Source Attribution: Recovering the Press Releases Behind Health Science News

Ansel MacLaughlin,[1] John Wihbey,[2] Aleszu Bajak,[2] David A. Smith[1]
[1]Khoury College of Computer Science, Northeastern University, Boston, MA
[2]College of Arts, Media and Design, Northeastern University, Boston, MA

Abstract

We explore the task of intrinsic source attribution: inferring which portions of a derived document were adapted from an *unobserved* source document. Specifically, we model the relationship between news articles and their press release sources using a dataset of 64,784 health science news articles and 23,068 press releases. We approach the problem at the sentence level and work with science journalism professors to develop a four point Likert scale describing the extent to which a news article sentence is derived from the content in the corresponding press release. Because manual annotation of news article - press release pairs is time-consuming, we turn to a mix of expert, non-expert, and heuristic-based annotation to label our dataset. After a small pilot study, which found that humans, when only able to view the text of the news article, struggle to identify which content is derived or not, we compare four different sentence regression models on the task. We find that modeling a sentence's context in the entire document is important, with the best performing model, a sequence regression model with BERT token representations, achieving a spearman's ρ of 0.49 and $NDCG@1$ of 0.60 on the expert-labeled test set. Examining the model's predictions, we find that it successfully identifies copied or closely paraphrased sentences in articles with a mix of derived and original content, but struggles to differentiate between loosely paraphrased and original sentences in articles with mostly original writing.

1 Introduction

Authors typically rely on a wide array of sources when writing a new document. These source documents serve a variety of purposes, from simply being general inspiration, to containing specific text and ideas to paraphrase, summarize, and potentially copy. A historian, for example, may quote, paraphrase, and discuss a number of primary and secondary sources in a journal article. Similarly, a blogger posting about a recent newsworthy event could quote, discuss, and be inspired by news articles and blog and social media posts on the same topic. Tracking this flow of information and ideas across a corpus of texts can be difficult, however, since parent-child relationships between documents are often unclear and similarity between documents can range from duplication to paraphrasing and topical resemblance (Hamid et al. 2009; Metzler et al. 2005). Approaches using social media and news data (Leskovec, Backstrom, and Kleinberg 2009; Seo and Croft 2008) track the spread of short pieces of text ("memes") by measuring similarity between different copies of the same meme across documents. Other approaches use word overlap to trace longer instances of text reuse, such as in congressional bills (Wilkerson, Smith, and Stramp 2015) and 19th C. newspapers (Smith, Cordell, and Dillon 2013).

In contrast to work on retrieving and analyzing reused text, we propose to perform the task of **source attribution**, inferring which portions of a document were derived from a source, from internal evidence alone, i.e., when the source documents are **not** available at test time. As described in the results of our pilot study (§6), this task is difficult for humans – without viewing the underlying source(s), it is not clear what portions of a document are derived or newly written. Though difficult, this problem setting is more realistic since many authors do not fully cite sources and some source documents are unpublished or not easily accessed. For instance, given a senator's press release after a closed-door meeting, we might infer which portions may have been copied from the meeting's unreleased talking points. Similarly, given a recently published news article, determining which sections of the article were adapted from press releases or other articles would help to identify sources of fake news and misinformation present in social and news media (Southwell, Thorson, and Sheble 2017; Vosoughi, Roy, and Aral 2018; Chou, Oh, and Klein 2018).

Validating results from a source attribution model, however, is difficult, as the source documents are often unpublished and the relationships between documents unknown. Therefore, we focus on a specific domain, health science news articles, for which the relationship between documents is somewhat clearer and some of the underlying source documents with apparent relationships to derived documents are published widely. When writing a science news article, journalists rely on a variety of sources, including, but not limited to, the scientific article itself, press releases issued by the university or journal, interviews with prominent scientists, reports by government agencies, and other news articles on the topic (Wihbey 2019; Len-Rios et al. 2009). Modeling how journalists write a piece, therefore, is hard, as they not only combine multiple input sources, but they may dramatically transform them, simplifying, paraphrasing, and explaining. Further, we often don't have access to all of the sources a journalist may use, such as interview transcripts and emails. Therefore, we simplify the problem, focusing only on modeling the relationship between university or journal press releases and news articles written about the same scientific article. Press releases are an important source for journalists as their mere existence indicates potentially important research (Kiernan 2003b; Wihbey 2019), and they serve as a high level summary of the work.

In the domains of journalism and science communication, in particular, there are ongoing debates about how to more accurately render new research findings to the public and to improve what has been called the "science of science communication" (National Academies of Sciences, Engineering, and Medicine 2017; Bubela et al. 2009). Greater understanding of how foundational knowledge is processed and disseminated to the public remains vitally important as researchers, policy makers, and public officials attempt to grapple with rampant misinformation on topics such as vaccine effectiveness and human-induced climate change (Southwell, Thorson, and Sheble 2017; Chou, Oh, and Klein 2018).

Exaggerated claims in science-related press releases themselves, as well as faulty replication and conveyance of scientific research information by journalists, can lead to compounding misinformation for the public (Sumner et al. 2014; Caulfield et al. 2016). Therefore, research on the informational network connecting academic findings, public relations materials such as press releases, and news media publication may be highly useful to journalists who are looking to improve their practice and better inform their audiences, as well as to consumers, media critics, and other watchdogs who are on guard for potentially damaging misinformation and the process by which it is generated.

2 Problem Formulation

In this paper we explore the problem of source attribution using health science news data, predicting which sentences of a news article (NA) have likely been adapted from the underlying press release (PR). Concretely, given a newly published NA, we split it into sentences, then, without access to the text of the PR, predict, for each sentence, how likely it is that its content is either derived from the source PR, wholly novel, or somewhere in between. We label our data on a 0-3 Likert scale developed in collaboration with 2 science journalism professors (§5). Unlike common paraphrase and plagiarism detection tasks, which assume access to both a source and derived document, we assume that the underlying source PR is *not* available at test time. We argue that is a more realistic version of the problem since many science journalists do not exhaustively cite and link to the full text of their sources. We perform our analysis and train our models using 64,784 NAs and 23,068 PRs written about 20,271 scientific articles.

To better understand the task, consider these excerpts from 6 different PR-NA pairs with differing amounts of content overlap, paraphrasing, and summarization:

Example 1: Label 2

- **PR:** The taller you are, the more likely you may be to develop blood clots in the veins, according to new research in the American Heart Association journal Circulation: Cardiovascular Genetics. In a study of more than two million Swedish siblings, researchers found that the risk of venous thromboembolism - a type of blood clot that starts in a vein - was associated with height, with the lowest risk being in shorter participants.
- **NA:** Taller people are at higher risk for venous thrombosis, according to a study of siblings in Swedish national registry databases reported in Circulation: Cardiovascular Genetics.

Example 2: Label 2

- **PR:** Neurologists have long believed RLS is related to a dysfunction in the way the brain uses the neurotransmitter dopamine, a chemical used by brain cells to communicate and produce smooth, purposeful muscle activity and movement. ... The small new study, headed by Richard P. Allen, Ph.D., an associate professor of neurology at the Johns Hopkins University School of Medicine, used MRI to image the brain and found glutamate – a neurotransmitter involved in arousal – in abnormally high levels in people with RLS.
- **NA:** New research suggests that insomnia caused by restless leg syndrome (RLS) is strongly linked to high levels of the brain chemical glutamate, contradicting long - held assumptions that the neurotransmitter dopamine is the main culprit of the symptoms.

