly separated the dataset of 2100 quotes into a testing and training set of 420 and 1680 attributions each. Using the training set, we used three learners to optimize the performance on models using the baseline set of features, and a rich set of features (those contributing positively to model accuracy in the context of other features, see third column of **Table 2**). The learners included linear regression with lasso regularization, a support vector classifier (SVC) that predicts the quintile from which an attribution was drawn (and returns the median score), and a support vector regressor (SVR). The support-vector-based models used linear, quadratic, and radial basis function as kernels. Using cross-validation on the training set, we optimized the learner selection, kernel selection, and learner hyperparameters, and performed ablation testing for the feature-rich model. Optimization was based on minimizing the RMSE. SVR performed best for both the baseline and ablation-optimized feature sets.

The optimized baseline and feature-rich model were then each run once on the test set, with the results summarized in **Table 3**.

RMSE gives a measure of error between the model's predicted scores, and the true verifiability scores. It's dimensionality and scale are equivalent to those of the variable predicted, so it can be compared to the range of values across which verifiability varies; the feature-rich RMSE was 5.7% of the prediction range. In many applications, the absolute verifiability may be less important than the relative verifiability. Both the baseline and the feature-rich model achieve relatively high Spearman's rank correlations ($p \ll 0.001$). The feature-rich model provides a substantial improvement in performance over the baseline, both in RMSE and rank correlation. This shows that a richer set of features, beyond source anonymity and quote directness, is needed to explain the perception of attribution verifiability.

Verif.: quintile actual predicted	Attribution
1 0.316 0.482	It is rumored *to be bound for a new model in the luxury Acura line in the U.S.*
2 0.698 0.676	**Earlier U.S. trade reports** have complained of *videocassette piracy in Malaysia and disregard for U.S. pharmaceutical patents in Turkey*
3 0.802 0.766	**South Korea** announced *$450 million in loans to the financially strapped Warsaw government.*
4 0.884 0.887	*Mr. Paul has been characterised as "the Great Gatsby or something,"* complains **Karen E. Brinkman, an executive vice president of CenTrust**
5 0.960 0.959	*"It has an archival, almost nostalgic quality to it,"* says **Owen B. Butler, the chairman of the applied photography department at Rochester Institute of Technology.**

Table 4: Selected attributions from each quintile of the verifiability-scored subset of PARC3 along with model predictions.

A selection of attributions from each quintile, along with their human-judged and model-predicted verifiability scores are shown in **Table 4**. These examples demonstrate how the model has learned to consider various features in regressing verifiability. Aside from the first, each of the examples in **Table 4** contains a named entity, however, it would appear that the model has learned to attribute less verifiability to location names than names of individuals. Additionally, in the example from quintile 2, we can see that the head of the source is the word "reports", which is likely what has led to its appropriately lower predicted score: it would be quite difficult, though possible, to comb through a sufficient number of U.S. trade

reports to reach a verdict about the fidelity of this attribution.

5 Application to the analysis of mass media

Journalists are frequently forced to decide whether given sources are sufficiently credible and relevant to cite, while balancing transparent attribution against the source's potential interest in remaining anonymous. It is reasonable to wonder what influences and biases exert themselves on such decisions.

If there are systematic influences at play, it should be possible to find evidence in the distribution of verifiability, and its correlation with publishers, topics, positions on given issues, and political alignments. To look for such patterns at scale, it will be necessary to create an end-to-end system for attribution extraction and verifiability regression.

Although prior work demonstrates good performance on attribution extraction, and we demonstrate accurate verifiability regression here, our initial investigations of an end-to-end extraction and regression system show that errors during extraction lead to large negative errors in verifiability (i.e. underestimates) during regression. This is especially true when there are errors in extracting the source span. Investigating verifiability at scale will require some combination of: (a) further improvements to extraction accuracy, (b) discarding poorly extracted attributions (with loss of recall), and (c) adjustment of the extraction / regression models to reduce error cascading, which we hope to investigate in future work.

6 Conclusion

Attribution is a critical feature of journalism, and a fundamental, challenging natural language phenomenon. We have introduced a new NLP task consisting of the prediction of attributions' perceived verifiability according to human judgments. We provide a dataset of verifiability-scored attributions based on a subset of PARC3.

Our models show that source anonymity and quote directness alone are insufficient to explain the continuum of perceived verifiability, but a richer set of linguistic features enables accurate verifiability regression. The source appears to be the dominant factor determining an attribution's verifiability, with an important contribution also coming from the cue, and a slight contribution from the content.

This new task, along with existing work in attribution extraction, creates a new opportunity to study attribution practices in mass media, at scale, and shed light on the shifting landscape of journalistic norms.

Acknowledgments

We thank Silvia Pareti for helpful discussions about PARC3 and attribution extraction.

