Quotebank: A Corpus of Quotations from a Decade of News

Timoté Vaucher
EPFL
timote.vaucher@epfl.ch

Andreas Spitz
EPFL
andreas.spitz@epfl.ch

Michele Catasta
Stanford University
pirroh@cs.stanford.edu

Robert West
EPFL
robert.west@epfl.ch

ABSTRACT

We present *Quotebank,* an open corpus of 178 million quotations attributed to the speakers who uttered them, extracted from 162 million English news articles published between 2008 and 2020. In order to produce this Web-scale corpus, while at the same time benefiting from the performance of modern neural models, we introduce *Quobert,* a minimally supervised framework for extracting and attributing quotations from massive corpora. Quobert avoids the necessity of manually labeled input and instead exploits the redundancy of the corpus by bootstrapping from a single seed pattern to extract training data for fine-tuning a BERT-based model. Quobert is language- and corpus-agnostic and correctly attributes 86.9% of quotations in our experiments. Quotebank and Quobert are publicly available at https://doi.org/10.5281/zenodo.4277311.

ACM Reference Format:
Timoté Vaucher, Andreas Spitz, Michele Catasta, and Robert West. 2021. Quotebank: A Corpus of Quotations from a Decade of News. In *The Fourteenth ACM International Conference on Web Search and Data Mining (WSDM '21), March 8–12, 2021, Virtual Event, Israel.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3437963.3441760

1 INTRODUCTION

> *"Quotations will tell the full measure of meaning, if you have enough of them."* —James Murray

This is a sentiment that might resonate with an academic reader, even without attribution. However, the weight of these words is likely to increase with the awareness that they were given voice by lexicographer James Murray, the original editor of the Oxford English Dictionary. While quotations can be important pieces of wisdom by themselves, their meaning is enriched by the context of an attributed speaker. As a result, and not least to follow Murray's paradigm of collecting *enough* quotations, the automated extraction and attribution of quotations is the subject of ongoing research, which can be considered a special case of relation extraction [10].

Prior work on this topic largely utilized either supervised machine learning approaches on the one hand, or unsupervised pattern matching on the other, with the typical advantages and disadvantages that these entail. More recently, unsupervised pattern learning has been proposed to increase scalability [17], based on the principle of bootstrapping [6] to iteratively learn patterns for quotation attribution, similar to the seminal Snowball system for relation extraction [1]. However, at its core, even the bootstrapping of patterns still requires pattern matching on text and is thus incapable of coping with the full complexity of natural language, which is evident in its low recall performance [17].

Here, recent advances in natural language processing offer an attractive way forward, with modern NLP techniques such as transformers [23] in general, and BERT [4] in particular, which produce state-of-the-art results on tasks that are closely related to quotation attribution, such as relation extraction and question answering. The downside of these system, of course, is their hunger for training data, which easily exceeds the amount of (often manually annotated) data that is available for many tasks.

To address this problem, we propose to combine the best of two worlds and use pattern-based bootstrapping to generate training data for a supervised neural model, following a paradigm of distant supervision [10]. Not only does such an approach avoid the costly overhead of manually labeling training data, it also allows to use state-of-the-art neural models to maximize recall and create large repositories of attributed quotations, for example from abundant news article data. The resulting framework proposed in this paper, *Quobert,* is sketched in Figure 1. It starts by extracting an initial set of quotation–speaker pairs from a large news corpus via Quootstrap [17], an unsupervised quotation attribution system trained with a single seed pattern. We then sift through the news corpus to identify additional occurrences of the initial pairs in new contexts that were not matched by any Quootstrap pattern. These additional occurrences are vastly more diverse than the pattern-based ones and thus constitute an ideal data set for training a powerful machine learning model with the ability to attribute quotations to speakers across a wide range of contexts. In doing so, we build on BERT [4], a popular pretrained transformer-based language model, which we fine-tune on the task of quotation attribution.

Equipped with this framework, and circling back to Murray's call for a sufficient number of quotations, we put aside the philosophical question of just how much is actually enough and instead aim to extract and attribute as many quotations as possible at Web scale. Applying Quobert to 162 million news articles collected from the online content aggregation service Spinn3r gives rise to Quotebank, a massive corpus of 178 million speaker-attributed quotations. We envision that, in future work, these quotations can be leveraged by the community as essential components to scale up a wide range of quotation-centric applications, such as the construction of claim provenance graphs [25] for fact checking and combating fake news, or the analysis of the propagation and coverage of events in the news [13, 21], among others.

WSDM '21, March 8–12, 2021, Virtual Event, Israel

ACM ISBN 978-1-4503-8297-7/21/03…$15.00
https://doi.org/10.1145/3437963.3441760

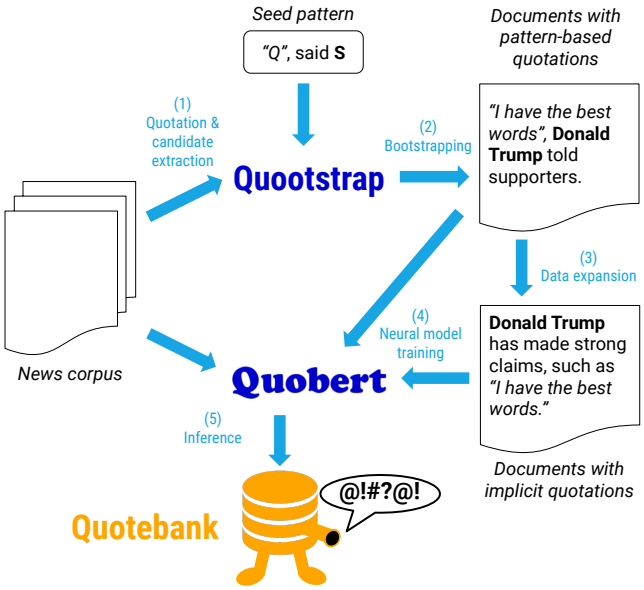


Figure 1: Overview of Quobert (for details, see Section 4). Based on (1) news articles annotated with quotations and candidate speakers, we (2) use Quootstrap [17], an unsupervised quotation attribution system trained with a single seed pattern, to bootstrap an initial set of quotation–speaker pairs. (3) We expand the data by identifying additional occurrences of the initial pairs, which were not recognized by any Quootstrap pattern, and (4) use the expanded set of occurrences to fine-tune a BERT-based model for the task of quotation attribution. (5) The resulting Quobert model is applied to the full news corpus to compile the Quotebank quotation corpus.