Example 3: Label 3

- **PR:** The study found significant improvements among participants in mental health, aerobic endurance and outcome expectations for exercise (for example, perceived benefit of exercise participation), based on assessments completed by the participants.
- **NA:** Significant improvements were also found among participants in mental health, aerobic endurance and outcome expectations for exercise.

Example 4: Label 3

- **PR:** Preeclampsia is a complex form of high blood pressure in pregnancy that can damage the kidneys, liver and brain and lead to fetal complications such as premature delivery, low birth weight and stillbirth.
- **NA:** Pre-eclampsia can damage the kidneys, liver and brain, and lead to foetal complications such as premature delivery, low birth weight and stillbirth, experts say.

Example 5: Label 1

- **PR:** Tresiba(r) Trial Shows that People with Type 2 Diabetes who Avoid Severe Hypoglycaemia have a Reduced Risk of Death. Novo Nordisk today announced new analyses from the multinational, double-blinded DEVOTE trial showing that people with type 2 diabetes who experience severe hypoglycaemia (low blood sugar levels) are at greater risk of death.
- **NA:** In adults with type 2 diabetes, higher day-to-day fasting glycemic variability and severe hypoglycemia are independently associated with all-cause mortality, according to two secondary analyses from the DEVOTE trial presented at the European Association for the Study of Diabetes Annual Meeting and published simultaneously in Diabetologia.

Example 6: Label 1

- **PR:** The researchers found that PM continued studying addiction through the 2000s to develop successful and potentially safer nicotine products, and that from the mid-1990s to at least 2006, Philip Morris's internal models of addiction regarded psychological, social, and environmental factors as comparable in importance to nicotine in driving cigarette use. Elias and colleagues argue that PM's outward support for nicotine's role in driving smoking allowed the company to redirect policy away from proven social and environmental interventions and toward the promotion of potentially reduced harm industry products.
- **NA:** In other words, they said, PM's' opportunistic shift from denying to affirming nicotine's addictiveness was driven not by a substantive change in scientific understanding but by public, regulatory, and legal pressures.

In example 1, the journalist summarizes the 2 corresponding PR sentences and includes a small additional detail from the study about the Swedish national registry database. The journalist similarly summarizes two non-adjacent PR sentences in example 2. In examples 3 and 4, the journalists nearly exactly copy excerpts from the corresponding PR sentences. In example 5, the journalist paraphrases some of the content from the corresponding PR sentences, but adds significantly more specific details pulled from the actual study. Finally, in example 6, the journalist both summarizes the corresponding PR sentences and adds their own interpretation. Each example is labeled on a 0-3 Likert scale described in §5. In order to perform this labeling, annotators view the full NAs and corresponding PRs side-by-side and examine lexical and semantic similarities. At test time, however, our pilot study subjects (§6) and models (§7) make predictions using *only* the language of the NA.

3 Related Work

3.1 Automatic Fact-Checking

Related to our source attribution problem is work focused on automatic fact-checking of news content (Rashkin et al. 2017; Potthast et al. 2018). Verifying an article's claims and determining their sources are related, but distinct, tasks – for example, a sentence derived from a source could still be untrue if the author misrepresents its meaning (taking out of context, exaggerating, etc); conversely, a sentence with no clearly attributable textual source could still be factually accurate. Rather, results from both systems are complementary. For instance, to combat misinformation in anti-vaccine literature, veracity and source attribution models could be run on NAs in question, and readers could use both scores to help determine what information is trustworthy.

3.2 Text Reuse & Similarity Identification

Text Reuse Detection Measuring the flow of information and reuse of ideas and text is well studied (Hamid et al. 2009; Metzler et al. 2005), with applications from short "memes" in news and social media (Seo and Croft 2008; Leskovec, Backstrom, and Kleinberg 2009) to longer instances of reuse such as sentences, paragraphs or large portions of entire documents (Wilkerson, Smith, and Stramp 2015; Smith, Cordell, and Dillon 2013; Nicholls 2019). Many text reuse systems, however, identify reuse only through lexical similarity (e.g. word overlap, edit distance), thereby missing instances where the author of the derived document substantially changes the text of the source through paraphrase, summarization, etc. In further contrast to our source attribution task, text reuse systems also assume the presence of both the source and target documents at prediction time.

Plagiarism Detection There are two commonly studied settings for plagiarism detection: 1) extrinsic – given a new document D_n and a corpus of source documents C_s, detect which sections of D_n are plagiarized, if any, and find their sources in C_s (Belyy, Dubova, and Nekrasov 2018) 2) intrinsic – given only D_n, detect which sections of D_n are plagiarized (Eissen and Stein 2006; Potthast et al. 2009). Our source attribution task is most similar to intrinsic detection since inference is performed on just the NA. Prior work in intrinsic detection has approached the problem as binary classification on text segments (plagiarized or not), training classifiers such as Naive Bayes and Gradient Boosted Regression Trees with features such as bag-of-words (BOW), POS, and readability scores (Bensalem, Rosso, and Chikhi 2014; Rahman 2015; Stein, Lipka, and Prettenhofer 2011). Although similar to intrinsic plagiarism detection, as discussed in §5, our source attribution task captures a larger variety of source-derived relationships beyond just wrongful and unattributed use of other's language and thoughts (plagiarism). We examine multiple levels of reuse, ranging from near or exact copying to paraphrasing, summarization, interpretation, and general inspiration, regardless of if the underlying source is properly cited or not.

Paraphrase Identification As noted above, our source attribution problem includes instances where the author of the derived document paraphrases the source. There exists significant prior work in paraphrase identification (Cer et al. 2017; Dolan and Brockett 2005; Jurgens, Pilehvar, and Navigli 2014; Iyer, Dandekar, and Csernai 2017). Labeled datasets include pairs of texts with either binary or ordinal (e.g. 1-5) ratings indicating whether they are paraphrases of each other and, in the case of the ordinal ratings, to what extent. Current state-of-the-art (SoTA) models train neural architectures on this sequence-pair classification or regression task (Devlin et al. 2019; Reimers and Gurevych 2017; Liu et al. 2019). However, similar to text reuse and extrinsic plagiarism detection, since this problem setup assumes a pair of texts as input, these models would not be directly applicable to our problem. Instead, as described in §5, we explore applications of SoTA neural models to help generate training data for our source attribution system.

4 Dataset

Our dataset comes from Altmetric[1], a company that tracks mentions of scientific research online. Altmetric monitors mentions of research in over 2000 news sources of various types, including, for instance, traditional national and local news outlets, university and journal press offices, and niche science news websites, such as Colon Cancer News Today. Altmetric finds mentions of scientific articles in news articles by searching for direct hyperlinks to scholarly papers and by extracting potential journal, article and author names and performing a search on CrossRef's scientific article database (English news articles only).

