Dataset of Quotation Attribution in German News Articles

Fynn Petersen-Frey, Chris Biemann
House of Computing and Data Science & Language Technology Group
Universität Hamburg
{fynn.petersen-frey, chris.biemann}@uni-hamburg.de

Abstract

Extracting who says what to whom is a crucial part in analyzing human communication in today's abundance of data such as online news articles. Yet, the lack of annotated data for this task in German news articles severely limits the quality and usability of possible systems. To remedy this, we present a new, freely available, creative-commons-licensed dataset for quotation attribution in German news articles based on WIKINEWS. The dataset provides curated, high-quality annotations across 1000 documents (250,000 tokens) in a fine-grained annotation schema enabling various downstream uses for the dataset. The annotations not only specify who said what but also how, in which context, to whom and define the type of quotation. We specify our annotation schema, describe the creation of the dataset and provide a quantitative analysis. Further, we describe suitable evaluation metrics, apply two existing systems for quotation attribution, discuss their results to evaluate the utility of our dataset and outline use cases of our dataset in downstream tasks.

Keywords: dataset, quote, quotation, attribution, German, news, annotation

1. Introduction

Ever-increasing amounts of data including discourses in natural language are produced in today's digital era. When scientists or journalists want to analyze this data such as online news articles, they are facing the issue that it is infeasible to manually work through the enormous amounts of data. Extracting who says what to whom is a crucial part in analyzing human communication, how the discourse changes over time or what quotations are reproduced by which media etc. Although the field of natural language processing has made huge leaps forward with the introduction of transformers, there is no suitable, annotated data to train a transformer-based system to extract who said what to whom for modern German news articles.

In this paper, we present a creative-commons-licensed dataset for quotation attribution in German news articles.[1] The dataset consists of 1000 manually annotated articles from the German WIKINEWS website[2]. In total, these annotated articles contain almost 250,000 tokens. We manually annotated and curated *Quotes* in different forms of speech such as *Direct*, *Indirect*, *Free Indirect*, *Indirect/Free Indirect*, *Reported* together with the corresponding *Frame*, *Speaker*, *Cue* and *Addressee*.

An overview of the span annotation classes can be found in Table 1. This includes short descriptions, number of occurrences in the data and their average length. Table 2 gives an overview of quote types including short descriptions, the number of occurrences and average length. In addition, we provide a number of annotated sentences in Examples 1.1–1.5 to get a quick intuition on the dataset and its annotations. These examples are modeled after cases from the curated dataset. We shortened or changed the content as needed to be presentable in this text while keeping the structure and grammatical phenomenon as it was.

Example 1.1 (*Direct*)

Zitat von Merkel: „Wir schaffen das.“
Cue *Speaker* *Direct*
Frame

Quote from Merkel: „We can do this.“
Cue *Speaker* *Direct*
Frame

Example 1.2 (*Indirect*)

Der Nachrichtenagentur sagte er, dass man eine Lösung finden werde.
Addressee *Cue* *Speaker* *Indirect*
Frame

He told the news agency that a solution would be found.
SpeakerCue *Addressee* *Indirect*
Frame

Example 1.3 (*Reported*)

Die Firma forderte eine schnellere Entscheidung.
Speaker *Reported*

The company demanded a quicker decision.
Speaker *Reported*

Example 1.4 (*Free Indirect*)

Ein Sprecher stellte gestern die neuen Ziele vor. Es soll mehr Geld in die Bildung fließen.
Speaker *Free Indirect*

A spokesman presented the new goals yesterday. More money is to flow into education.
Speaker *Free Indirect*

[1] Available at https://github.com/uhh-lt/german-news-quotation-attribution-2024

[2] URL: https://de.wikinews.org

LREC-COLING 2024, pages 4412–4422
20-25 May, 2024.

annotation	short description	count	avg. len.
Quote	the quotation uttered by the *Speaker*, fine-grained labels in Table 2	4182	16.69
Speaker	entity in the text that utters the quotation	3908	3.53
Cue	words that are part of a *Frame* and signal a *Quote* construction	2929	1.57
Frame	part of a sentence including *Cue* & *Speaker*, but not the quotation	3038	8.95
Addressee	entity in the text that the quotation is directed at	337	2.72

Table 1: Overview of quotation attribution spans

type	short description	count	avg. len.
Direct	actual words of an utterance, usually in quotation marks	873	17.54
Indirect	content-wise equivalent utterance using different words, usually part of a sentence together with a *Frame*	2250	14.71
Reported	report of a speech action, possibly far from the original quote, usually a full sentence, no *Frame*	454	18.01
Free Indirect	mix of article author & actual speaker, typically construct with "sollen" (shall) or "müssen" (must), full sentence	171	20.42
Indirect/ Free Indirect	content-wise equivalent utterance written in conjunctive mood, full sentence	434	22.33

Table 2: Overview of the quotation types

Example 1.5 (*Indirect/Free Indirect*)

Ein Passant *schilderte die Situation.* *Die Polizei habe den Bereich großräumig abgeriegelt.*
Speaker *Indirect/Free Indirect*

A passerby described the situation. The police had cordoned off the area over a wide area.
Speaker *Indirect/Free Indirect*

In the following, we review related work on quotation detection and attribution before describing our annotation schema. Then, we describe the creation of our dataset and perform experiments including a quantitative analysis as well as an application of two existing systems for quotation attribution. Before concluding, we describe use cases for our dataset.