We emphasize the elegant use of bootstrapping in our approach: Quobert bootstraps training examples from the output of Quootstrap, which can in turn bootstrap its entire pattern database from a single, manually specified seed pattern—*"Q", said S*—, to the effect that all of Quotebank's 178 million quotation–speaker pairs can ultimately be traced back to that one simple pattern.

To summarize, our main contributions are the following:

(1) We introduce *Quobert*, a distantly and minimally supervised end-to-end framework for extracting and attributing quotations in large news corpora that starts from a single seed pattern and is language-agnostic (Section 4).
(2) We provide a new, manually annotated evaluation data set (Section 5) and use it to demonstrate the performance of Quobert (Section 6), showing that it correctly attributes 86.9% of quotations.
(3) Applying Quobert to 162 million news articles from Spinn3r, we extract *Quotebank*, a massive corpus of 178 million attributed quotations covering over a decade of news (Section 7).

The Quotebank corpus and the Quobert model and code are publicly available at https://doi.org/10.5281/zenodo.4277311.

2 RELATED WORK

Related work on quotation attribution can be divided by *domain* (i.e., the type of corpus for which it is designed), and by the employed *method*. With regard to the domain, previous work has predominantly focused on either literary texts [5, 11] or, more commonly, on news articles [2, 12, 14, 17–19]. The differences between these domains are substantial, and models that are designed for and trained on one domain tend to perform poorly on another [14]. From a technical perspective, existing approaches can be grouped by their method into *unsupervised models*, which are typically pattern-based, and *supervised models*. With regard to the type of extracted and attributed quotations, some previous approaches have considered the problem of attributing *indirect quotations* in an effort to improve the overall recall of quotation extraction [16]. In contrast, we focus on the *direct quotations* that are considered by most existing approaches and aim to improve recall through a better methodology.

Supervised quotation attribution. A typical approach to quotation attribution with supervised learning is to model the process as a token classification or sequence labeling task [12, 14]. Other approaches have proposed to train a model that simultaneously performs quotation attribution and coreference resolution based on a selection of engineered features to attribute even quotations without explicitly mentioned speakers [2]. More recently, Muzny et al. proposed a two-stage sieve approach that first links quotations to speaker mentions and subsequently mentions to entities [11]. Common to all these approaches is the requirement of substantial amounts of labeled training data or engineered, domain-specific features, which limit the usability on diverse corpora in practice.

Unsupervised quotation attribution. As an alternative to supervised methods, unsupervised quotation attribution typically relies on regular expressions or on handwritten rules. In an early contribution, Pouliquen et al. compiled a set of manually created attribution patterns in 11 languages [18]. In a statistically driven analysis of sentences, Salway et al. explored the structure of quotations to leverage them for attribution [19]. More recently, Pavllo et al. introduced Quootstrap as a scalable model for quotation attribution in large news corpora [17], which leverages bootstrapping to learn quotation attribution patterns [6]. While this approach also relies on rules for the extraction and attribution of quotations, it starts with just a single seed pattern and bootstraps additional rules, which significantly reduces the effort of manually creating patterns. As such, Quootstrap can be used to extract a large quantity of quotation–speaker pairs from news corpora with high precision, yet the achieved recall of even this state-of-the-art method is still mediocre (the authors estimate it to be 40%).

In contrast to strictly supervised or unsupervised methods, we consider distant supervision as a combination of the two. The proposed method, Quobert, minimizes the supervision effort and thus avoids the scarcity of training data. We leverage the bootstrapping principle to generate training data with minimal supervision, which we then utilize to train a supervised model to improve the overall performance in general, and recall in particular. We utilize BERT [4] as a pretrained model that is fine-tuned for the task of quotation attribution, which also addresses the prevalent problem that coreference resolution poses in quotation attribution [15].

Table 1: Examples of different types of direct quotations. Candidate speakers are highlighted in bold.

Type	Quotation example
Syntactic pattern	In 2016, **Donald Trump** said, "I know words. I have the best words."
Multi-entity	**Donald Trump** ran against **Clinton**. "I know words. I have the best words", **Trump** claimed during a campaign rally.
Anaphora	**Donald Trump** ran for office in 2016, when **he** said, "I know words. I have the best words."
Implicit	**Donald Trump** has made strong claims, such as "I know words. I have the best words." Does he really?
No speaker	Remember that quote "I know words. I have the best words." from way back when we still had political rallies?
Non-quotation	We recently watched the 2012 movie "The Words" and rather liked it.

3 QUOTATION ANNOTATION AND TYPES

Before we introduce Quobert, we briefly review the different types of quotations that such a model is likely to encounter.

3.1 Direct vs. Indirect Quotations

Quotations can be categorized into *direct* and *indirect* quotations (also called explicit and implicit quotations, respectively). Direct quotations are verbatim reproductions and are typically encased in quotation marks within the context, such as in the example sentence *He said, "I know words. I have the best words."* In contrast, indirect quotations are more difficult to extract since they are only indicated through syntactic construction, such as in the sentence *He said that he knew words and that he had the best words*, which conveys the same information by paraphrasing the speaker. There are, of course, also contexts in which combinations of both types of quotations occur as partially indirect quotations, such as *He said that he knew words, before emphasizing, "I have the best words."*

For Quobert, we focus on the attribution of quotations, and not the extraction of quotations. Due to a lack of highly reliable extraction frameworks for indirect quotations, we use only direct quotations in our experiments, but note that Quobert itself is agnostic to this distinction and could easily be adapted to work with indirect or even partially indirect quotations.

