Don't Annotate, but Validate: a Data-to-Text Method for Capturing Event Data

Piek Vossen, Filip Ilievski, Marten Postma, Roxane Segers
Vrije Universiteit Amsterdam
De Boelelaan 1105, 1081 HV Amsterdam
The Netherlands
{piek.vossen, f.ilievski, m.c.postma, r.h.segers}@vu.nl

Abstract

In this paper, we present a new method to obtain large volumes of high-quality text corpora with event data for studying identity and reference relations. We report on the current methods to create event reference data by annotating texts and deriving the event data a posteriori. Our method starts from event registries in which event data is defined a priori. From this data, we extract so-called Microworlds of referential data with the Reference Texts that report on these events. This makes it possible to easily establish referential relations with high precision and at a large scale. In a pilot, we successfully obtained data from these resources with extreme ambiguity and variation, while maintaining the identity and reference relations and without having to annotate large quantities of texts word-by-word. The data from this pilot was annotated using an annotation tool created specifically in order to validate our method and to enrich the reference texts with event coreference annotations. This annotation process resulted in the *Gun Violence Corpus*, whose development process and outcome are described in this paper.

Keywords: event coreference, text corpora, structured data

1. Introduction

Events and entities are central to referential semantics. Semantic parsing of news articles not only concerns detecting the meaning of words and their relations, but especially establishing the referential relations to the outside world. For entities, it is straightforward what this referential world is. However, compared to entities, events are less tangible (Guarino, 1999; Hovy et al., 2013) for various reasons: 1. we use a small vocabulary to name events, which results in large referential ambiguity 2. events are more open to interpretation and framing, which leads to more variation in making reference 3. events are less persistent in time than entities 4. each event has many idiosyncratic properties, e.g. unique participants playing different roles in a unique spatio-temporal context, making generalization harder. Due to these properties, textual data on events is more fragmented than textual data on entities. In news, events are mentioned during a very short period of time and they rapidly lose their news value (except for a few events like *9/11*), whereas popular entities tend to be mentioned across different texts over longer periods of time. We thus do not find a typical Zipfian distribution for events, with a few events that dominate the news (the head) and a long tail of low-frequent events, but a more even low-frequent distribution.

Given this fragmented distribution, it is not surprising that NLP tasks on event detection, event relation detection, and event coreference are difficult, which is reflected by relatively low inter-annotator-agreements and small amounts of data with event annotations.

Most corpora with event annotations do not consider how they relate to the same or similar events in the world, e.g. PropBank (Kingsbury and Palmer, 2002), NomBank (Meyers et al., 2004), and FrameNet (Baker et al., 2003). In these corpora, syntactic structures are taken as the starting point and all predicates mark mentions of events. Similarity of events follows from the assigned event type, the meaning of the word or frame, and from having similar argument structures. These corpora, however, lack a notion of event reference and are not very well-suited for studying the different ways we describe the same or similar events. For the latter purpose, specific event coreference corpora have been created: ECB+ (Cybulska and Vossen, 2014), RED (O'Gorman et al., 2016), among others. Event coreference annotations have been created using what we call a **text-to-data (T2D)** approach. In the T2D approach, annotators start from the text and first decide what phrases are labeled as event mentions after which different event mentions are related to each other through an event coreference relation. The coreference relations establish event identity across event mentions a posteriori by chaining event mentions that share coreference relations. Due to the complexity and labor-intensity of this T2D approach, only a limited amount of referential event data has been created so far (see Table 1).

The research on event coreference faces a data bottleneck because it is both too difficult and too costly to gather sufficient data following the traditional T2D method. We therefore present a novel semi-automatic method, called **structured-data-to-text (D2T)** to address this data problem. Instead of deriving event identity a posteriori after the text annotation, this approach starts from event registries (structured data on events) in which the events are defined a priori and the texts that report on these events are mapped to the event data. By preserving the relation between the world and texts reporting on it, D2T allows us to create large volumes of referential event data in a more efficient way, with high agreement, and capturing more variation in the language making reference.