2. Related Work

The task of quotation detection and attribution to a speaker has been tackled by numerous approaches, usually with the goal to extract information from the data such as news articles (Krestel et al., 2008; Pareti et al., 2013; Almeida et al., 2014; Scheible et al., 2016). Earlier approaches were purely rule-based and only dealt with quotation detection. More recent works used data-driven methods (often based on the PARC dataset by Pareti (2012)) to detect quotations and their respective speakers. However, there is still a strong focus on resources for direct quotations such as the software by Pouliquen et al. (2007) or the datasets by O'Keefe et al. (2012) and Zhang and Liu (2022) while only few resources provide indirect quotations as they are not as easily extracted automatically.

A similar task is speaker attribution in literary works. As the literary domain differs from the news domain e.g. in the author perspective, in the type of quotations and the focus on characters as implicit or explicit speakers, the field of computational literary studies has seen numerous studies dealing with speaker attribution in literary works (Elson et al., 2010; He et al., 2013; Muzny et al., 2017).

While many works have addressed quotation detection and attribution in English, less work and resources have been created for other languages.

For historical German texts, Brunner (2015); Krug et al. (2018), Brunner et al. (2019) and Brunner et al. (2020) have created a number of resources. The DROC corpus (Krug et al., 2018) consists of 90 fragments of German novels and includes about 2000 annotated direct quotes and annotations for speakers and addresses. The Redewiedergabe corpus (Brunner et al., 2020) extends this work by creating a historical corpus (mostly literary domain, but also some news articles) with fine-grained annotations for speech, thought and writing. Bögel and Gertz (2015) created a system to extract statements from German news articles. For Finnish, Janicki et al. (2023) recently created an annotated dataset of news articles with quotations and their speakers.

Our research is focused on who says what to whom according to German news media. Since no suitable dataset exists for this purpose, we created a new manually annotated and curated dataset for quotation attribution in German news articles.

3. Annotation Schema

The annotation schema is inspired primarily by the Redewiedergabe project (Brunner et al., 2020) and also by the work of Bögel and Gertz (2015). We annotate five different (possibly discontinuous) spans: Beside the actual quotation annotated as *Quote*, spans of *Speaker*, *Cue*, *Frame* and *Addressee* (not part of the Redewiedergabe project) are annotated as optional roles for each *Quote*. For a *Quote* two additional dimensions are coded: Five different types of speech (*Direct*, *Indirect*, *Reported*, *Free Indirect*, *Indirect/Free Indirect*) as well as six media (*Speech*, *Thought*, *Writing*, and not part of the Redewiedergabe project *Speech/Thought*, *Speech/Writing*, *Writing/Thought*). This produces 30 different combinations of quotations. Examples 1.1–1.5 provide short sentences showing the annotations. While we reused some class names from the Redewiedergabe project and tried to align our classes, we modified definitions from the Redewiedergabe annotation guidelines or created new definitions suited for news articles and nested quotations. In the next sections, we define the roles, types and media.

3.1. Roles

A quotation in itself is of little value when it is not known who said it or in what context the statement was made. Thus, we provide roles as additional spans that are linked to one or more quotations. The *Speaker* identifies who said something, the *Cue* describes how it was uttered (possibly negating/adversary!) and the *Frame* provides context so that a quotation is not free-floating.

Speaker The speaker is the linguistic phrase in the text that utters one or more *Quotes*. This is typically a personal pronoun, named entity or a noun phrase subject in a sentence. Explanatory relative clauses as well as content clauses (clauses with *dass* (that) or *ob* (whether) making an attribution to a subject) are not annotated. However, noun phrase modifiers (explanatory attributions of subjects) are annotated (e.g. *der zufälligerweise anwesende Doktor* (the doctor who happens to be present) is the full *Speaker*. Usually, the speaker of *Direct* and *Indirect* speech is located within the associated frame. For *Reported*, the speaker can be found within the quotation span. The *Speaker* of *Free Indirect* and *Indirect/Free Indirect* is outside the respective quotation span.

Cue A *Cue* consists of signal words in a frame that announce a *Direct* or *Indirect* speech. These are usually verbs. However, they can also be specified expressions (e.g.: *laut, nach, so, zufolge* (according to). The *Cue* span can also be split within a frame (typically for German verb pre- or suffixes). Besides the reflexive pronoun, a verb can also include other parts of speech such as prepositions, adverbs, nouns and adjectives which distinguish the verb from similar verbs; e.g.: *von . . . die Rede sein* (talk of . . .), *für wahrscheinlich halten* (consider likely).

Addressee The *Addressee* is the linguistic phrase in the text that a quotation is directed at. It is typically found within a *Frame* or within the quotation span in case of *Reported*.

Frame The part of a sentence that is outside the quotation and contains *Cue*, *Speaker* and possibly *Addressee*. The frame provides context for the quotation. It can be at the beginning, in the middle or at the end of a sentence. It is also possible to split the *Frame* within a sentence if it is interrupted by the quotation. The *Frame* in *Indirect* and *Direct* speech is usually a clause, annotated with its comma or colon, which separates the quotation from the *Frame*.