References

John B Adams. 1962. The relative credibility of 20 unnamed news sources. *Journalism Quarterly* 39(1):79–82.

David E Boeyink. 1990. Anonymous sources in news stories: Justifying exceptions and limiting abuses. *Journal of Mass Media Ethics* 5(4):233–246.

Larry L Burriss. 1988. Attribution in Network Radio News A Cross-Network Analysis. *Journalism and Mass Communication Quarterly* 65(3):690.

Luís António Diniz Fernandes de Morais, Sérgio Sobral Nunes, et al. 2009. Automatic extraction of quotes and topics from news feeds. In *DSIE09-4th Doctoral Symposium on Informatics Engineering*.

D. Deutsch. 1997. *The Fabric of Reality: The science of parallel universes and its implications*. Allen Lane, New York.

Matt J. Duffy and Carrie P. Freeman. 2011. Unnamed Sources: A Utilitarian Exploration of their Justification and Guidelines for Limited Use. *Journal of Mass Media Ethics* 26(4):297–315.

David K Elson and Kathleen McKeown. 2010. Automatic attribution of quoted speech in literary narrative. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*.

Frank Esser and Andrea Umbricht. 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.

Norman Fairclough. 1995. *Critical discourse analysis*. Longman, New York.

J.A. Hampton. 2007. Typicality, graded membership, and vagueness. *Cognitive Science* 31:355–383.

David R Hunter. 2004. MM algorithms for generalized bradley-terry models. *Annals of Statistics* pages 384–406.

Paula M. Jullian. 2011. Appraising through someone elses words: The evaluative power of quotations in news reports. *Discourse & Society* 22(6):766–780.

Svetlana Kiritchenko and Saif M. Mohammad. 2016. Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling. In *Proceedings of The 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*. San Diego, California.

Seok Ho Lee and Qian Wang. 2016. A Comparative Investigation Into PressState Relations: Comparing Source Structures in Three News Agencies Coverage of the North Korean Missile Crisis. *International Journal of Communication* 10:22.

Jenn Burleson Mackay and Erica Bailey. 2012. Succulent Sins, Personalized Politics, and Mainstream Medias Tabloidization Temptation. *International Journal of Technoethics* 3(4):41–53.

Nuno Mamede and Pedro Chaleira. 2004. Character identification in children stories. In *Advances in natural language processing*, Springer, pages 82–90.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*. pages 55–60.

Drew B. Margolin and Peter Monge. 2013. Conceptual retention in epistemic communities. In Patricia Moy, editor, *Communication and community*, Hampton Press, New York, pages 1–22.

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*. Association for Computational Linguistics, pages 790–799.

Silvia Pareti. 2012. A database of attribution relations. In *Language Resources and Evaluation Conference*. pages 3213–3217.

Silvia Pareti. 2015. Attribution: a computational approach .

Silvia Pareti, Timothy O'Keefe, Ioannis Konstas, James R Curran, and Irena Koprinska. 2013. Automatically detecting and attributing indirect quotations. In *Empirical Methods in Natural Language Processing*. pages 989–999.

Roberta Piazza. 2009. News is Reporting What was Said. Techniques and Patterns of Attribution. *Evaluation and Stance in War News* pages 170–194.

Ivanka Pjesivac and Rachel Rui. 2014. Anonymous sources hurt credibility of news stories across cultures: A comparative experiment in America and China. *International Communication Gazette* 76(8):641–660.

K.R. Popper. 2003. *The Logic of scientific discovery*. Routledge, New York.

Bruno Pouliquen, Ralf Steinberger, and Clive Best. 2007. Automatic detection of quotations in multilingual news. In *Proceedings of Recent Advances in Natural Language Processing*. pages 487–492.

Daniel Riffe. 1980. Relative credibility revisited: How 18 unnamed sources are rated. *Journalism Quarterly* 57(4):618–623.

Nate Silver. 2017. Why You Shouldnt Always Trust The Inside Scoop.

Miglena Mantcheva Sternadori and Esther Thorson. 2009. Anonymous sources harm credibility of all stories. *Newspaper Research Journal* 30(4):54–66.

S. Shyam Sundar. 1998. Effect of source attribution on perception of online news stories. *Journalism and Mass Communication Quarterly* 75(1):55–68.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29(1):24–54.

K Tim Wulfemeyer. 1985. How and Why Anonymous Attribution Is Used by "Time" and "Newsweek". *Journalism and Mass Communication Quarterly* 62(1):81.

K. Tim Wulfemeyer and Lori L. McFadden. 1986. Anonymous attribution in network news. *Journalism and Mass Communication Quarterly* 63(3):468.

Jason Y Zhang, Alan W Black, and Richard Sproat. 2003. Identifying speakers in children's stories for speech synthesis. In *Interspeech*.