The paper is structured as follows. In Section 2., we analyze the current state for the event coreference data collected so far and summarize the major issues and bottlenecks. In Section 3. we explain our proposal to follow a D2T approach and give an overview of the potential data archives that can be used. Section 4. reports on a pilot study to cre-

ate an event coreference corpus, called the *Gun Violence Corpus (GVC)*, following this method. An annotation tool was specifically designed to pair structured data with reference texts and annotate the event mentions per incident. We analyzed the annotation results and efforts in terms of volume, speed, agreement and variation of referring expressions. Finally, we conclude and discuss our future plans in Section 5.

2. From text to data

In this Section, we first give an overview and the overall statistics of the most studied datasets for event coreference, which are all created using a T2D method. Next, we discuss their annotation process.

2.1. State of T2D datasets

In Table 1, we present an overview of the text corpora that dominated the event coreference research in the last decade. The Table shows the number of documents in each dataset, the number of mentions of events, and the number of so-called coreference clusters (groups of mentions that refer to the same event). The final column indicates if the coreference clusters span across documents (cross-document coreference) or only within a single document (within-document coreference). We observe that the number of documents and mentions is small for both within- and cross-document relations: less than four thousand documents and less than forty thousand mentions in total (10 mentions per document on average). The ratios between mentions and clusters vary considerably, which is due to the different ways in which the datasets have been compiled: either subsets of the sentences and/or event types were annotated or all mentions in a full article.

Cross-document data is more sparse than within-document data, as can be seen in Table 1. ECB+ (Cybulska and Vossen, 2014) therefore extended the Event Coreference Bank (ECB) (Bejan and Harabagiu, 2008; Lee et al., 2012) from 482 articles to 982 articles by including more events of the same type. This slightly increased the referential ambiguity and variation, but, nevertheless, only a few sentences per article were annotated (1.8 sentences per article on average in ECB+). The Rich Entities, Relations and Events corpus (Song et al., 2015) and the Richer Event Description corpus (O'Gorman et al., 2016) are two recent initiatives to manually create similar annotations for *all* sentences in articles and also partially across documents, but the number of documents covered is small.

We analysed the referential annotations in a number of these datasets, revealing that they, despite efforts such as the creation of ECB+, hardly reflect referential ambiguity and show very little variation (Cybulska and Vossen, 2014; Ilievski et al., 2016). For example, ECB with 482 documents contains 8 news articles on one specific murder, but since there are no other murders in the dataset, searching for the word "murder" results in almost all mentions of that specific incident with high accuracy: one-form-one-referent and one-referent-one-form. Cybulska and Vossen (2014) demonstrated that the so-called lemma baseline to establish coreference relations[1] scores already very high in this dataset and is difficult to beat by state-of-the-art systems. From the perspective of a real-world situation and the many different ways in which events can be described and framed in language, these datasets are far too sparse and do not reflect true ambiguity and variation. Partly due to this lack of data and variation, automatic event coreference detection has made little progress over the years, especially across documents (Chen and Ji, 2009; Bejan and Harabagiu, 2010; Lee et al., 2012; Liu et al., 2014; Peng et al., 2016; Lu and Ng, 2016; Vossen and Cybulska, 2016).

All data listed in Table 1 are created according to the T2D approach: a selection of text is made and interpreted by annotators who add an annotation layer. Creating data following a T2D approach is expensive and labor-intense, as all mentions of events need to be cross-checked against all other mentions across documents for coreference relations. With the size of the data, the effort increases exponentially.