3.2. Quotation Types

Quotations come in various forms and shapes. These differ in their level of truthfulness of the reproduction to the original utterance. To account for this, we marked a span of text not only as a *Quote* but also labeled it according to the five classes described in the following paragraphs. This additional information per quotation allows to use the dataset for downstream tasks that are only interested in a specific type of quotation. When further processing extracted quotations, systems can consider the reproduction truthfulness in their methods and differentiate e.g. between a *Direct* quotation and only vaguely related *Reported* speech.

Direct The *Direct* label is used for verbatim reproductions of quotations that usually occur enclosed in quotation marks. Typically, it either directly follows an introductory *Frame* or the associated *Frame* immediately follows the quotation. *Frame* and *Direct* speech can be separated either by colons or commas that belong to the *Frame*. In other cases, fragments of *Direct* quotations are integrated into a sentence and cannot stand on their own. Then, the *Direct* speech is nested in a longer quotation of any of the remaining four quotation types.

Indirect A quotation is labeled as *Indirect* whenever the author of a text indicates in an associated *Frame* that the utterances of another person are reproduced as a paraphrase, not verbatim. It is the only type of *Quote* that always requires a *Frame*.

The usual type of *Indirect* speech is a partial sentence that, together with a *Cue* in the *Frame*, forms the complete sentence. In a special case, the *Cue* is a single word reference (e.g. *wonach, demnach, danach* whereupon, thus) to a *Speaker* in the previous sentence.

Reported A *Reported* quotation is a summary report of a statement made by another person that is reproduced in a free manner possibly far from the original statement. Because a reporting style is common in news texts, *Reported* speech is annotated only in cases where, first, there is a clearly identifiable *Speaker* within the quotation span, and second, the quotation contains information uttered by the *Speaker* – not only a description of an action.

Free Indirect A quotation is labeled *Free Indirect* when statements, writings or thoughts of a person are reproduced who is not the article author, but the quotation is nevertheless written from the author's perspective. Mostly formulations with *sollen* (shall) and *müssen* (must) that reflect foreign thoughts, statements or writings are considered. A *Free Indirect* quotation is usually a complete sentence that is not enclosed by quotation marks, nor does it have a *Frame* in the same sentence. However, it has a *Speaker* that is outside the quotation span and thus also outside the sentence.

Indirect/Free Indirect This class is used for those forms of speech reproduction which, without an introductory *Frame*, reproduce statements of another entity in a sentence in the subjunctive mood, but not verbatim. An *Indirect/Free Indirect* quotation occurs only when any of the other four types of quotation occur in the preceding or succeeding sentence that also provide a *Speaker*.

3.3. *Quote* Media

The media of speech reproduction indicate whether a quotation is a *Speech*, *Thought* or *Writing* action. These three media are only annotated if it can be clearly determined which particular medium was used in a quotation. If the media cannot be clearly determined or two different media apply at the same time, the mixed forms are chosen.

Speech A *Quote* is labeled *Speaker* when the reproduced utterance was originally oral. This medium of speech reproduction is often recognizable in newspaper texts by *Cue* verbs that are unambiguous for a spoken reproduction; e.g.: *sagen* (say), *sprechen* (speak). But also the *Speaker* can give clarity about a spoken utterance; e.g. *der Sprecher* (the speaker).

Thought The *Thought* class marks a reproduced cognitive process where the statement originally occurred mentally. Consequently, *Thought* is rarely found in news articles as the author cannot reproduce the thoughts of another person.

Writing This class labels a reproduced writing process or a written form of language. Similar to *Speech*, the *Cue* or *Speaker* can be indicators.

Speech/Thought This mixed medium is used to label a person's oral statement in which they have expressed their thoughts.

Speech/Writing This mixed medium marks a *Quote* a) when it's uncertain whether the original quote comes from a written or oral source or b) if a quotation is made as a combination of texts as well as oral statements.

Writing/Thought This mixed medium is chosen when a person's writing is cited in which he or she has reproduced his or her thoughts. Our dataset of news articles has no instances of this class.

4. Dataset Creation

After describing the annotation schema, we provide details on the source data, its pre-processing, the annotation process, the inter-annotator agreement and handling of disagreement between annotators in the following sections.

4.1. Source Data

The data originates from news articles published on the German WIKINEWS website. We used the XML dump[3] available through the Wikimedia foundation. Our dataset is based on the dump from April 2022 that consists of 13,001 published articles. From these published articles, we randomly sampled 1000 articles for annotation to stay close to the original distribution while reducing the data size to an amount manageable in our project timeframe. These articles range from December 2004 to March 2022.

4.2. Data Pre-Processing

As articles stored in MediaWiki markup contain custom macros for the German WIKINEWS, we wrote a program to obtain plain text. The conversion is a recursive procedure to support the nested macros present in the markup. Using this approach, we stripped all markup like formatting (e.g. bold,

[3]URL: https://dumps.wikimedia.org/dewikinews/

italic), semantic information (e.g. links to entities on Wikipedia) and non-textual content (e.g. pictures, tables) from the documents. Further, we removed any text not belonging to the main text body such as publication metadata, comments, links to related articles or sources. The resulting plain text was tokenized and split into sentences using spaCy (Honnibal et al., 2020).

4.3. Annotation Process

The annotation was carried out by three annotators with a background in German studies or Linguistics. The annotators were selected after performing a trial annotation on a handful of articles. The annotation team received extensive training during a preliminary annotation before the actual annotation begun. Further, we held weekly meetings during the main annotation to discuss open questions and uncertain cases, thereby providing ongoing training to all annotators. The annotation quality of our annotators did not differ in a noticeable way after training. Neither in the discussions nor in the curation did it became evident that the annotations of one annotator were preferred over annotations of another annotator.