3.2 Types of Direct Quotations

Although we focus on direct quotations, there are still numerous different types of varying complexity to consider, which we briefly discuss in the following. For examples, see Table 1.

Syntactic patterns. The low-hanging fruit for quotation attribution consists of quotations that follow strict syntactic patterns, such as *"Q", said S.* These patterns can be learned or designed easily and integrated in rule-based attribution approaches.

Multi-entity contexts. In many contexts, multiple entities may be mentioned and could be considered as speaker candidates for attribution. Furthermore, even the correct speaker might be mentioned multiple times, indicating that a pooling of the signal that is generated for individual candidates may be useful.

Anaphora. When speakers are not mentioned explicitly, for example due to the use of pronouns, pattern-based attribution approaches require the resolution of such anaphora in preprocessing. The difficulty of coreference resolution in particular has significant impact on the quality of quotation attribution [15].

Implicit attribution. Frequently, quotations are not attributed explicitly by the use of an obvious syntactic pattern, yet the attribution may still be clear from context through patterns of phrasing, punctuation, or through deduction by the reader. Designing patterns for such cases manually seems infeasible, yet they can potentially be learned, given sufficient amounts of training data.

No speaker. In some cases, a quotation may be present without a clearly attributed speaker. More pragmatically, such cases may arise from the failure to even recognize the speaker as a candidate in the entity recognition phase. In both cases, a model should be able to identify the fact that the correct speaker cannot be attributed.

Non-quotations. Finally, not everything that is encased in quotation marks is in fact a quotation. In practice, it may be difficult to draw a clear distinction between quotations and non-quotations, especially in cases of quotations that are partially direct and partially indirect, which tends to result in small sentence fragments that are recognized as quotations.

4 THE QUOBERT FRAMEWORK

Based on the above considerations, it is obvious that machine learning stands to provide massive gains in recall over pattern-based approaches, as long as sufficient amounts of training data can be generated that cover the diverse types of quotations. In the following, we show how such data can be generated with minimal effort by bootstrapping the process with pattern-based attribution. In particular, we leverage the redundancy of quotation–speaker pairs in the corpus. Given the entangled nature of contemporary news streams that are published by a diverse set of news outlets, each of which provides a different yet partially redundant facet of the same event and the same entities on a daily basis [20], identified quotation–speaker pairs will frequently also occur in contexts that pattern-based approaches fail to identify. These contexts provide precisely the training data from which a neural architecture may learn further extraction criteria.

Our proposed framework, Quobert, consist of five steps, detailed in the following subsections.

(1) *Quotation and candidate extraction.* Given a news corpus, we tokenize the text, extract quotations, and identify entities that constitute candidate speakers (Subsection 4.1).
(2) *Bootstrapping.* Using Quootstrap [17], we learn patterns for extracting quotation–speaker pairs and extract pairs from the corpus (Subsection 4.2).
(3) *Data expansion.* We identify additional occurrences of the extracted pairs that were not recognized by any Quootstrap pattern (*positive mentions*), as well as occurrences in which the speaker known from other occurrences is not recognized as a candidate (*negative mentions*) (Subsection 4.3).

(4) *Neural model training*. Using the output of (2) and (3) as training data, we fine-tune a pretrained BERT model for the task of quotation attribution (Subsection 4.4).
(5) *Inference*. We apply the learned model to a massive news corpus to attribute quotations to speakers (Subsection 4.5).

An overview of the pipeline is shown in Figure 1.

4.1 Quotation and Candidate Extraction

Quotation extraction. We use Quootstrap's [17] preprocessing code to identify direct quotations. We consider only quotations whose length lies between $l_{\min}$ and $l_{\max}$ tokens. In the case of nested quotations, we retain only the innermost quotation. On either side of the quotation, we retain a context window of at least w tokens from the surrounding text. In choosing w, we aim to maximize the context window within the technical limitations of the transformer model. (The actual sizes of the left and right context depend on the sentence boundaries: if we can identify the beginning/end of the sentence containing the quotation within $w + 50$ tokens we return this larger context; otherwise we return a context window of size w to each side.)

Candidate speaker extraction and linking. Quootstrap [17] links entities to the Freebase knowledge base, which has since been discontinued and integrated into Wikidata [22]. Since we require an up-to-date repository of candidate speaker names to attribute the quotations contained in contemporary news articles, we therefore use Wikidata instead of Freebase. Although any named-entity detector could be used, we take a simple gazetteer-based approach, where we simply match all names and aliases of people listed by Wikidata as alive at the onset of the news corpus, and link a mention to the set of all candidate entities without disambiguation.

4.2 Bootstrapping

To generate the base set of quotation–speaker pairs, we train Quootstrap [17] on the news article section of the ICWSM 2011 Spinn3r corpus [3], which contains about 14 million news articles in a variety of languages. After deduplication and focusing on English text, we run Quootstrap on 3.8 million articles. Starting with a single seed pattern—*"Q", said S*—and running 4 iterations results in 1,405 quotation attribution patterns. We then apply this pattern base to our large news corpus (see Section 5.1) to identify pairs of quotations and speakers. Quotations that Quootstrap fails to attribute to a unique speaker are discarded. Similarly, we discard quotation–speaker pairs if the speaker that is identified by Quootstrap does not occur in the article from which the context was extracted since it is likely to be falsely attributed.

4.3 Data Expansion

By design, Quootstrap identifies contexts in which patterns can match both a quotation and an attributed speaker. We refer to this case as a *positive mention*. To also include the more complex implicit positive mentions, we identify contexts in which known quotation–speaker pairs occur that were *not* recognized by the pattern-based approach since no matching pattern was available (see Figure 1 for an example).