2.2. State of text-to-data guidelines

Besides meta-level choices on what needs to be annotated, guidelines and annotations tend to differ in criteria for deciding on the text span to be annotated as a mention. In some cases, complete phrases (e.g. "inflicted a fatal gunshot wound") or even whole sentences are annotated, whereas in other cases only semantic main verbs ("pull the trigger") are annotated; implicit events ("murderer", "killer", "victim") are included or excluded; coreference, subset and subevent relations are lumped together ("attack", "pointing a weapon", "shootings", "5 shots", "4 hits one fatal", "3 injured, 1 killed, 1 died in the hospital"); quantification of events is or is not ignored (the phrase "the 2 earthquakes" refers to two different earthquake events) or generic events ("measles is a deadly disease") are excluded; only realis events are annotated or also irrealis events; aspectual verbs ("begin", "stop", "continue", "happen", "take place") are sometimes seen as events and sometimes not; adjectival or adverbial modifiers ("fatal accident") are not marked, etc. Such choices are based on a priori criteria regardless of the types of events annotated and they tend to vary depending on the specific task for which the data were annotated e.g. semantic role detection (Kingsbury and Palmer, 2002), detecting temporal and causal event relations (Boguraev et al., 2007; Pustejovsky and Verhagen, 2009; Bethard et al., 2015; Caselli and Morante, 2016), or event coreference relations (Hovy et al., 2013).

Besides the differences in guidelines, annotators following guidelines may also have different interpretations, which may lead to relatively low inter-annotator-agreement and conservative annotation strategies. Due to the complexity of the task, annotators may for example stay on the safe side and create identity relations only when the same word is used, hence eliminating variation. Such difficulties in defining events, event relations, and event coreference have led to the creation of the KBP2015 dataset (Mitamura et al., 2015) in which a weaker definition of an event has been applied, so-called Event Nuggets, to ease the annotation and the task for establishing coreference relations. In the KBP2015 dataset, "attack", "shooting", and "murder" do not represent separate event instances, but are considered

[1] all occurrences of the same word, e.g. "murder", mention a single unique event and hence are coreferential

Table 1: Event coreference corpora for English created by a text-to-data method

Name	Reference	nr. docs	nr mentions	mention/ docs.	nr clusters	mention/ cluster	cross doc.
ACE2005	(Peng et al., 2016)	599	5268	8.79	4046	1.30	NO
KBP2015	(Mitamura et al., 2015)	360	13113	36.43	2204	5.95	NO
OntoNotes	(Pradhan et al., 2007)	1187	3148	2.65	2983	1.06	NO
IC	(Hovy et al., 2013)	65	2665	41.00	1300	2.05	NO
EECB	(Lee et al., 2012)	482	2533	5.26	774	3.27	YES
ECB+	(Cybulska and Vossen, 2014)	982	6833	6.96	1958	3.49	YES
MEANTIME	(Minard et al., 2016)	120	2096	17.47	1717	1.22	YES
EER	(Hong et al., 2016)	79	636	8.05	75	8.48	YES
RED	(O'Gorman et al., 2016)	95	8731	91.91	2390	3.65	YES
Total		3874	36292	9.37	15057	2.41	
GVC	this publication	510	7298	14.31	1411	5.17	YES

as mentions of the same underspecified event represented at a more coarse-grained level of granularity, so-called event-hoppers.

In all the T2D approaches described above, event reference is established a posteriori after annotating texts word-by-word and sentence-by-sentence to mark events and event relations. Figure 1 gives a schematic overview for the T2D approach that indirectly constructs a referential representation from annotated mentions of events and participants. First, event and participant mentions need to be annotated in a single document and after that all these annotations need to be compared across all news articles to establish cross-document coreference. The more documents are included in the data set, the more comparisons need to be made. In the next section, we propose a new method that starts from registered events that are given a priori when annotating event references in texts so that we only need to compare mentions across relevant documents.

3. From data to text

For the reasons discussed in the previous Section, T2D methods do not provide the means nor the datasets to study *identity*, *reference* and *perspectives* of events on a large scale, since they are too small and lack sufficient ambiguity and variation. We therefore propose a novel *structured-data-to-text* (D2T) methodology, based on the notions **microworlds** and **reference texts**. *Microworlds* are structured representations of referents related to specific world events (e.g. human calamities or economic events). *Reference texts* are documents reporting on this data, e.g. news articles, blogs, and Wikipedia pages. In the D2T method, we start from some event registry that has been created by people a priori by hand and is publicly available as structured data. From these registries, we derive microworld representations of the unique event instances, their participants, location and date as a referential graph, as shown in Figure 2. Assuming that reference texts mainly refer to the corresponding microworld and not to other events and participants, we can establish the referential relation relatively easily and partially automatically.