In an initial preliminary annotation, we tested the suitability of the annotation schema in the news domain. We iteratively tested which attributes of the schema are necessary and which additional options we needed. Finally, we settled on the medium and type attribute for a *Quote* and *Frame*, *Cue*, *Speaker* and *Addressee* as the other annotation components (roles).

For the annotation, we used the annotation software INCEpTION (Klie et al., 2018). The different components are modeled as span annotations with relations between them to indicate e.g. which *Speaker* belongs to which *Quote*. We divided our sampled documents into six parts to a) annotate and curate in parallel, b) allow to adapt the annotation schema early in the process if needed and c) track the inter-annotator agreement over time. We decided against automatic highlighting of candidate annotations etc. to not introduce any automatic processing bias. Thus, the instances were always manually identified by searching for the suitable grammatical structures.

4.4. Inter-Annotator Agreement

We use Krippendorff's Alpha to compute the agreement between two annotators per part. The measure includes both the quality of the span annotation offsets (overlap) and their labels, but does not include the relations between the span annotations. However, the relations were typically made identical given the same annotation spans and their labels. Moreover, for different annotation spans, there is no sensible way to compute an inter-annotator agreement on the relations.

part	type	medium	roles
Part 1	0.56	0.37	0.61
Part 2	0.76	0.51	0.75
Part 3	0.77	0.40	0.76
Part 4	0.77	0.68	0.76
Part 5	0.86	0.51	0.83
Part 6	0.78	0.61	0.78

Table 3: Krippendorff's Alpha agreement between the annotators on the six parts

Table 3 shows the inter-annotator agreement values for the six parts into which we divided the 1000 documents. The inter-annotator agreement values increased strongly after the first part, slightly increasing with additional experience and training over the course of the remaining parts. As such, the first part required significant curation effort and discussion that ultimately led to improved skills of our annotators. The inter-annotator agreement values for the medium fluctuate and show the lowest numbers in general because annotating the correct medium proved to be difficult depending on the context. The documents in part 3 (with the drop to 0.4) had many quotations where the medium was challenging for the annotators to select.

Neither the Redewiedergabe project nor Bögel and Gertz (2015) report inter-annotator agreement scores to compare to. As our annotation is a lot more complex than most span-based annotations (e.g. named-entity recognition) it is to be expected that our scores a lower. With levels around 0.76 for type and roles, the scores are only slightly lower than the typical scores achieved in simpler span annotations tasks.

4.5. Disagreements between Annotators

During the annotation phase we held weekly meetings to discuss general questions how would we best annotate a specific phenomenon within our annotation schema. After two annotators had finished annotating the documents, we employed curation by a third person to resolve differences in the annotations. In situations where the curator was not certain who (or if any) of the two annotators had correctly annotated the sentences in question, we discussed the issue in detail to resolve the disagreement, thereby potentially defining our annotation guidelines more precisely.

One of the most frequent reasons of disagreement during the early phases of the annotation was the difficulty of choosing the correct medium, usually the choice was between *Writing* or *Speech*. After many discussions, we concluded that it is