In contrast to positive mentions, there are also contexts in which only a quotation occurs, but the speaker (known from a different context) does not. We refer to these contexts as *negative mentions*. While Quootstrap is not designed to identify such negative mentions, they are key to training a model that can decide when no speaker should be attributed to a quotation. In practice, such instances may occur due to two reasons: (*i*) the correct speaker is not mentioned in the context window of a quotation, or (*ii*) named-entity recognition failed to identify the speaker as a candidate. We generate negative training examples from both cases.

Additionally, given contexts in which both a quotation and its attributed speaker are correctly recognized, we generate artificial negative examples by deleting the entity annotation of the speaker to simulate a failure in the named-entity recognition phase. Since some degree of failure in named-entity recognition is to be expected in real applications, this type of negative example serves to train the model to avoid erroneous predictions.

4.4 Neural Model Training

In preparation for training a BERT-based model [4], we format the input accordingly. In addition to adding the `[CLS]` and `[SEP]` tokens to the beginning and end of contexts, we also perform masking.

Quotation masking. In order to force the model to focus on the structure and prevent it from learning the content of quotations, all quotations are replaced by special tokens in all contexts. The target quotation, for which we are aiming to detect the correct speaker, is replaced by a `[TARGET_QUOTE]` token. All other quotations in the context are replaced by `[QUOTE]` tokens. As a side effect, this replacement of sequences of tokens by a single token also enables us to fit larger contexts within the 512-token limit of BERT.

Entity masking. Similar to quotations, we mask annotated entities (i.e., speaker candidates) by replacing them with `[MASK]` tokens, in order to make the model agnostic to the identity of specific speakers, thus preventing model bias towards the most commonly quoted speakers and making the model transferable across corpora.

Neural architecture. For a given input context with a target quotation, the intended output of Quobert is a probability distribution over the candidate speakers that are contained in the context. That is, for each of the n entity masks m_i in the context, we require a probability p_i that this entity is the correct speaker. Additionally, we use the context's `[CLS]` token m_0 with associated probability p_0 to represent the case in which none of the candidates is the correct speaker. As we require a probability distribution over the entity masks m_i, we must enforce the stochastic constraint $\sum_{i=0}^{n} p_i = 1$ for each context.

By passing all context tokens through BERT's transformer layers, we obtain a d-dimensional contextual embedding vector for each token. As we require a probability distribution over the candidate speakers, we add, on top of the BERT encoder, a softmax layer that considers only the entity masks m_i and ignores all other tokens. All model weights are fine-tuned in order to maximize the log likelihood of the entity mask corresponding to the ground-truth speaker using the cross-entropy loss. For our implementation, we use the BERT base model from Hugging Face's `transformers` library [24], where $d = 768$.

Table 2: Available data in the Spinn3r corpus (in millions), including the number of quotations that can be attributed by Quootstrap and by the expansion heuristic (see Section 4.3).

	Phases A–D	Phase E	Full
Articles	66.5	96.3	162.8
Contexts	254.2	422.6	676.8
Unique quotations	173.1	221.8	394.9
Attr. quotations (Quootstrap)	29.4	47.2	76.6
Attr. quotations (expansion)	4.6	12.2	16.8
Negative mentions	0.3	0.6	0.9

In order to maintain a one-to-one mapping between entity masks and candidate speakers, we restrict the training data to contexts in which each speaker occurs only once. This restriction is lifted at inference time, as described next. (On a validation set, the restriction was found to not decrease performance on multi-speaker contexts.)

4.5 Inference

As described above, Quobert outputs a probability distribution over entity masks. In order to handle contexts in which the same candidate is mentioned multiple times (and thus represented by multiple entity masks), we consider two combination models: (*i*) In the *max* model, we represent each candidate speaker via their highest-scoring entity mask. (*ii*) In the *sum* model, we first sum the probabilities of all entity masks representing the same candidate and then associate each candidate with their aggregated probability. The sum model has the immediate advantage of returning a probability distribution over candidates, whereas the candidate probabilities may sum to less than 1 in the max model.

5 NEWS DATA AND GROUND TRUTH

We now introduce the news data used for training and evaluating Quobert and for compiling Quotebank.

5.1 Spinn3r News Data

We train Quobert and construct Quotebank on a large corpus of English-language news articles from Spinn3r.com, spanning from August 2008 to April 2020. We preprocess the data by removing HTML tags and tokenizing sentences with the Stanford NLP Penn Treebank tokenizer [9]. The data was collected over the course of 12 years and stored in multiple phases of character encoding with varying quality. Most notably, only the data in Phase E is correctly cased, whereas data in Phases A–D is mostly lower-cased. Line breaks were removed in all phases and are not reconstructible. Table 2 shows an overview of the data. Since Phase E is the cleanest part of the data, we exclusively use data from Phase E for training and testing Quobert in our evaluation. When compiling Quotebank, we, however, include all phases, fine-tuning `bert-base-uncased` for Phases A–D, and `bert-base-cased` for Phase E (see Section 4.4).

5.2 Handling Training Data Imbalances

As evident from the statistics in Table 2, we could potentially use $47.2 + 12.2 + 0.6 = 60$ million quotation contexts as training data. However, due to the distribution of the data, this is unlikely to be helpful in training the model, and the selection of training data

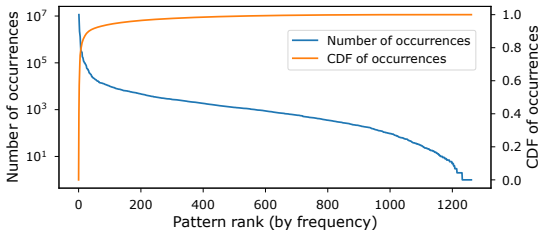


Figure 2: Number of individual occurrences per pattern identified by Quootstrap and cumulative frequency of pattern occurrences. Patterns are sorted from most to least common.