Altmetric has provided us with a database snapshot from October 8th, 2019. This contains metadata (title, abstract, journal, etc.) for 26,222,754 scientific articles along with URLs of corresponding news and social media articles. Most of the scientific articles, however, do not garner news coverage. For our analysis, we first filter this dataset to 5,649,276 health science articles (articles with categorized as "Health Science" by Scopus), then remove scientific articles with no news coverage, leaving 343,245 scientific articles with 1,277,646 distinct news article URLs. We use lazyNLP[2] to crawl and extract article content, and successfully crawl 544,777 articles. We do not download the full text of the scientific articles due to copyright limitations.

Data Cleanup As we aim to model how journalists transform PRs into new content, we need to first separate our list of articles into PRs and NAs. For PRs, we utilize results from a survey of science journalists (Wihbey 2019) to select a list of 13 PR publishers and aggregators[3] and mark the rest as NAs, yielding 56,369 PRs and 488,408 NAs. For each NA, we identify all PRs written about the same scientific article(s) as potential sources and remove NAs and PRs with no corresponding source or target. This yields 49,725 PRs and 238,029 NAs written about 38,656 scientific articles. Inspecting the dataset, however, we find that it contains many NAs which are either near copies of other NAs or related PRs, or off-topic boilerplate. Since we want to model how journalists transform source content and create a new piece, we do not want to include articles from outlets which are copies of other documents or unrelated to the PR. Thus, based on manual inspection of the dataset, we devise a set of document- and sentence-level word-overlap rules to cleanup the data. We split documents into sentences and tokenize using spaCy (Honnibal and Montani 2017), then remove stopwords and punctuation. At the document level, we transform NAs and PRs into TF-IDF weighted BOW. At the sentence level, we transform each sentence in a document into its own TF-IDF weighted BOW.

Split	# Docs	# Sentences	Avg. Sent Length
Train	58,134	1,770,348	25.2
Dev	6,600	179,945	25.6
Test	50	1,140	28.0

Table 1: Dataset Summary Statistics

To remove NAs which copy substantial PR content, we remove NAs with document-level cosine similarity > 0.8 to any corresponding PR or where $> 20\%$ of its sentences have similarity > 0.8 to any PR sentence. To find NAs that are unrelated to the PRs, we remove NAs with document-level similarity < 0.2 to all relevant PRs or where $< 20\%$ of its sentences have similarity > 0.2 to any PR sentence. The sentence-level cutoffs are necessary to identify copying NAs with uncleaned boilerplate content and unrelated NAs, such as those on a related topic that only briefly mention the underlying scientific paper as related work. In order to deduplicate the NAs, we group them by scientific article and remove those with document-level similarity above 0.8 to another NA. We similarly deduplicate PRs.

Finally, we manually inspect a sample of articles from each of the 24 outlets with at least 500 articles in the dataset. This includes major outlets such as NYT, CNN, and BBC, newswires such as UPI, and science-specific outlets such as The ASCO Post and Healio. We identify 3 PR-copying outlets and remove all of their corresponding $\approx$ 4k NAs. This leaves our final dataset of 64,784 NAs and 23,068 PRs written about 20,271 scientific articles. We cap each NA and PR at 75 sentences, greater than 95% of all documents.

Dataset Splits We split our dataset into train, dev, and test sets so that no pair of NAs, for example, in the train and test sets, have the same corresponding PR(s) or cite the same scientific articles. We create a test set of 50 articles for expert annotation by journalists (§5.2) and split the remaining 64,734 articles into train and dev (Table 1). NAs contain, on average, 30 sentences with mean 25 tokens per sentence.

5 Annotation

5.1 Annotation Scale Development

We examine the task of attributing portions of a derived document to a source document at the sentence-level. For each NA sentence, we wish to label it on a scale indicating how original vs. derived its content is. To this end, we recruit two science journalism professors (both authors of this paper) to examine a sample of 5 NA-PR pairs. Following work in semantic similarity scale development (Jurgens, Pilehvar, and

[1] altmetric.com/audience/researchers

[2] https://github.com/chiphuyen/lazynlp

[3] EurekAlert!, Science Daily, AlphaGalileo, Newswise, Nature, PR Newswire, Journalist's Resource, Science News, Science/AAAS, Kaiser Health News, Business Wire, Mayo Clinic, The New England Journal of Medicine.

Navigli 2014; Cer et al. 2017), we develop and validate a 4 point Likert scale:

0. Novel or unrelated: not derived from passage(s) in the press release.
1. Partially derived: journalist has used the PR as a source, but has substantially changed or added to the content.
2. Mostly derived: journalist has paraphrased or repackaged parts of the PR and not added much new content.
3. Derived: journalist has nearly or exactly copied the PR.

This is a difficult annotation task, even for journalists – it is easy to miss similarities between documents, and there can be too much to keep in mind at a time going back and forth between the full NA and PR texts. To aid the annotation effort, we iterate with the journalists to devise a small annotation tool which displays a NA and all corresponding source PRs side-by-side. When an annotator clicks on a NA sentence, all of the overlapping tokens in the PRs are highlighted – all overlapping tokens are colored using a sequential, single hue (blue) color scale where rarer words (low document frequency) are highlighted using a darker shade of blue. Further, when a span of at least 3 tokens occurs in both the NA sentence and one of the PRs, that span of tokens in the PR is bold faced for easier identification. Though this tool makes it easier to identify lexically similar content between a PR and NA, NA sentences that are derived from the PR, but paraphrased, might be missed. Therefore, each annotator is instructed to read each PR and NA in full before beginning annotation and to re-examine the entire PR for both lexical and semantic similarities when labeling each NA sentence. For reference, the NA-PR pairs in §2 contain examples of NA sentences with labels 1-3.

5.2 Train-Dev-Test Annotations

Since we wish to build a model that learns to identify which portions of a NA are derived from the corresponding PR, we need labeled data, with each NA sentence labeled on our 0-3 scale. However, labeling each NA requires a substantial amount of time (≈ 15 - 20 min.) and thought. In order to scale to our dataset of ≈ 65k NAs, we turn to a mix of expert and non-expert human annotation along with heuristic-based methods. Specifically, we evaluate our models on a test set labeled by expert journalists, and we use a set of non-expert annotations on a sample of our dev set to find an automatic metric to label our training data.

Test Set: Expert Annotation To evaluate our model on the cleanest data possible, we create a test set of 50 NAs (1,140 total sentences) for expert annotation. We manually clean these 50 articles, ensuring that all boilerplate content is removed. The same two science journalism professors who developed the annotation scheme also annotated the test set. Each journalist annotated 25 NAs. To evaluate annotator consistency, one journalist annotated 3 randomly selected NAs (58 sentences) from the other's list of 25. We use tie-corrected Spearman rank correlation (ρ) to assess agreement, which is 0.74.

As seen in table 2, the dataset is imbalanced, with 67% of sentences labeled 0. However, nearly half (24) of the NAs have at least one label 3 sentence – 16 NAs with at least one label 2, and 10 with at least one label 1. The median NA has 39% of its sentences labeled as derived (labels 1-3). Figure 1 shows the average label at each quintile of an NA – we sort each NA's sentences by label then calculate the average label at each quintile in the ranked list. The top ranked sentence has an average label of 2.28, with a gradual decrease to an average label of 0.1 for the least derived sentence. In order to examine whether sentence position in the NA has any relationship to its label, we iterate over each NA and calculate the relative [0,1] positions of its derived sentences (labels 1-3). We find that the derived sentences occur slightly more frequently in the first half of the NA, but are mostly uniformly distributed.