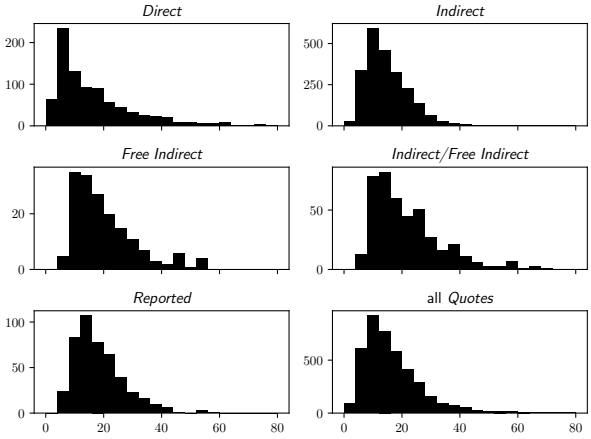

Figure 1: *Quote* token length histograms

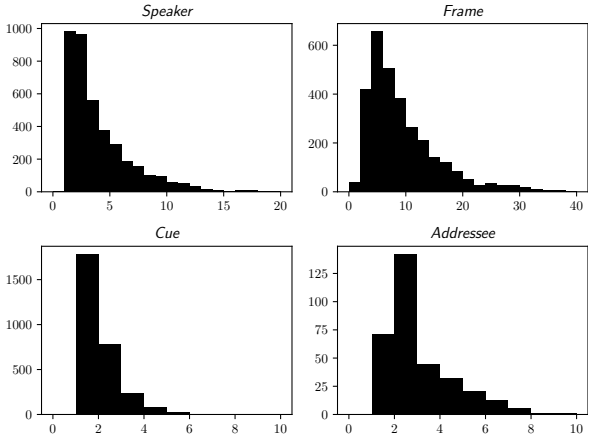

Figure 2: Role token length histograms

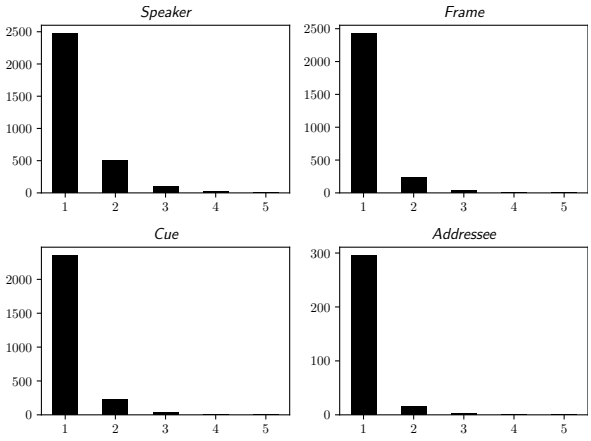

Figure 3: Number of *Quotes* per role span

sometimes impossible to decide from the text alone whether an utterance was produced in spoken or written form. As such, we modified our annotation schema by adding three new labels to medium. While this increased the annotation consistency considerably, it did not completely resolve the issue as the inter-annotator agreement shows.

4.6. Final Dataset

We exported and converted the curated articles into a JSON representation. During the conversion, we applied automatic checks for potential annotation errors and manually resolved true errors in the curated documents. The relations between the annotation spans allow us to build tuples where each tuple consists of one quotation with type and medium as well as all linked roles. Each text span is provided with character, token and sentence offsets enabling easy usage in various NLP frameworks.

5. Experiments

In this section, we present the experiments we performed on the dataset. First, we conduct a quantitative analysis of the annotations. Second, we evaluate two systems on the dataset after explaining the systems and defining evaluation metrics.

5.1. Dataset Analysis

In this section, we provide a quantitative view of the annotations in our dataset. The total count and average length of each *Quote* type is shown in Table 2. Table 1 provides the equivalent data for the role annotations. While most *Quotes* have a *Speaker*, only 70% have a *Cue* or *Frame*.

Figure 1 shows histograms of the token lengths for the different types of quotations. Overall, quotations lengths approach a heavily skewed normal distribution with some very short spans and a long tail for rare, long spans. Most quotations consist of 5 to 20 tokens. *Direct* quotations contain both the shortest and the longest quotations in the dataset. *Indirect* quotations are the shortest on average as they are usually fragments in a single sentence. *Reported*, *Free Indirect* and *Indirect/Free Indirect* (in increasing order of average length) have similar distributions leaning towards longer spans since they normally consist of at least one but sometimes a few sentences.

Figure 2 shows the equivalent histograms for the roles. *Speaker* follows a Poisson distribution where most speaker spans are shorter than 5 tokens. Yet, some *Speaker* spans include descriptive phrases leading to more than 15 tokens (see Section 3.1 for details). The *Frame* annotations follow a skewed normal distribution like the *Quotes*. Since they can be a full sentence in length, they are the longest of the role spans with up to 40 tokens. *Cue* spans are the shortest annotations; typically a single token. However, around 33% of the *Cues* are multi-token expressions. The *Addressee* is a very rare annotation with lengths between one and eight tokens.

Figure 3 shows the number of quotations each role annotation is attributed to. In the overwhelm-

ing majority of cases each role span is only used for a single quotation. However, some spans are attributed to two *Quote* spans. This is especially true for the *Speaker* where up to five quotations are attributed to a single *Speaker* annotation.

We further analyzed the amount of nested *Quotes*, the number of sentences in a *Quote* and the distance between a role and its corresponding *Quote*. 10% of all *Quotes* are nested inside another *Quote* or *Frame*. The majority of these cases are instances of *Direct* speech fragments. While most *Quotes* span only one sentence, about 10% span two or more sentences (up to 11 sentences). 11% of all role spans and 21% of all *Speaker* spans are one or more sentences apart from the corresponding *Quote* (up to 7 sentences).

From these quantitative observations it becomes apparent that a system requires the following traits to be able to perform well: 1) Find *Quotes* and match the *Speaker* without any *Cue*. 2) Support for multi-word *Cue* spans. 3) Support *Quote* spans over multiple sentences. 4) Support finding roles for a *Quote* in other sentences. 5) Find *Indirect* and *Direct* quotations, also *Reported* and *Indirect/Free Indirect* as they account for 10% each, while *Free Indirect* makes up only 4% of all *Quotes*. 6) Support role spans to be used for multiple *Quotes*. 7) Support nested *Quote* spans.

5.2. Baseline Systems

In order to evaluate the utility of our dataset, we apply two baseline systems that can extract quotations with attributed roles from news articles. The first system is a purely rule-based system that does not need any training data. The second system uses a data-driven machine learning approach.

An apparently similar problem to the annotations in our dataset is semantic-role-labeling (SRL). However, we did not evaluate SRL systems because the typical SRL datasets and thus available systems are limited to work on a single sentence as a unit. This is not suitable for our dataset that requires a document-wide context (or at least multi-sentence context) as quotes span multiple sentences and roles appear in yet other sentences.