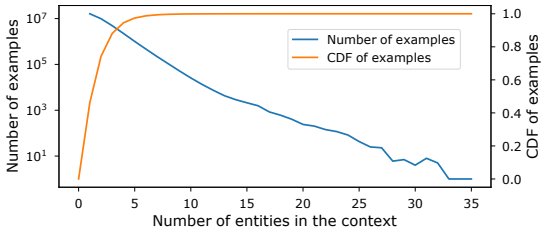


Figure 3: Distribution and cumulative distribution of the number of speaker candidates in quotation contexts.

requires additional consideration. Three types of imbalance occur in the quotation–speaker pairs extracted by Quootstrap and the expansion heuristic, with respect to (*i*) the number of pairs of different types, (*ii*) the frequency of patterns that are successfully matched, and (*iii*) the number of candidate speakers in the context.

Imbalance by quotation type. As shown in Table 2, we have substantially more training data from Quootstrap than from the expansion steps. This would result in a distinct bias towards quotations that can be identified with syntactic patterns, while training data for negative mentions (i.e., the no-speaker class) and implicit attributions would be under-represented.

Imbalance by pattern frequency. The majority of quotation attributions in the Quootstrap output occur due to a small number of very frequently occurring patterns (see Figure 2). If left unchecked, this would create a bias towards the most common and simple patterns and negate any advantage of employing a neural architecture.

Imbalance by candidate entity type. Considering the number of candidate speakers in quotation contexts, we find that most quotations have a simple context with only a few candidates, while quotations with many candidate speakers in the context are rarer (see Figure 3). Using this data for training without accounting for the candidate frequency would lead to a bias towards quotation contexts with few candidates.

To account for these imbalances and learn a classifier that also works on difficult cases, and to keep the training times manageable, we undersample the data to 1 million contexts, as follows.

Sampling the Quootstrap output. We first compute the frequencies of contexts by the pattern that was used to match the context, and by the number of candidate speakers. Due to sparsity, contexts with patterns occurring less than 500 times or with more than 20 candidates are grouped in this step. We then sample quotation contexts by their inverse frequency.

Sampling the expanded data. We sample the expanded data by inverse frequency in the same way as above (but note that we do not have pattern information for this data by design, since pattern-based extraction cannot match these contexts).

Sampling negative mentions. Since the number of contexts with negative mentions is low overall, we include the entire set in the training data. We additionally generate artificial negative mentions (see Section 4.3) in equal proportion from Quootstrap data and the expanded data, such that each source of negative mentions represents a third of the negative class.

Balanced training sample. In order to train a model that is optimized for the data that will be encountered at deployment, we compose a training sample whose proportions reflect the empirical distribution of the ground truth data (see Section 5.3). Negative mentions contribute 25% towards the 1 million training contexts. The remainder is split into 30% data from Quootstrap and 45% data from the expanded data.

Uniformly random training sample. In order to evaluate the effectiveness of the above-described balanced sampling strategy, we also draw another, unbalanced sample of the same size (1 million) from the entire data uniformly at random. By design, this data set reflects all discussed imbalances.

5.3 Ground Truth

While there are strong similarities between our considered task and data and those of Quootstrap [17], using Quootstrap's evaluation data is not feasible. The evaluation of Quootstrap is geared towards measuring the precision of detecting positive mentions for a subset of selected speakers. To take a more fundamental approach that includes negative mentions, we thus require a new ground truth.

We select three types of data for the ground truth to account for multiple evaluation focus areas: implicit contexts, multiple candidate speakers, and a representative setting.

(1) *Implicit* consists of contexts for which the pattern-based approach failed to attribute a speaker.
(2) *Many-choice* consists of contexts that contain a large number of speaker candidates. Given the imbalance towards contexts with few speakers, we expect this to be a more difficult than average subset. We select data from the long tail of the distribution of contexts by sampling by inverse speaker candidate frequency (selected from contexts with 5–20 candidates).
(3) *Representative* mirrors the distribution of contexts in the entire data set and is extracted by drawing random samples of equal size for all months.

For each of the three cases, we sample 500 contexts from the Spinn3r corpus and collect labels via the Amazon Mechanical Turk crowdsourcing platform. We collect labels from 3 crowd workers per context. Workers were United States residents, had an approval rating of at least 99%, and at least 5,000 completed prior assignments. The task included a context with known speaker attribution as an attention check. We re-ran batches with failed attention checks.

For each context, we asked the workers to identify (1) which (if any) candidate speaker uttered the target quotation, (2) if there was no valid target speaker (either because it was not in the context, because it was not in the list of identified candidates, or because it could not be decided from the given context), or (3) if the quotation was invalid because it was not a quotation or not in English.

Based on the answers, we group the crowd annotations into three categories:

(+) *Positive mentions*, if the target speaker was among the candidates that named-entity recognition provided.
(−) *Negative mentions*, if the true speaker was not among the candidates or there was no speaker.
(○) *Other*, if it was not a quotation or the context was ambiguous.

For these labels, we observe a moderate inter-annotator agreement (Fleiss' $\kappa = 0.65$), which speaks to the difficulty of the data set, in particular for the many-choice setting. We retain all contexts for which at least two workers agreed (1,399 out of 1,500 cases, or 93.3%). In Table 3, we show an overview of the distribution of cases.

Table 3: Distribution of positive (+) and negative (−) mentions and other cases (○) for each ground-truth context type.

	+	−	○	∑
Implicit	249	153	51	453
Many-choice	269	64	143	476
Representative	335	103	32	470
Total	853 (61%)	320 (23%)	226 (16%)	1399 (100%)

6 EVALUATION

We briefly introduce the evaluation setup and the evaluated Quobert and baseline models, before discussing their performance.

6.1 Evaluation Setup

We approach the problem of quotation attribution as a multi-class classification problem. For models that assign probabilities to candidates (such as Quobert), we predict the candidate speaker with the highest output probability. As the evaluation metric, we use accuracy, i.e., the fraction of contexts for which the correct candidate speaker was predicted.