By combining microworlds for similar but different events with their paired reference texts, we increase the referential ambiguity for systems that need to reconstruct the microworld from the texts, hence approximating the complexity of reference relations in reality across large volumes of text. By collecting news from different sources on the same or similar events, we approximate true variation in making reference from different perspectives. Furthermore, the fact that the data on events from which we start has been created from the perspective of general human interest (e.g. gun violence incident reports) avoids the never-ending discussion on what establishes an event in text. More practically, the D2T method is much less labor-intensive than T2D, because a rich and consistent set of event properties and links to its supporting documents are often provided within a microworld by the original data author. Finally, since the underlying data is often created manually, its quality is very high.

3.1. Desiderata

Our method operates best on resources with: 1. *links between structured data and reporting texts* 2. *disambiguated/unique and consistently defined* events and event properties following Linked Data principles 3. *open, available* data 4. *high volume*, since more data typically exhibits higher referential ambiguity and variation. If all four desiderata are fulfilled, the conversion of the data to microworlds and reference texts is a matter of writing data manipulation scripts. In practice, resource properties are often not ideal, thus requiring some additional work - however, the amount of annotation or retrieval needed is far lower/incomparable to the exhaustive annotation processes in T2D.

3.2. Resource availability

Table 2 provides description of several public resources that satisfy most of the desiderata. The resources register event incidents with rich properties such as participants, location, and incident time, and they provide pointers to one or more reference texts. The number of events and documents is usually high, for instance there are $\sim$9K incidents in RA, and $\sim$231K incidents in GVA.

In a pilot, we successfully obtained data from these resources with extreme ambiguity and variation, while maintaining the identity and reference relations and without having to annotate large quantities of texts word-by-word. Following the D2T method, we obtained over ten thousand news articles and over five thousand incidents from GVA and FR. The data from this pilot was used as basis for a referential quantification task entitled "counting events and

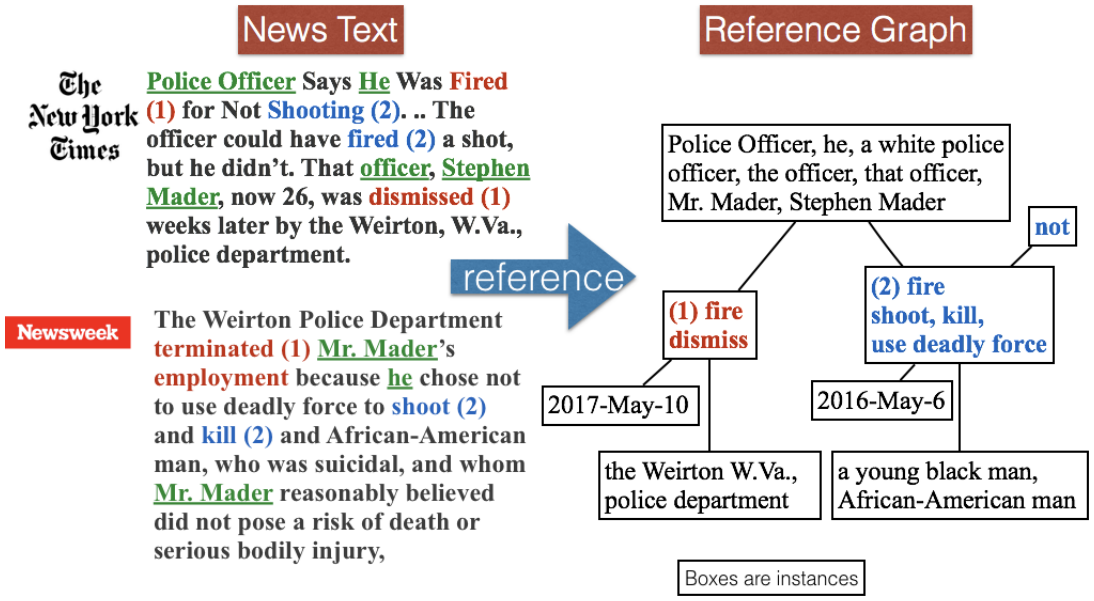

Figure 1: Overview of the T2D method: deriving a referential graph from mentions across different news text.