Rule-based system We developed a rule-based system (RBS) building on top of spaCy (Honnibal et al., 2020) to extract direct and indirect quotations with the speaker from text. The system follows ideas of an older system presented by Bögel and Gertz (2015). It uses rules and word lists on top of neural components for dependency parsing and named-entity recognition. *Direct* speech is identified by regular expressions looking for quotation marks. The *Speaker* of the quotation (i.e. the speaker) is searched in the proximity, preferring candidates in the same sentence but outside the quotation span. *Indirect* speech is identified through the grammatical structure of a sentence (using dependency parsing) and the main or auxiliary verb being a cue word that is looked up in a word list. The word list contains utterance verbs (*verba dicendi*) that can be used to indicate (in)direct speech. In addition, the system finds sentences in subjunctive mood that occur directly before or after a sentence containing another quotation. These sentences are typically marked as *Indirect/Free Indirect* in the dataset. Lastly, the system combines *Direct* and *Indirect* speech, enriching the information of identical quotations. The system does not handle the *Addressee* span. Since it is a rare class, we simply ignore it and do to not predict any *Addressee*. However, *Frame* is a frequent role that the system predicts by marking all tokens of a sentence as the *Frame* that do not belong to the *Direct* or *Indirect* quotation.

Citron The system was created by the BBC (Newell et al., 2018) to extract quotations with their *Speaker* from English news articles. It consists of several components that are built on top of spaCy (Honnibal et al., 2020) and are trained individually: *Cue* classifier, *Speaker* classifier, *Speaker* resolver, *Quote* classifier, *Quote* resolver. The resolvers link the classified spans to a *Cue*. The system can only find quotations that have a *Cue* – with the additional constraint that a *Cue* is single token verb. We modified Citron to work with German texts, accept any single token as a *Cue* and trained it using the subset of all quotations that have a *Cue* in our dataset (70%). As with RBS, we ignore any *Addressee* and predict the *Frame* to span all tokens in a sentence not belonging to the *Quote*. To predict the quotation type, we use *Direct* for spans with quotations marks and *Indirect* otherwise.

5.3. Evaluation Metrics

We use the usual precision, recall and F1-metrics on token overlap of possibly discontinuous spans (thereby creating ordered sets of tokens). For most *Quote* types, all roles are optional. Thus, predicted spans of roles can only be matched to the reference roles if they belong to a correctly matched *Quote*. A span representing a role can be related to multiple *Quote* spans, i.e. the same *Speaker* can utter multiple *Quotes*. Roles or *Quote* spans can be nested within another *Quote* or *Frame*. To perform an evaluation, *Quotes* from system and reference are assigned via linear sum assignment of the *Quote* span's token overlap using type and medium as tie-breakers. Each *Quote* can only be matched to at most one other *Quote*. The tie-breakers are needed to correctly assign *Quotes* in rare cases as

they can have the same offsets, yet are of a different type or medium. If a system predicts a *Quote* that has no matching *Quote* in the reference annotations, this increases the false positives for *Quote* and each role the system predicted as belonging to the unmatched *Quote*. Vice versa, if a *Quote* from the reference annotation has no match in the system prediction, the false negatives are increased. A correctly matched *Quote* yields true positives for all correct roles according to the fraction of overlap and false negatives resp. false positives for tokens that were not identified resp. wrongly predicted by the system.

5.4. Results

To evaluate the two baseline systems, we divided our dataset into three parts: A training set of 700 documents, 150 documents for the development set (653 quotations, 1567 roles) and 148 documents in the test set (652 quotations, 1605 roles). Table 4 (upper half) shows the results for the two baseline systems on the development/test set. We do not report scores on the medium because neither system is capable to predict it. The rule-based system is not tuned on the development set (and not even trained on the training set). Consequently, there should be almost no difference between the scores on the test and development set.

Overall, the results show that both systems achieve between decent and good precision while clearly suffering from low recall. Compared to Citron, the rule-based system has lower precision, but higher recall of *Quote* resulting in a slightly better F1 score. For the roles, Citron has both better precision and better recall than RBS. Together (joint measure of *Quote* and roles), Citron again surpasses RBS in both precision and recall. As for predicting the type of a *Quote*, RBS has slightly higher recall, but greatly lower precision than Citron leading to slightly better F1 score of Citron.

For the rule-based system, the low recall mainly results from two causes. First, the system is not capable of predicting certain types of speech (*Reported* and *Free Indirect*) or roles (*Addressee*) that are present in the dataset. Second, the system was designed to prefer quality to quantity when automatically extracting quotations from large amounts of raw text. As such, the system has a preference for precision over recall even for types of speech that it can predict.

For Citron, the low recall also has two reasons. First, the system only predicts quotes that have a *Cue* – but only 70% of all *Quotes* have a *Cue*. Second, the *Cue* recall itself is low because Citron's *Cue* classifier a) cannot detect a multi-word *Cue* and b) was designed only for verbs as *Cue*.

While RBS should produce highly similar results on the test resp. dev set, there is a difference in the performance. This deviation in precision, recall and F1 between the test and dev dataset for RBS can be solely attributed to natural variations in the data, e.g. the different quotations contained in the documents of the test resp. dev set. The documents in the dev set contain quotations that happen to be more aligned with the rules implemented in RBS, thus reaching a slightly higher scores.

In summary, the data-driven machine learning approach clearly outperforms the rule-based system. However, both existing systems do not provide a high recall level thereby motivating the need for our presented dataset to enable the creation of new systems providing higher recall (and more fine-grained annotations etc.). From our experimentation with the systems, we believe that a machine learning approach actually designed for the annotation schema will significantly improve the recall to a usable level.