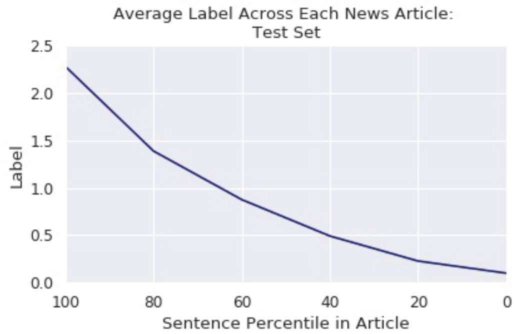


Figure 1: Test set: avg. sentence label across each quintile.

	Label 0	Label 1	Label 2	Label 3
Dev	862 (66%)	204 (15%)	201 (15%)	46 (4%)
Test	763 (67%)	172 (15%)	92 (8%)	113 (10%)

Table 2: Breakdown of the total number of label 0, 1, 2 and 3 sentences in each of the human-labeled datasets – 1,313 total dev sentences and 1,140 test sentences.

Train & Dev Sets: Non-Expert Annotation & Heuristics Though we have created an expert-labeled, clean test set for evaluation, we still need labeled data to train models. In order to scale to tens of thousands of NAs, we turn to text similarity heuristics. Many measures, such as ROUGE (Lin 2004) and BLEU (Papineni et al. 2002), and sentence embedding models, such as Universal Sentence Encoder (Cer et al. 2018), exist to measure similarity between two pieces of text. However, it is not clear a priori which metric most closely resembles human judgement on this task. We therefore sample 50 articles from the dev set for annotation in order to find the metric which best approximates human labels. As our expert annotators have limited time, 4 non-expert, fluent English speakers (two of them authors of this paper) annotated data for this task. Using the same tool and instructions as the expert journalists, each annotator annotated 12 or 13 NAs for a total of 50. To evaluate the agreement of the annotations, we randomly select 3 NAs (112 sentences) for annotation by all 4 annotators. The average pairwise ρ is 0.75, nearly identical to the 0.74 ρ between the journalists.

Inspecting the annotations, we discover 5 NAs with label "0" for all of their sentences (none of them were used to calculate annotator agreement). We find that these NAs are on related topics to the corresponding PRs, but focus on separate, distinct scientific articles. Since there is no source-target relationship between these pairs, we exclude them from our analysis, yielding a labeled dataset of 1,313 sentences across 45 NAs. As can be seen in table 2, this dataset also has a heavy class imbalance: 66% of sentences are labeled 0. However, we also find that there is a difference between the percentage of label 2 and 3 sentences between the dev and test sets. Examining the labels and their corresponding sentences, we find that although this is partially due to chance, with the test set containing articles that copy more content, there is discrepancy between the groups of annotators on how to label sentences with high content overlap that fall between a 2 and 3. The journalists were more likely to label such sentences as 3s (near exact copies), whereas the non-experts as 2s (close paraphrases).

We then automatically generate a label for each of the 1,313 sentences as follows – we calculate its similarity (under some unsupervised model) to each sentence in the corresponding PR(s), and label it with the maximum similarity score from across all of the PR sentences. We opt for this all-sentence-pairs comparison since PRs in our dataset are very long (avg. 31 sentences), and each NA sentence is likely only derived from and similar to a specific subset of the PR, if at all. This setup, though, will miss NA sentences that partially excerpt, paraphrase, and/or summarize content from multiple, sometimes non-adjacent, PR sentences. We performed initial experiments calculating the similarity between a NA sentence and an entire PR document using TF-IDF cosine similarity (ie. vector space model for information retrieval), but achieved better results with the all-sentence-pair comparisons – we thus stick with this setup for the following experiments. We leave applications of supervised cross-level (document-to-sentence) semantic similarity identification and retrieval systems to future work.

Following work in paraphrase identification (Jurgens, Pilehvar, and Navigli 2014; Cer et al. 2017; Huang and Chang 2014; Sarkar et al. 2016; Ferrero et al. 2017) and text reuse detection (Metzler et al. 2005), we calculate similarity scores between each of the 1,313 NA sentences and their corresponding PR sentences under 8 similarity metrics – five using BOW representations and three using continuous representations:

- ROUGE: -1, -2, and -L, with and without stemming (Lin 2004)
- BLEU (Papineni et al. 2002)
- Meteor (Banerjee and Lavie 2005)
- Token-level Levenshtein similarity $= 1 -$ (Levenshtein distance/ length of longer sequence)
- Bag-of-ngrams cosine similarity, $n \in \{1, 2, 3\}$, with and without TFIDF-weighting
- Word Mover's Distance (WMD) (Kusner et al. 2015) with GloVe vectors (Pennington, Socher, and Manning 2014). We use negative WMD to ensure that more similar sequences have higher (closer to zero) scores.
- Universal Sentence Encoder (USE) (Cer et al. 2018) cosine similarity[4]
- Sentence-BERT[5] (Reimers and Gurevych 2019) cosine similarity

Heuristic	ρ
ROUGE-1	59.0
ROUGE-2	59.2
ROUGE-L	55.4
BLEU	53.0
Meteor	57.0
Levenshtein	47.8
Bag-of-ngrams	61.2
TFIDF-ngrams	**62.5**
WMD	53.5
USE	59.1
Sent-ROBERTA	59.1
Sent-BERT	52.5

Table 3: Spearman's $\rho \times 100$ between heuristic-derived similarities and non-expert labels on 1,313 sentences from the dev set.

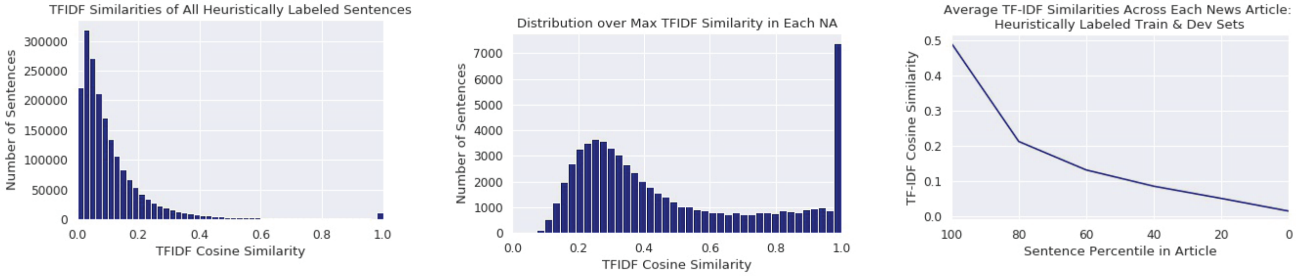
For each metric, we measure the similarity between its scores and the ground truth, non-expert human labels using tie-corrected ρ (Table 3). Correlation is commonly used to measure the similarity between human-rated Likert scale similarity judgements and machine output similarity scores (Cer et al. 2017; Reimers, Beyer, and Gurevych 2016; Jurgens, Pilehvar, and Navigli 2014; Reimers and Gurevych 2019). We opt for ρ since it is not sensitive to outliers, non-linear relationships or non-normally distributed data, unlike Pearson's correlation (Reimers, Beyer, and Gurevych 2016; Zesch 2010).