Quobert models. We evaluate Quobert for the two samples of training data (*balanced* and *random*) introduced in Section 5.2, and for the two different multi-speaker inference models (*max* and *sum*) introduced in Section 4.5, which results in the following four models:

- $\text{Quobert}_{(\text{BAL,max})}$ is trained on the balanced training sample and uses the maximum individual probability for aggregating multiple mentions of the same candidate.
- $\text{Quobert}_{(\text{BAL,sum})}$ is trained on the balanced training sample and uses the sum of individual probabilities for aggregating multiple mentions of the same candidate.
- $\text{Quobert}_{(\text{RAN,max})}$ is trained on the uniformly random training sample and uses maximum aggregation.
- $\text{Quobert}_{(\text{RAN,sum})}$ is trained on the uniformly random training sample and uses sum aggregation.

We use a batch size of 24 examples per GPU, and Adam [7] as the optimizer, with a learning rate of 10^{-7} and $\epsilon = 10^{-8}$. The learning rate follows a linear schedule with a warmup of 25% of the steps of the first epoch. We set Quobert's parameters (see Section 4.1) to $l_{\text{min}} = 6$, $l_{\text{max}} = 500$, and $w = 50$. All models are trained for three epochs on two Nvidia GeForce GTX Titan (with each epoch taking

Table 4: Speaker-attribution accuracy on subsets of ground truth (see Section 5.3), one subset per row. Best-performing method of each row in bold. Contexts with no actual quotation or with ambiguous speaker attribution (∘) omitted from evaluation.

		Quootstrap	Baseline	CoreNLP	Quobert$_{(\text{BAL,max})}$	Quobert$_{(\text{BAL,sum})}$	Quobert$_{(\text{RAN,max})}$	Quobert$_{(\text{RAN,sum})}$
Implicit	+	0	0.763	0.635	0.835	0.855	0.896	**0.920**
	−	0	0	0.494	**0.857**	**0.857**	0.468	0.422
	All	0	0.471	0.581	0.844	**0.856**	0.732	0.730
Many-choice	+	0.316	0.606	0.606	0.896	0.911	0.944	**0.952**
	−	0	0	0.194	**0.731**	**0.731**	**0.731**	**0.731**
	All	0.253	0.485	0.592	0.863	0.875	0.902	**0.908**
Representative	+	0.331	0.803	0.773	0.872	0.899	0.943	**0.952**
	−	0	0	0.466	**0.825**	0.806	0.485	0.447
	All	0.253	0.614	0.701	0.861	**0.877**	0.836	0.833
Overall		0.167	0.528	0.629	0.856	**0.869**	0.819	0.819

around 4–5 hours to complete) and evaluated on a held-out validation set in the middle and at the end of each epoch. For each model we choose the version that maximizes validation performance.

Nearest-speaker baseline. As a naive baseline, we use a model that works on the assumption of proximity and simply predicts the candidate speaker that is closest to the target quotation (with distance measured as the number of tokens in between).

Quootstrap. We use the pretrained version of Quootstrap [17] as described in Section 4.2 for comparison. We acknowledge that Quootstrap was not designed to work on many of the more difficult cases in our ground truth, and include it primarily to show the gain in recall by training a neural model on its output.

CoreNLP. We also compare to the sieve-based approach by Muzny et al. [11], which is available via Stanford CoreNLP, but we acknowledge that it was designed for literary texts instead of news. We use it with its default settings.

6.2 Evaluation Results

Table 4 shows the results of the evaluation. Quobert's accuracy is clearly superior to that of the other methods, with Quobert$_{(\text{BAL,sum})}$ correctly attributing 86.9% of quotations. The nearest-speaker baseline performs surprisingly well on the representative sample, since the closest candidate is the correct speaker in 80% of the contexts in which a quotation has an attributed speaker, but it obviously fails when no speaker should be predicted.

The accuracy of Quootstrap, which is geared towards high precision and is known to have low recall [17], is expectedly poor. While Quootstrap alone is thus clearly insufficient for constructing a large, complete quotation repository, it still enables the latter by providing high-precision seed quotation–speaker pairs that allow Quobert to catch many of those pairs that Quootstrap missed.

In a comparison between Quobert variants, we find that training on a balanced sample is clearly advantageous for identifying implicit attributions as well as those that are most common in the corpus, while training on a random sample gives the model a slight edge in contexts with many candidates. With regard to the method for combining the signals of multiple mentions of the same speaker, the summation of probabilities clearly has the best overall performance.

Table 5: Attribution error sources for Quobert$_{(\text{BAL,sum})}$.

Error source	Number	Percentage
Failure in obvious pattern context	38	24.5%
Failure in obvious implicit context	21	13.5%
Assumption of implicit context	16	10.3%
Failure in coreference resolution	30	19.4%
Obvious NER target is missing	18	11.6%
Not a quotation / junk context	11	7.1%
Correct label / ground truth is wrong	20	12.0%

The performance of Quobert on the implicit subset of the ground truth in particular highlights the benefits of the underlying BERT architecture for quotation attribution in a large and diverse corpus.

6.3 Error Analysis

To obtain a better understanding of the inner workings of Quobert and investigate sources of errors, we manually evaluate all 155 contexts for which the best-performing model, Quobert$_{(\text{BAL,sum})}$, misattributed the quotation (see Table 5). In 48% of the cases, the error stems from treating an implicit context as though it followed a syntactic pattern, or vice versa. The cases are evenly split between errors due to failing to recognize (often extremely simple) pattern-based contexts and errors due to missing or assuming an implicit context. 31% of the errors are due to a failure of the model to resolve coreferences (although the majority would also have required extended world knowledge to be resolved) or by named-entity recognition (NER) errors for obvious attribution targets. Finally, 20% of the cases are not true errors and stem from junk contexts, misidentified quotations, or are actually correct attributions of quotations for which the crowdworkers agreed on the wrong label. Overall, we find no clear evidence of systematic error sources.