5.5. Ablation Study

To support our view and verify that the low recall of the Citron system is largely an effect of its *Cue* limitations, we performed additional experiments with modified versions of the dataset. First, we remove any quotations that do not have a *Cue*, i.e. that cannot be predicted by Citron. The results are shown in Table 4 (lower half) for the data *dev cue* (445 quotations, 1340 roles) and *test cue* (468 quotations, 1400 roles). Second, we further removed any quotations that have a multi-word *Cue* since Citron internally is limited to a single word *Cue*. We re-trained Citron with the new data. The results are in the same table with the data column *dev 1 cue* (255 quotations, 768 roles) and *test 1 cue* (300 quotations, 899 roles).

For Citron, the effect is as expected: The recall for both quotes and roles significantly increases while precision slightly decreases, leading to increased F1 scores across the board. Limiting the dataset to quotations with any *Cue*, Citron sees +7.5 recall, +7.5 F1 on the dev set resp. +7.6 recall, +7.5 F1 test set joint scores combining quotations and roles. As a comparison, for our rule-based system precision decreases (-9.3/-6.9), recall increases (+2.6/+2.2) and F1 remains unchanged (+0.0/+0.2) for dev/test joint scores. For limitation to a single-word *Cue*, RBS performance degrades as its rule set is no longer compatible with the artificially reduced dataset: Joint precision -21.1/-19.3, recall +3.2/+1.8 and F1 -6.6/-6.0. Citron, however, improves another +8.7 recall, +5.5 F1 for the dev set resp. +7.7 recall, +4.7 F1 for the test set joint scores. Together, this results in an increase of +16.2 recall, +13.0 F1 on the dev set resp. +15.3 recall, +12.2 F1 on the test set joint scores. Thus, we confidently attribute a large part of Citron's low recall to its *Cue* limitations.

		quotation			roles			joint			type		
system	data	prec.	rec.	F1	prec.	rec.	F1	prec.	rec.	F1	prec.	rec.	F1
RBS	dev	75.1	36.1	48.8	55.0	25.5	34.9	60.7	28.7	38.9	57.8	29.6	39.1
RBS	test	70.8	36.2	47.9	55.6	26.1	35.5	59.9	29.0	39.1	63.5	33.6	43.9
Citron	dev	91.5	27.6	42.4	79.3	31.5	45.1	82.4	30.3	44.3	87.0	26.6	40.8
Citron	test	88.2	30.1	44.9	77.9	34.2	47.6	80.5	33.0	46.8	86.5	29.6	44.1
Citron	dev cue	88.4	39.9	55.0	80.1	37.1	50.7	82.2	37.8	51.8	91.5	41.1	56.7
Citron	dev 1 cue	79.0	44.6	57.0	73.5	47.1	57.4	74.9	46.5	57.3	82.1	48.6	61.1
Citron	test cue	85.9	41.6	56.1	80.6	40.3	53.7	81.9	40.6	54.3	90.1	42.9	58.2
Citron	test 1 cue	80.0	47.6	59.7	74.7	48.3	58.8	76.0	48.3	59.0	85.4	50.7	63.6

Table 4: Evaluation results

Most multi-word *Cue* expressions in the dataset are either past tense constructions or common idioms (see Section 3.1) that could be replaced by a single verb. When manually examining the data, there is no inherent difference in the difficulty between quotations and roles used in single- or multi-word *Cue* expressions. Consequently, we are confident a better suited system is capable to achieve strong results on our dataset and thereby create a system that can automatically extract quotations with attributions from German news articles.

6. Use cases

In this section, we outline envisioned use cases with project partners from the Digital Humanities and Computational Social Sciences. We demonstrate how our dataset can help researchers work on their research questions. Note that we do not intend the data (WIKINEWS articles) to be analyzed directly. While this may be interesting for specific research targeting WIKINEWS, we intend our annotated resource to be used to train machine learning systems which, in turn, enable the automatic creation of annotations on other data sources, thereby making it useful for a wide range of applications.

Such a system can produce a list of quotations/speaker pairs from a collection of documents. This allows researchers to quickly analyze the quotations contained in their data without laboriously reading every single document in a potentially large collection. The system can further provide the *Cue* and *Frame* to automatically mark negations, classify the type of quotation, aggregate quotations by their *Cue* word and provide statistics on these aspects. Thereby, researchers can both have a quantitative view on the quotations in their data of interest as well as qualitatively analyze individual quotations and/or speakers by filtering all detected quotations for certain aspects.

For example, social climate science researchers can compare statements after grouping the speakers into politicians, environmental activists, corporate representatives and other public figures (this is possible after performing co-reference resolution and entity linking on the documents).

Another example is the comparison of different news outlets based on the general frequency resp. fraction of text being a quotation as well as the attributes of quotations used: Type of quotation (e.g. *Direct* versus some *Indirect* form), presence versus absence of a *Speaker*. After collecting news articles on the same topic during the same timeframe for various media outlets, researchers can quantitatively compare the news outlets and analyze whether this correlates with news outlet metadata such as reach, geographical location, position in the political spectrum. Further, it is possible to check if individual quotations occur in multiple news outlets or only once – these cases could be candidates for a manual verification or otherwise of interest.

7. Conclusion

We presented a new dataset for quotation attribution in German news articles. The dataset is freely available under a Creative Commons license and provides curated, high-quality annotations. The fine-grained annotation schema allows the data to be used for various applications as it includes not only specifies who said what but also how, in which context, to whom and the type of quotation.