As seen in Table 3, we find that TFIDF-weighted bag of n-grams (unigrams and bigrams, no stopping or stemming) cosine similarity has the highest absolution correlation, with a ρ of 0.625 with the human labels. Thus, we select it generate labels for each of the 1,948,833 sentences from the 58,134 train and 6,550 dev NAs without human-annotated labels. We follow the same all-sentence-pairs setup as above, labeling each sentence with the maximum TFIDF cosine similarity across the corresponding PR sentences.

Heuristically-Labeled Data Description Figure 2a shows the distribution over the TF-IDF similarity scores for all 1,948,833 heuristically-labeled sentences. Approximately 61% of sentences have similarity ≤ 0.1 Only approximately 8% have similarity scores of at least 0.3, 5% of at least 0.5, and 1% of at least 0.9. These percentages are substantially lower than the proportions of label 1-3 sentences in the human-labeled dataset. This is due to multiple

[4] https://tfhub.dev/google/universal-sentence-encoder/3

[5] Tested the 2 best performing models on the STS benchmark data: bert-large-nli-mean-tokens and roberta-large-nli-stsb-mean-tokens. https://github.com/UKPLab/sentence-transformers#pretrained-models



(a) TF-IDF similarities of all heuristically-labeled sentences

(b) Maximum TF-IDF similarities in each NA

(c) Average TF-IDF similarities across each NA.

Figure 2: Distributions over TF-IDF cosine similarities in heuristically labeled sentences in train and dev sets.

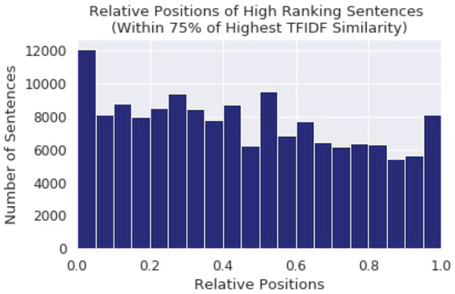


Figure 3: Relative positions of sentences in NAs with high cosine similarity scores in train and dev sets.

factors, but most significantly because journalists often summarize, paraphrase and simplify PR content. A BOW model applied only at the sentence-pair level cannot capture these complex relationships. However, as demonstrated by the ρ of 0.625 with the human labels, the BOW method still ranks sentences relatively well – on average, sentences that are labeled as derived (labels 1-3) have higher TF-IDF scores than the the unrelated or novel sentences (label 0).

Figure 2b shows the distribution over the max TF-IDF score across each NA. Of the 64,684, approximately 16% contain a sentence with a similarity score of at least 0.9. The mean NA contains a maximum sentence-level similarity of 0.49 (median 0.38). To generate figure 2c, we sort each NA's sentences by similarity score and calculate quintiles as we did for the test set. Average similarity at the 80th percentile is 0.2, then declines steadily to nearly 0.

Similar to the test data, we also examine the relationship between sentence position and TF-IDF similarity. We find the relative [0,1] positions of each sentence with similarity $\geq 75\%$ the maximum sentence similarity in that NA. As seen in figure 3, similar to the test data, these relatively high similarity sentences occur slightly more frequently in the first half of the NA, but also with spikes at the start (5% of top scoring sentences) and end (4%) of the NA.

6 Task Difficulty: Pilot Study

To assess the difficulty of this task, we conducted a small pilot study. We recruited 5 fluent English speakers to perform the source attribution task – given a NA, predict which sentences are derived from the unseen, underlying press release using *only* the text of the NA. None of the participants are authors of the paper or were previously involved in the labeling effort. We sampled 8 NAs from the test set for the study – 3 to familiarize participants with the task and 5 for evaluation. Participants first examined the 3 example NAs, their corresponding PRs and the expert sentence labels using the same tool used by the annotators. Then, for each of the 5 evaluation NAs, the participants each predicted a 0-3 score for all of the sentences (121 total sentences) without looking at the PRs. As in §5.2, we use tie-corrected ρ to evaluate the pilot participant's predicted scores vs. the ground truth. Since, unlike the text similarity metrics used to label the training data, participants are making integer predictions on the same ordinal scale as the labels, we also evaluate using F1 score. We compute F1 score under two settings: binary F1 (derived vs. not) by collapsing labels 1-3 to a single positive label, and macro-averaged multi-class F1.

As can be seen in table 4, with a maximum ρ of 0.4, binary F1 of 0.626 and multiclass F1 of 0.361 across all participants, this task is difficult. In general, participants overestimated the proportion of derived sentences (labels 1-3) and struggled to differentiate between the different levels of derived content (hence the similarities in multiclass F1 scores). Some participants, however, excelled with respect to others in differentiating between non-derived and derived content, leading to the disparities in ρ and binary F1. Participants noted that they thought usage of direct quotes and listing of specific numbers and figures were signals of content copying, but that this intuition was not perfect since journalists could use a variety of sources while writing an article (the original scientific article, interviews with non-affiliated experts, other news articles, etc).

7 Models

As demonstrated by our pilot study results, non-expert human readers cannot highly accurately identify which sentences in a NA are derived from the underlying PR. We thus argue that a model trained to identify derived PR content

Subject	ρ	F1 (binary)	F1 (multiclass macro)
A	25.5	58.3	28.4
B	25.7	55.4	31.7
C	40.0	61.0	36.1
D	32.6	62.6	31.9
E	13.7	32.9	33.0

Table 4: Pilot study results: 5 non-expert raters predicted 0-3 labels for sentences in a sample of 5 expert-labeled NAs from our test set. ρ, F1 $\times$100.

would be useful for such readers to help them better understand and analyze the NAs they read. We evaluate the performance of 4 different models on this task – 2 neural models and 2 feature-based models. As noted in §3, although our task is quite similar to other, common NLP tasks such as paraphrase identification and extrinsic plagiarism detection, since the source PR is not available at test time, we are unable to apply models designed for these tasks to our dataset.

In order to generate labels for our training and dev data that are on a similar scale as our 0-3 human labels, we multiply each NA sentence's TF-IDF cosine similarity score by 3.

Fine-Tuned BERT We select BERT$_{BASE}$ (Devlin et al. 2019) as BERT-based architectures have recently achieved SoTA performance on a variety of NLP tasks, including sentence classification and sequence tagging (Devlin et al. 2019). Optimally, we would use a model that could efficiently and effectively train on whole NAs, extracting a contextual representation for every sentence and making a prediction for each one. However, the NAs in our dataset are long (average 857 WordPiece tokens), well above the 512 WordPiece limit of the pre-trained BERT checkpoint released by Google[6]. We thus fine-tune BERT on individual NA sentences. Following Devlin et al. (2019), we feed each WordPiece tokenized (Wu et al. 2015) NA sentence into BERT independently, use the final [CLS] embedding as the representation for the entire sentence, and feed that into a linear layer to make our final regression prediction. We train the models using Adam (Kingma and Ba 2014) and, following work on ordinal regression with class-imbalanced data (Baly et al. 2019; Rosenthal, Farra, and Nakov 2017), optimize mean absolute error (MAE) loss.