7 QUOTEBANK

With Quobert, it is not our goal to simply add another application to the rising pile of fine-tuned BERT models. Rather, the goal is to create *Quotebank* as a large, community-usable repository of attributed quotations from news. In the following, we give an overview of this repository and highlight one of many potential applications.

Table 6: Ten most frequent speakers by unique quotations.

Speaker	Quotations	Speaker	Quotations
Barack Obama	1,509,759	George W. Bush	177,464
Donald Trump	782,457	John McCain	161,908
Mitt Romney	283,117	Pope Francis	144,453
Hillary Clinton	232,156	Benjamin Netanyahu	136,619
Narendra Modi	203,742	Joe Biden	128,651

7.1 Data Extraction and Overview

Given the results of Section 6.2, we use the Quobert$_{(\mathrm{BAL,sum})}$ variant and extract quotation–speaker pairs from all 162 million articles of the Spinn3r news corpus (Section 5.1). For quotations that occur multiple times, we sum the speaker probabilities across all contexts to attribute a single speaker to each unique quotation.

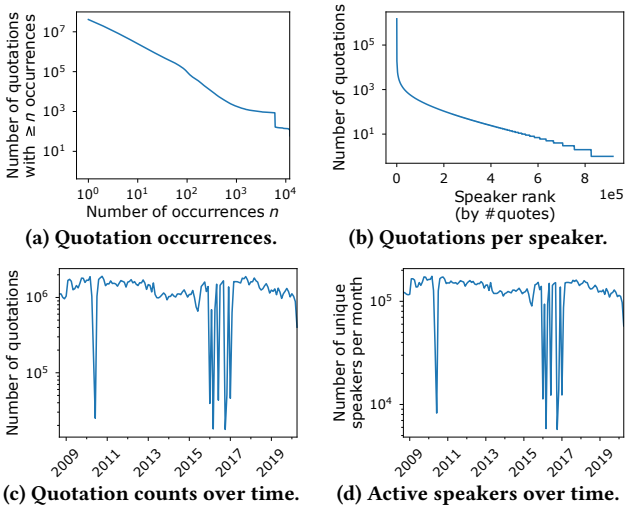
From 235,036,977 quotation contexts with at least one candidate speaker mention, we extracted 178,557,135 unique quotations, each attributed to one of 918,286 speakers in the data who uttered at least one quotation. We extracted quotations from news articles published on 377,065 unique Web domains. In Table 6, we show the most common speakers. In Figure 4, we show the distribution of quotation-occurrence and speaker frequencies, as well as the number of quotations and speakers over time. (The sharp dips are due to data outages on behalf of Spinn3r.)

7.2 Investigating a Shift in Political Attitude

To showcase a potential application for Quotebank, we analyze the changes in political attitude expressed in the words of U.S. Senator Lindsey Graham, who achieved a level of popularity for the complete reversal of his stance towards Donald Trump shortly after the 2016 election [8]. Given this popularity, we expect his change in opinion to be visible in his public statements.

To investigate this event in Quotebank, we extract all 41,700 quotations that were made by Graham between July 2015 (when he first publicly mentioned Trump) to the end of the data in April 2020, four years after the election. After removing near-duplicates with locality sensitive hashing and filtering for quotations that contain mentions of *Trump* or *Donald*, we obtain 3,254 unique quotations, which we transform into 768-dimensional embedding vectors with a pretrained BERT model and mean-aggregate by month. We stack the monthly vectors into a matrix, use principal component analysis to reduce the dimensionality from 768 to 2, and plot the results in Figure 5, which shows that quotations before and after the election are perfectly separable (the single outlier in 2016 is the month of December). In order to confirm that the separation is indeed due to a change in attitude, we also embed contrived positive ("Trump is an example for the Americans", "Trump is the best president we could have hoped for", "Donald Trump has propelled the United States to economic success") and negative statements ("Trump is the worst", "Donald Trump is hurting the country", "Trump's administration is a nightmare"), whose respective means are visualized as "+" and "−", respectively, in Figure 5. This clearly shows that pre-election quotations are close to the contrived negative statements, and post-election quotations close to the contrived positive statements.

While this is but one example of an already known opinion of a single politician towards a single topic, it highlights the potential of



(a) Quotation occurrences. (b) Quotations per speaker.

(c) Quotation counts over time. (d) Active speakers over time.

Figure 4: Quotation and speaker frequencies in the Quotebank repository extracted from the Spinn3r news corpus.

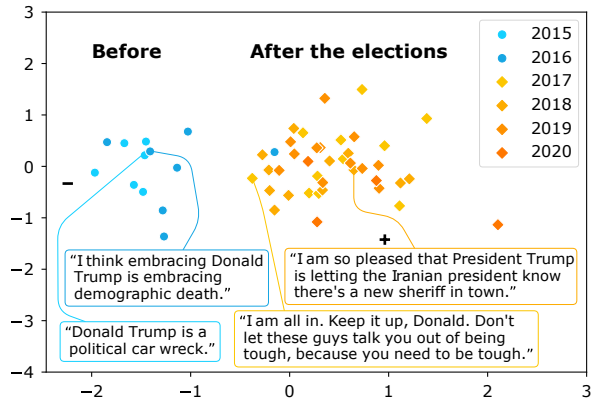


Figure 5: PCA projection of vector embeddings of quotations by Lindsey Graham about Donald Trump. Selected examples highlighted. "+", "−" denote embeddings of contrived positive and negative statements, respectively (see Section 7.2).

Quotebank to enable data-driven analyses of the public statements of hundreds of thousands of speakers, and much beyond.