We described our annotation schema and dataset creation in detail, provided inter-annotator agreement and performed a quantitative analysis of the final dataset. Finally, we evaluated two existing systems on our new dataset showing that a new approach is required to provide a high quality automatic detection of quotations. While the systems managed to achieve an acceptable precision, they were only able to detect a subset of all annotations leading to low recall.

In the future, we want to create a system using the full potential of the dataset to be able to automatically obtain attributed quotations from news articles.

8. Ethical Considerations and Limitations

Automating tasks to scale to large data collections always carries a certain risk. In the case of this paper, the dataset is the foundation to create a system that can extract attributed quotations from German news articles with high precision and recall (but certainly not error free). Identifying who said what to whom according to news media on a large scale poses only a small risk compared to generating fake quotations (instead of extracting real ones) with already available state-of-the-art large language models. Moreover, there are already existing rule-based systems (with precision and/or recall issues) to extract quotations and speakers automatically. In any case, our dataset also provides the type of quotation so when using the identified quotations for a further analysis, it is possible to interpret the results more appropriately.

The dataset in itself is based on a freely available resource and uses a random sample without any focus on particular topics, speakers, authors or sources. However, the articles in WIKINEWS might include certain biases as some articles will be written by the same authors, have the same source news agencies etc. Being an open, collaborative platform, the raw articles should still be less biased than relying only on licensed articles from one specific news outlet.

Our annotations are likely neither perfectly error/bias free nor all-encompassing. Sometimes, it is a balancing act to decide whether a certain sentence contains a quotation or the article author only phrased the sentence in a certain way to suggest a quotation. Yet, we employed all means to create high-quality, fine-grained annotations to mitigate such issues by relying on skilled annotators, using annotation guidelines, weekly discussion meetings, curation and thorough handling of disagreements between annotators.

Overall, the possibility to extract who said what to whom according to news media can be an invaluable tool for researchers and journalists in their work to analyze the vast number of online media, help with identification of fake news based on their quotations or ease verification of quotations.

9. Bibliographical References

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*, pages 39–48, Gothenburg, Sweden. Association for Computational Linguistics.

Thomas Bögel and Michael Gertz. 2015. Did i really say that? – combining machine learning and dependency relations to extract statements from german news articles. In *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology*, pages 13–21, Duisburg-Essen, Germany. German Society for Computational Linguistics and Language Technology.

Annelen Brunner. 2015. Automatic recognition of speech, thought, and writing representation in german narrative texts. *Literary and Linguistic Computing*, 28(4):563 – 575.

Annelen Brunner, Stefan Engelberg, Fotis Jannidis, Ngoc Duyen Tanja Tu, and Lukas Weimer. 2020. Corpus REDEWIEDERGABE. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC), May 11-16, 2020, Palais du Pharo, Marseille, France*, pages 803 – 812, Paris. European Language Resources Association.

Annelen Brunner, Ngoc Duyen Tanja Tu, Lukas Weimer, and Fotis Jannidis. 2019. Deep learning for free indirect representation. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Short Papers*, pages 241–245, Erlangen, Germany. German Society for Computational Linguistics & Language Technology.

David Elson, Nicholas Dames, and Kathleen McKeown. 2010. Extracting social networks from literary fiction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 138–147, Uppsala, Sweden. Association for Computational Linguistics.

Hua He, Denilson Barbosa, and Grzegorz Kondrak. 2013. Identification of speakers in novels. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1312–1320, Sofia, Bulgaria. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, et al. 2020. spaCy: Industrial-strength natural language processing in Python.

Maciej Janicki, Antti Kanner, and Eetu Mäkelä. 2023. Detection and attribution of quotes in Finnish news media: BERT vs. rule-based approach. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*,

pages 52–59, Tórshavn, Faroe Islands. University of Tartu Library.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico. Association for Computational Linguistics.

Ralf Krestel, Sabine Bergler, and René Witte. 2008. Minding the source: Automatic tagging of reported speech in newspaper articles. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Markus Krug, Lukas Weimer, Isabella Reger, Luisa Macharowsky, Stephan Feldhaus, Frank Puppe, and Fotis Jannidis. 2018. Description of a corpus of character references in german novels – DROC [Deutsches ROman Corpus]. In *DARIAH-DE Working Papers*, 27. Göttingen: DARIAH-DE.

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

Chris Newell, Tim Cowlishaw, and David Man. 2018. Quote extraction and analysis for news. In *Proceedings of the Workshop on Data Science, Journalism and Media, KDD*, pages 1–6, London, UK.

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*, pages 790–799, Jeju Island, Korea. Association for Computational Linguistics.

Silvia Pareti. 2012. A database of attribution relations. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3213–3217, Istanbul, Turkey. European Language Resources Association (ELRA).

Silvia Pareti, Tim O'Keefe, Ioannis Konstas, James R. Curran, and Irena Koprinska. 2013. Automatically detecting and attributing indirect quotations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 989–999, Seattle, Washington, USA. Association for Computational Linguistics.

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

Christian Scheible, Roman Klinger, and Sebastian Padó. 2016. Model architectures for quotation detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1736–1745, Berlin, Germany. Association for Computational Linguistics.

Yuanchi Zhang and Yang Liu. 2022. DirectQuote: A dataset for direct quotation extraction and attribution in news articles. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6959–6966, Marseille, France. European Language Resources Association.