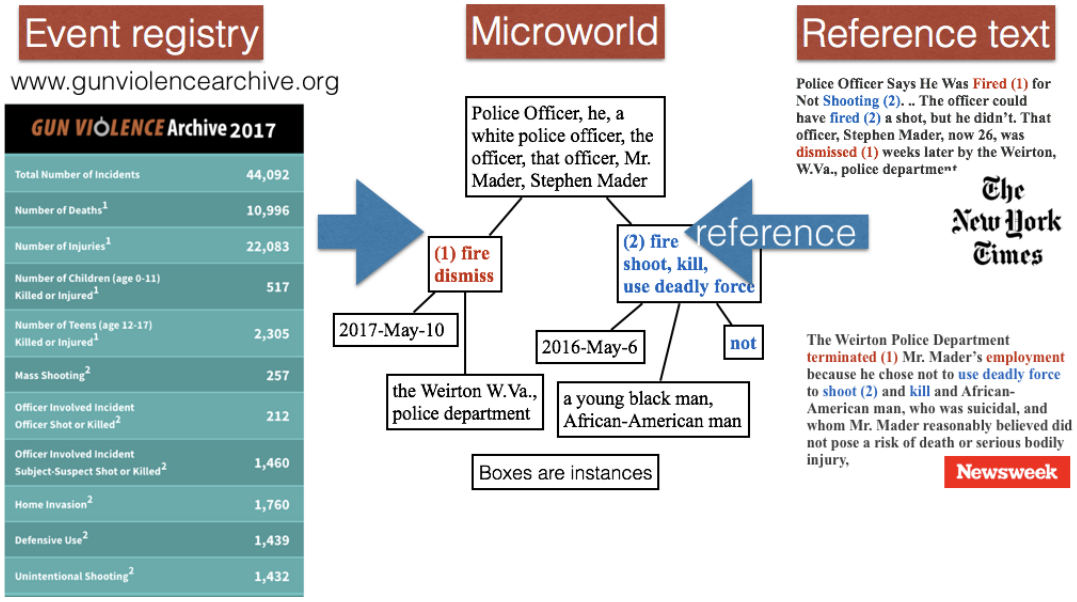

Figure 2: Overview of the D2T method: representing structured event data first as microworlds and secondly pairing it with reference texts.

participants within highly ambiguous data covering a very long tail", hosted at SemEval-2018.[9]

In addition to the resources presented in Table 2, data with similar properties can be obtained from Wikipedia and structured databases such as Wikidata[10], Yago2[11], and DBpedia[12] with little effort. This can either be done through direct extraction, or through smart querying of the data (Elbassuoni et al., 2010; Knuth et al., 2015; Hewlett et al., 2016). For example, a simple query on Wikidata for event instances belonging to certain event classes (i.e. explosion, crime, natural disaster, accident, sport, election), already yields over 70k events with structured data (type of event, location and time) that can form the basis for creating microworlds. Many of these events can be traced back to Wikipedia pages, that describe these events in textual form. Such Wikipedia pages often include further links to news articles as references to substantiate the information given. By using Wikipedia as the glue between the structured microworld data and the reference texts, one can obtain a reliable mapping of texts with framings and representations of the referential events.

[9] https://competitions.codalab.org/competitions/17285

[10] https://query.wikidata.org

[11] http://www.yago-knowledge.org

[12] http://events.dbpedia.org/ns/core#Event

Table 2: Potential event data for extracting microworlds and reference texts. Numbers marked with '*' are estimates.