8 CONCLUSION AND OUTLOOK

We introduced *Quobert,* a framework for language-agnostic quotation attribution with minimal supervision requirements. Based on this framework, we implemented an end-to-end pipeline for the attribution of quotations to speakers by bootstrapping the generation of training data for a neural transformer model from a single manually created quotation attribution pattern. We then used Quobert on a large corpus of 162 million English news articles published between 2008 and 2020 to create *Quotebank,* an open repository of 178 million unique quotation–speaker pairs in which speakers are linked to the Wikidata knowledge base. We believe this resource can be extremely useful for natural language processing and computational social science, and hope these communities will make ample use of Quotebank.

Acknowledgments. This work was partly supported by Collaborative Research on Science and Society, Swiss Data Science Center, Swiss National Science Foundation (grant 200021_185043), Facebook, Google, Microsoft. We thank Tiziano Piccardi, Lorenzo Tarantino, Dario Pavllo, Seth Vanderwilt for early help, and Spinn3r and Jure Leskovec for facilitating data access.

REFERENCES

[1] Eugene Agichtein and Luis Gravano. 2000. *Snowball*: Extracting Relations From Large Plain-text Collections. In *Proceedings of the Fifth ACM Conference on Digital Libraries*. https://doi.org/10.1145/336597.336644

[2] Mariana S. C. Almeida, Miguel B. Almeida, and André F. T. Martins. 2014. A Joint Model for Quotation Attribution and Coreference Resolution. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL*. https://doi.org/10.3115/v1/e14-1005

[3] Kevin Burton, Niels Kasch, and Ian Soboroff. 2011. The ICWSM 2011 Spinn3r Dataset. In *Proceedings of the Fifth Annual Conference on Weblogs and Social Media, ICWSM*.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*. https://doi.org/10.18653/v1/n19-1423

[5] David K. Elson and Kathleen R. McKeown. 2010. Automatic Attribution of Quoted Speech in Literary Narrative. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI*. http://www.aaai.org/ocs/index.php/AAAI/AAAI10/paper/view/1945

[6] Marti A. Hearst. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. In *14th International Conference on Computational Linguistics, COLING*. https://www.aclweb.org/anthology/C92-2082/

[7] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR*. http://arxiv.org/abs/1412.6980

[8] Mark Leibovich. 2019. How Lindsey Graham Went From Trump Skeptic to Trump Sidekick. *The New York Times* (February 25, 2019). https://www.nytimes.com/2019/02/25/magazine/lindsey-graham-what-happened-trump.html

[9] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL*. 55–60. https://doi.org/10.3115/v1/p14-5010

[10] Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics, ACL, and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. https://www.aclweb.org/anthology/P09-1113/

[11] Grace Muzny, Michael Fang, Angel X. Chang, and Dan Jurafsky. 2017. A Two-stage Sieve Approach for Quote Attribution. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL*. https://doi.org/10.18653/v1/e17-1044

[12] Chris Newell, Tim Cowlishaw, and David Man. 2018. Quote Extraction and Analysis for News. In *KDD Workshop on Data Science, Journalism & Media, DSJM*.

[13] Vlad Niculae, Caroline Suen, Justine Zhang, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. QUOTUS: The Structure of Political Media Coverage as Revealed by Quoting Patterns. In *Proceedings of the 24th International Conference on World Wide Web, WWW*. https://doi.org/10.1145/2736277.2741688

[14] Timothy 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, EMNLP-CoNLL*. https://www.aclweb.org/anthology/D12-1072/

[15] Timothy O'Keefe, Kellie Webster, James R. Curran, and Irena Koprinska. 2013. Examining the Impact of Coreference Resolution on Quote Attribution. In *Proceedings of the Australasian Language Technology Association Workshop, ALTA*. https://www.aclweb.org/anthology/U13-1007/

[16] Silvia Pareti, Timothy O'Keefe, Ioannis Konstas, James R. Curran, and Irena Koprinska. 2013. Automatically Detecting and Attributing Indirect Quotations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP*. https://www.aclweb.org/anthology/D13-1101/

[17] Dario Pavllo, Tiziano Piccardi, and Robert West. 2018. Quootstrap: Scalable Unsupervised Extraction of Quotation-Speaker Pairs from Large News Corpora via Bootstrapping. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM*. https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17827

[18] Bruno Pouliquen, Ralf Steinberger, and Clive Best. 2007. Automatic Detection of Quotations in Multilingual News. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP*.

[19] Andrew Salway, Paul Meurer, Knut Hofland, and Øystein Reigem. 2017. In *Proceedings of the 21st Nordic Conference on Computational Linguistics, NODALIDA*. http://www.ep.liu.se/ecp/article.asp?issue=131&article=041&volume=

[20] Andreas Spitz and Michael Gertz. 2018. Exploring Entity-centric Networks in Entangled News Streams. In *Companion of the The Web Conference, WWW*. https://doi.org/10.1145/3184558.3188726

[21] Caroline Suen, Sandy Huang, Chantat Eksombatchai, Rok Sosic, and Jure Leskovec. 2013. NIFTY: A System for Large Scale Information Flow Tracking and Clustering. In *22nd International World Wide Web Conference, WWW*. https://doi.org/10.1145/2488388.2488496

[22] Thomas Pellissier Tanon, Denny Vrandecic, Sebastian Schaffert, Thomas Steiner, and Lydia Pintscher. 2016. From Freebase to Wikidata: The Great Migration. In *Proceedings of the 25th International Conference on World Wide Web, WWW*. https://doi.org/10.1145/2872427.2874809

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, NeurIPS*. http://papers.nips.cc/paper/7181-attention-is-all-you-need

[24] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *CoRR* abs/1910.03771 (2019). arXiv:1910.03771 http://arxiv.org/abs/1910.03771

[25] Yi Zhang, Zachary G. Ives, and Dan Roth. 2020. "Who said it, and Why?" Provenance for Natural Language Claims. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL*. https://www.aclweb.org/anthology/2020.acl-main.406/