BCL: BERT-CNN-LSTM As noted above, we hypothesize that learning representations of NA sentences that are sensitive to the entire document context will improve performance. Thus, instead of operating on each NA sentence independently, our model must take an entire NA as input, then predict a score for each sentence. We use BERT as the basis of our model and, again, due to the sequence length limitations, input each NA sentence to BERT independently. Further, due to memory constraints of fine-tuning BERT with all of the sentences of an NA in a single batch, we opt for the more efficient feature-based approach, which has been shown to have comparable performance to fine-tuning on some tasks (Peters, Ruder, and Smith 2019). Use of feature-based BERT requires a task-specific architecture on top of the extracted embeddings. Much of the prior work on feature-based BERT for sequence tagging (Devlin et al. 2019; Peters, Ruder, and Smith 2019) has focused on token-level tasks, such as NER, and has utilized a BiLSTM (and sometimes CRF) as the task-specific architecture. As we are making predictions on sequences of sentences, we must first aggregate a sentence's token representations into one sentence-level representation. Thus, we extract and concatenate BERT's last 4 hidden layers (dimensionality 3072) as features for each token, then use a CNN to learn a sentence-level representation. We select CNNs as they have proved effective on many sentence-level classification tasks (Kalchbrenner, Grefenstette, and Blunsom 2014; Kim 2014; Johnson and Zhang 2015). We use ReLU as the CNN activation function and 1-max pooling over time to learn a fixed-length representation of each sentence. We then pass all sentence representations for a given NA through a single layer BiLSTM (dim. 200 each direction) to encode each sentence with information about surrounding sentences. We use a final linear layer for prediction. We train to minimize MAE. For clarity, figure 4 shows the design of the entire model.

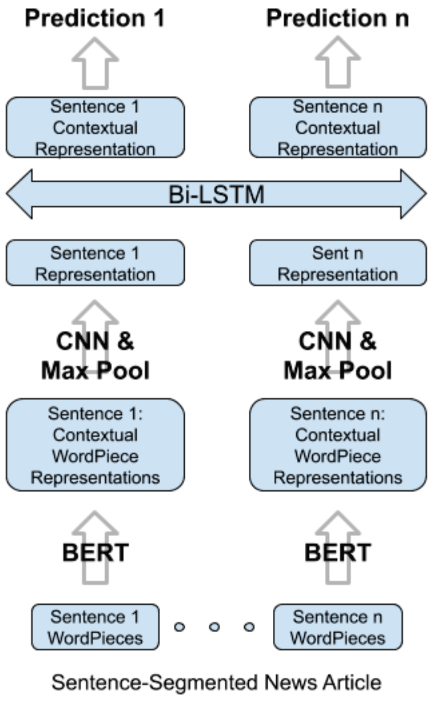

Prediction 1
Prediction n
Sentence 1 Contextual Representation
Sentence n Contextual Representation
Bi-LSTM
Sentence 1 Representation
Sent n Representation
CNN & Max Pool
CNN & Max Pool
Sentence 1: Contextual WordPiece Representations
Sentence n: Contextual WordPiece Representations
BERT
BERT
Sentence 1 WordPieces
Sentence n WordPieces
Sentence-Segmented News Article

Figure 4: BCL model design.

Feature-Based Models: SVR & GBR We explore two feature-based models: Support Vector Regression and Gradient Boosted Regression Trees, both trained with MAE loss functions. Similar to BERT (and unlike BCL), we train each feature-based model on individual sentences. Thus, they cannot leverage the text of surrounding NA sentences when

[6] storage.googleapis.com/bert_models/2018_10_18/uncased_L-12_H-768_A-12.zip

	Corpus-Level		$NDCG_{doc}$		
	ρ	$NDCG@1140$	@1	@3	@5
In-Order	–	–	24.3	36.0	40.1
SVR	37.7	85.2	57.0	56.7	55.6
GBR	43.0	86.4	59.7	54.9	56.4
BERT	37.7	85.3	55.3	54.2	55.1
BC	40.7	85.9	57.0	56.7	55.6
BCL	**48.8**	**87.6**	**60.3**	**62.6**	**62.7**

Table 5: Results on test set – 1,140 sentences across 50 NAs. ρ and NDCG ×100.

making their predictions.

We design 7 sentence-level features: sentence length, presence of a quote, and 5 measures of position (absolute position, relative position, is 1st sentence, is last sentence, and a 1-hot vector indicating in which positional quartile in the NA the sentence occurs). We also include BOW features for each sentence, extracting counts of n-grams. The number of BOW features for each model is tuned on our dev set.

8 Evaluation Settings

Hyperparameters: For all models we select the hyperparameter configuration with lowest MAE on the dev set.

BERT: For efficiency, we limit each sentence to 57 WordPiece tokens, excluding the special [CLS] and [SEP] tokens ($>$95% of sentences). As suggested by Devlin et al. (2019), we search over: batch size $\in \{16, 32\}$, learning rate $\in$ {5e-5, 3e-5, 2e-5} and train up to 4 epochs.

BCL: We train using the same 57 WordPiece token max sentence length and Adam optimizer (learning rate 1e-3) as fine-tuned BERT. As suggested by Zhang and Wallace (2017), we perform a grid search over the hyperparameters for our convolutional layer, with filter sizes $\in \{1, 3, 5, 7, 10\}$ and feature map sizes $\in \{100, 200, 400, 600\}$. We use dropout (0.1) after the CNN and LSTM layers. We train on mini-batches of size 32 for a maximum of 10 epochs.

SVR & GBR: We perform a grid search, optimizing maximum and minimum document frequency, with/without TF-IDF weighting, with/without stopping, order of n-grams $\in \{1, 2, 3\}$, maximum number of BOW features, and model specific hyperparameters.

Metrics: We evaluate all models using ρ, as described in §5.2. Also, following Reimers, Beyer, and Gurevych (2016), we use the ranking metric $NDCG@k$ (Järvelin and Kekäläinen 2002) as another corpus-level statistic. Specifically, we rank all 1,140 sentences by their predicted scores and compute $NDCG@1140$ for the entire list.

Since users of a source attribution system would likely only use and evaluate models on individual NAs at a time, we also compute $NDCG$ scores for each NA and average those scores across documents. We argue that this document-level evaluation is more realistic than the corpus-level as it measures how well a model rates all of the sentences in an NA with respect to each other, specifically how well it identifies the *k most* derived sentence(s). We compute $NDCG_{doc}$ for $k \in \{1, 3, 5\}$.

Non-Model Baseline: As derived sentences often occur earlier in the NA (§5.2), we include predicting the sentences of a NA in their original order as a non model baseline for $NDCG_{doc}$. This is similar to the commonly used Lead baseline in summarization, since the lead paragraph of a NA summarizes and introduces the main elements of the story.