Name	Topic	Structured data	Nr docs	Nr incidents	From year	To year	Locations	Reference texts
ASN incident database[2]	aircraft safety occurrences	fatalities, locations, time, other domain data	32K	21K	1919	2017	world	news, reports, social media
ASN Wikibase[3]	aircraft safety occurrences	fatalities, locations, time, other domain data	310K	207K	1905	2017	world	news, reports, social media
Fire Incident Reports (FR)[4]	fire disasters	publishing time and location	1K	1K	2004	present	USA	reports
Global nonviolent action DB[5]	social justice/protests	incident location and time	*6K	1K	1955	present	world	various
Gun Violence Archive (GVA)[6]	gun violence	fatalities, locations, time, participant roles, weapon information	*462K	231K	2012	present	USA	news
Legible news[7]	science, sports, business, economics, law, crime, disasters, accidents, ...	/	*20K	*15K	2014	present	world	news
Railways Archive (RA)[8]	railway accidents	casualties, locations, time, vehicle operators	5K	9K	1803	present	UK, Ireland	news
TOTAL			***836K**	***485K**				

4. The Gun Violence Corpus

D2T resources provide a very valuable link between microworlds and reference texts. We hypothesize that this link is also very useful for mention annotation of reference texts, because the microworld information provides a summary of the incidents reported. Annotation of mentions can then be seen merely as a task of marking evidence for the incident and its characteristics in the supporting text documents. Following this new view on the annotation process, we created the Gun Violence Corpus on top of our pilot data. This Section describes the details of its development.

4.1. Annotation Task and Guidelines

The Gun Violence Corpus (GVC) consists of 241 unique incidents for which we have structured data on a) location, b) time c) the name, gender and age of the victims and d) the status of the victims after the incident: killed or injured. For these data, we gathered 510 news articles following the D2T approach. The structured data and articles report on a variety of gun violence incidents, such as drive-by shootings, murder-suicides, hunting accidents, involuntary gun discharges, etcetera.

The documents have been manually annotated for all mentions that make reference to the gun violence incident at hand. More specifically, the annotation process involved three basic steps:

- Annotating the *event type* of every mention that refers to a gun violence incident in the structured data;
- Annotating the *victim(s)* involved in the mention referring to a shooting in the structured data;
- Annotating every mention related to gun violence but NOT referring to the incident in the structured data (*other incidents or generic mentions*).

Based on these annotations, we can infer coreference relations: in case that two or more mentions have the same annotations (event type and victims) AND they both relate to the same incident ID in the structured data, we can infer that these mentions are coreferential.

To further capture the referential complexity and diversity of event descriptions in text, we designed an event schema that captures subevent relations in addition to the above incident references, see Figure 3.

The main event ("the gun incident") is basically a container that can be split into several more fine-grained events that stand in some implication relation to each other. In this case the bag of events consists of five events: *Firing a gun*, *Hitting someone* or *Missing someone*. From *Hitting someone* follows *Injuring* and in some cases *Death*. Apart from these events, many articles also contain references to gun violence in a more general way or not related to the structured data. These have been labeled *Generic* and *Other*.

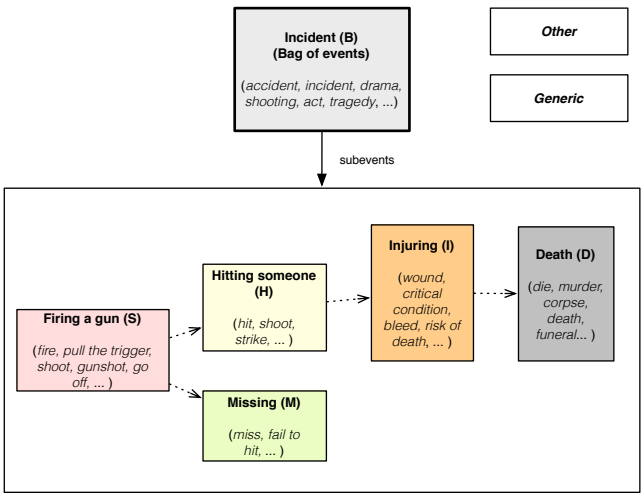

Figure 3: The event scheme used for the annotation of gun violence events.