9 Results

Table 5 shows the results of the four models and In-Order baseline on the test set, as well as an ablation experiment described below. BCL outperforms all other models and the baseline across all metrics. It outperforms the next best model, GBR, by a margin of approximately 0.5 to 8% absolute across the different metrics. As $NDCG_{doc}@1$ is simply the ratio of the label of the top predicted sentence and the maximum label sentence in the NA, BCL's $NDCG_{doc}@1$ of 0.6 is approximately equal to, on average, ranking a label 2 sentence as the most derived in a NA with one or more label 3 sentences. Furthermore, BCL's strong performance relative to fine-tuned BERT, a model with significantly more trainable parameters, demonstrates the importance of BCL's ability to leverage the context of an entire NA when predicting scores for each sentence.

Document-Context Ablation Experiment In order to more directly examine the impact of BCL's ability to model document-level context, we perform an ablation experiment, training a sentence-level BC model without the document-level LSTM. Thus, each sentence is processed *independently* by feature-based BERT, the CNN and the final dense layer for prediction. We use the same MAE loss and optimal hyperparameters from the best BCL model. As seen in table 5, our hypothesis of the importance of incorporating document context is confirmed, with performance relative to BCL decreasing across each metric by $\approx$ 2-8% absolute. Interestingly, the BC model outperforms BERT across all metrics.

Error Analysis To gain intuition into the performance of BCL, we analyze its errors on the test set. We compare the 13 NAs in the bottom quartile of $NDCG_{doc}@5$ performance ($NDCG_{doc}@5 < 0.4$) to the rest of the test set (≥ 0.4). We find that the primary difference between the two sets is the proportion of derived sentences in each NA. The bottom quartile NAs contain fewer derived sentences than the rest of the test set – each bottom quartile NA has, on average, only 16% of its sentences labeled 1-3, and the average sentence label across the NAs is 0.23. Only 7 of the 13 NAs contain a sentence labeled 2 or above, and only 2 contain a label 3 sentence. This compares to averages of 52% nonzero-labeled sentences and a 1.03 sentence label across the NAs on the rest of the test set. These results indicate that identifying highly derived (label 2-3) sentences in NAs with a mix of derived and non derived sentences is easier than differentiating between low and non-derived (label 0-1) sentences in NAs with mostly original writing. Examining the performance of the other models, we find similar, but not as extreme trends (slightly higher average sentence labels and proportion of derived sentences in the bottom quartile NAs).

We also examine whether sentence position in the NA has any relationship with the predicted scores on the test set.

As noted in our analysis of the expert labels (§5.2), the derived sentences occur slightly more frequently in the first half of the NA, but are mostly uniformly distributed. For each model's predictions, similar to figure 3, we calculate the relative positions of each sentence and identify those top-scoring sentences with a predicted score $\geq 75\%$ the maximum score in that NA. We find that all 4 models overestimate scores for sentences occurring earlier in the NA. On average, approximately half of the top scoring sentences identified by each model occur in the first third of a NA. BCL's predictions are the most biased towards earlier sentences – 13% of the top scoring sentences are the first sentence in their respective NAs.

10 Conclusion

We explore the task of intrinsic source attribution, with application to inferring which sentences in a NA were likely adapted from the underlying press release. We work with two science journalism professors to develop a 4 point Likert scale to measure how derived a given NA sentence is and to create an expert-labeled test set of news articles. We find that this task is difficult for non-experts participants in our pilot study, and thus explore the applications of 4 models to assist humans. We train our models with heuristically-derived sentence labels based on each sentence's TF-IDF similarity with the PR. We demonstrate the importance of modeling document-level context, with the best performing model using a document-level LSTM to encode sentences with information about surrounding content in the NA and achieving ρ of 0.49 and $NDCG@1, 3, 5$ of 0.60, 0.63 and 0.63, respectively, on the expert-labeled test set

Source attribution-related research may have uses for journalists and media watchdogs who are keen to improve the accuracy of scientific information in the public domain. Ultimately, models that can help unpack and explain the transformation of scientific information for public consumption may be used as part of a system to identify quality journalism or misinformation.

There are several potential directions for future research. First, we can explore finer-grained modeling of the task than at the sentence level, though getting human labels on individual tokens would be difficult. Next, as our results demonstrate the importance of modeling document-level context, we could test the performance of other contextual models, such as fine-tuning BERT on a sliding window of NA sentences (Wang et al. 2019). Further, we can examine uses of our labeled data for training paraphrase, summarization, and genre-transfer models. Finally, by modeling the portions of a NA's content not traceable to a PR, we can explore what content journalists add, potentially helping to identify and examine propagation of exaggerated or false claims (Bubela et al. 2009; Sumner et al. 2014; Kiernan 2003a; Caulfield et al. 2016; Chou, Oh, and Klein 2018).

Acknowledgments

Ansel MacLaughlin was supported by a National Endowment for the Humanities Digital Humanities Advancement Grant (HAA-263837-19). We would like to thank our non-expert annotators and pilot study participants: Ryan Muther, Rui Dong, Elizabeth Brydon, Elizabeth Barr, Amanda Khoo, Andy Yu, and Preston Tunnell Wilson.

References

Baly, R.; Karadzhov, G.; Saleh, A.; Glass, J.; and Nakov, P. 2019. Multi-task ordinal regression for jointly predicting the trustworthiness and the leading political ideology of news media. In *NAACL*, 2109–2116.

Banerjee, S., and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72. Ann Arbor, Michigan: Association for Computational Linguistics.

Belyy, A.; Dubova, M.; and Nekrasov, D. 2018. Improved evaluation framework for complex plagiarism detection. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 157–162. Melbourne, Australia: Association for Computational Linguistics.

Bensalem, I.; Rosso, P.; and Chikhi, S. 2014. Intrinsic plagiarism detection using n-gram classes. In *EMNLP*, 1459–1464.

Bubela, T.; Nisbet, M. C.; Borchelt, R.; Brunger, F.; Critchley, C.; Einsiedel, E.; Geller, G.; Gupta, A.; Hampel, J.; Hyde-Lay, R.; Jandciu, E. W.; Jones, S. A.; Kolopack, P.; Lane, S.; Lougheed, T.; Nerlich, B.; Ogbogu, U.; O'Riordan, K.; Ouellette, C.; Spear, M.; Strauss, S.; Thavaratnam, T.; Willemse, L.; and Caulfield, T. 2009. Science communication reconsidered. *Nature Biotechnology*.

Caulfield, T.; Sipp, D.; Murry, C. E.; Daley, G. Q.; and Kimmelman, J. 2016. Confronting stem cell hype. *Science* 352:776–777.

Cer, D.; Diab, M.; Agirre, E.; Lopez-Gazpio, I.; and Specia, L. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 1–14. Vancouver, Canada: Association for Computational Linguistics.

Cer, D.; Yang, Y.; yi Kong, S.; Hua, N.; Limtiaco, N. L. U.; John, R. S.; Constant, N.; Guajardo-Céspedes, M.; Yuan, S.; Tar, C.; hsuan Sung, Y.; Strope, B.; and Kurzweil, R. 2018. Universal sentence encoder. In *In submission to: EMNLP demonstration*. In submission.

Chou, W.-Y. S.; Oh, A.; and Klein, W. M. P. 2018. Addressing health-related misinformation on social media. *JAMA*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Dolan, W. B., and Brockett, C. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Eissen, S. M. z., and Stein, B. 2006. Intrinsic plagiarism detection. In Lalmas, M.; MacFarlane, A.; Rüger, S.; Tombros, A.; Tsikrika, T.; and Yavlinsky, A., eds., *Advances in Information Retrieval*, 565–569. Berlin, Heidelberg: Springer Berlin Heidelberg.

Ferrero, J.; Besacier, L.; Schwab, D.; and Agnès, F. 2017. Using word embedding for cross-language plagiarism detection. In *EACL*, 415–421.

Hamid, O. A.; Behzadi, B.; Christoph, S.; and Henzinger, M. R. 2009. Detecting the origin of text segments efficiently. In *WWW*.

Honnibal, M., and Montani, I. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Huang, P., and Chang, B. 2014. SSMT:a machine translation evaluation view to paragraph-to-sentence semantic similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 585–589. Dublin, Ireland: Association for Computational Linguistics.

Iyer, S.; Dandekar, N.; and Csernai, K. 2017. First quora dataset release: Question pairs. https://www.quora.com/q/quoradata/First-Quora-Dataset-Release-Question-Pairs.

Järvelin, K., and Kekäläinen, J. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.* 20(4):422–446.

Johnson, R., and Zhang, T. 2015. Effective use of word order for text categorization with convolutional neural networks. In *HLT-NAACL*.

Jurgens, D.; Pilehvar, M. T.; and Navigli, R. 2014. SemEval-2014 task 3: Cross-level semantic similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 17–26. Dublin, Ireland: Association for Computational Linguistics.

Kalchbrenner, N.; Grefenstette, E.; and Blunsom, P. 2014. A convolutional neural network for modelling sentences. In *ACL*.

Kiernan, V. 2003a. Diffusion of news about research. *Science Communication* 25:3–13.

Kiernan, V. 2003b. Embargoes and science news. *Journalism & Mass Communication Quarterly* 80:903–920.

Kim, Y. 2014. Convolutional neural networks for sentence classification. In *EMNLP*.

Kingma, D., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kusner, M. J.; Sun, Y.; Kolkin, N. I.; and Weinberger, K. Q. 2015. From word embeddings to document distances. In *ICML*.

Len-Rios, M. E.; Hinnant, A.; Park, S.-A.; Cameron, G. T.; Frisby, C. M.; and Youngah, L. 2009. Health news agenda building: Journalists' perceptions of the role of public relations. *Journalism & Mass Communication Quarterly* 86.

Leskovec, J.; Backstrom, L.; and Kleinberg, J. 2009. Meme-tracking and the dynamics of the news cycle. In *KDD*, 497–506.

Lin, C.-Y. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized BERT pretraining approach. *arxiv preprint: 1907.11692*.

Metzler, D.; Bernstein, Y.; Croft, W. B.; Moffat, A.; and Zobel, J. 2005. Similarity measures for tracking information flow. In *CIKM*.

National Academies of Sciences, Engineering, and Medicine. 2017. Communicating science effectively: A research agenda. *The National Academies Press*.

Nicholls, T. 2019. Detecting textual reuse in news stories, at scale. *International Journal of Communication*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 311–318.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *EMNLP*, 1532–1543.

Peters, M. E.; Ruder, S.; and Smith, N. A. 2019. To tune or not to tune? adapting pretrained representations to diverse tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, 7–14. Florence, Italy: Association for Computational Linguistics.

Potthast, M.; Stein, B.; Eiselt, A.; universität Weimar, B.; Barróncedeño, A.; and Rosso, P. 2009. P.: Overview of the 1st international competition on plagiarism detection. In *SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09), CEUR-WS.org*, 1–9.

Potthast, M.; Kiesel, J.; Reinartz, K.; Bevendorff, J.; and Stein, B. 2018. A stylometric inquiry into hyperpartisan and fake news. In *ACL*, 231–240.

Rahman, R. 2015. Information theoretical and statistical features for intrinsic plagiarism detection. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 144–148. Prague, Czech Republic: Association for Computational Linguistics.

Rashkin, H.; Choi, E.; Jang, J. Y.; Volkova, S.; and Choi, Y. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *EMNLP*, 2931–2937.

Reimers, N., and Gurevych, I. 2017. Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging. *arxiv preprint: 1707.09861*.

Reimers, N., and Gurevych, I. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *EMNLP-IJCNLP*, 3980–3990.

Reimers, N.; Beyer, P.; and Gurevych, I. 2016. Task-oriented intrinsic evaluation of semantic textual similarity. In *COLING*, 87–96.

Rosenthal, S.; Farra, N.; and Nakov, P. 2017. SemEval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 502–518. Vancouver, Canada: Association for Computational Linguistics.

Sarkar, S.; Das, D.; Pakray, P.; and Gelbukh, A. 2016. JUNITMZ at SemEval-2016 task 1: Identifying semantic similarity using Levenshtein ratio. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 702–705. San Diego, California: Association for Computational Linguistics.

Seo, J., and Croft, W. B. 2008. Local text reuse detection. In *SIGIR*.

Smith, D. A.; Cordell, R.; and Dillon, E. M. 2013. Infectious texts: Modeling text reuse in nineteenth-century newspapers. In *IEEE Workshop on Big Data and the Humanities*.

Southwell, B. G.; Thorson, E. A.; and Sheble, L. 2017. The persistence and peril of misinformation. *American Scientist*.

Stein, B.; Lipka, N.; and Prettenhofer, P. 2011. Intrinsic plagiarism analysis. *Language Resources and Evaluation* 45(1):63–82.

Sumner, P.; Vivian-Griffiths, S.; Boivin, J.; Williams, A.; Venetis, C.; Davies, A.; Ogden, J.; Whelan, L.; Hughes, B.; Boy, F.; and et al. 2014. The association between exaggeration in health related science news and academic press releases: retrospective observational study. *BMJ* 349.

Vosoughi, S.; Roy, D.; and Aral, S. 2018. The spread of true and false news online. *Science*.

Wang, Z.; Ng, P.; Ma, X.; Nallapati, R.; and Xiang, B. 2019. Multi-passage BERT: A globally normalized BERT model for open-domain question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5878–5882. Hong Kong, China: Association for Computational Linguistics.

Wihbey, J. P. 2019. *The social fact: news and knowledge in a networked world*. The MIT Press.

Wilkerson, J.; Smith, D. A.; and Stramp, N. 2015. Tracing the flow of policy ideas in legislatures: A text reuse approach. *American Journal of Political Science*.

Wu, Y.; Schuster, M.; Chen, Z.; Le, Q. V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; Klingner, J.; Shah, A.; Johnson, M.; Liu, X.; Kaiser, L.; Gouws, S.; Kato, Y.; Kudo, T.; Kazawa, H.; Stevens, K.; Kurian, G.; Patil, N.; Wang, W.; Young, C.; Smith, J.; Riesa, J.; Rudnick, A.; Vinyals, O.; Corrado, G.; Hughes, M.; and Dean, J. 2015. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Zesch, T. 2010. *Study of Semantic Relatedness of Words Using Collaboratively Constructed Semantic Resources*. Ph.D. Dissertation, Technische Universitat, Darmstadt.

Zhang, Y., and Wallace, B. 2017. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. In *IJCNLP*.