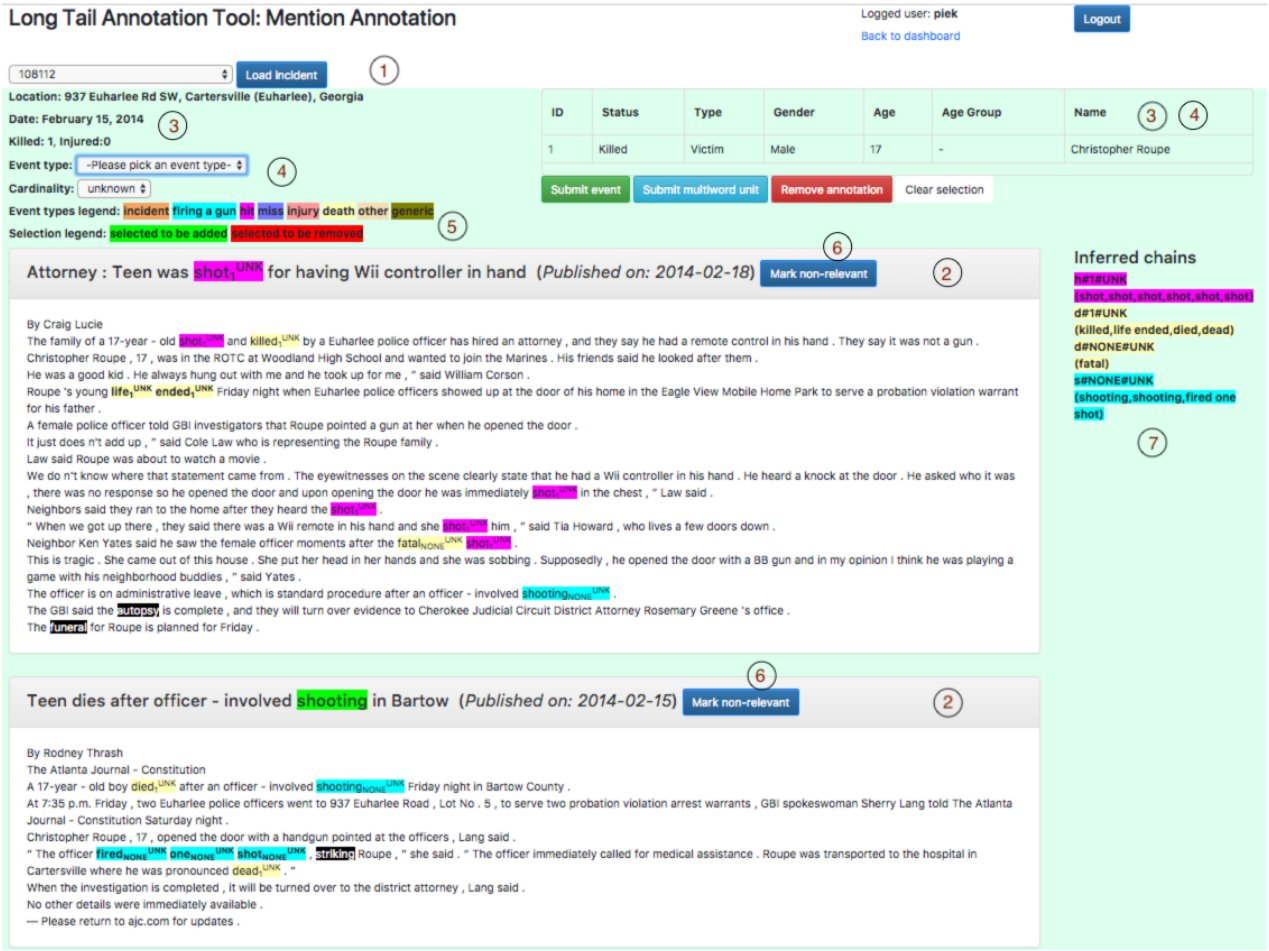

Figure 4: Annotation environment for annotating mentions in Reference Texts related to structured data.

4.1.1. The GVC Annotation Guidelines

We annotated all mentions denoting but also implying one of the predefined event classes. For example, a *funeral*, an *autopsy* or the process of *grieving* imply that someone died. A *shooter* and *killer* imply respectively the event types *Firing a gun* and again *Death* in the context of this domain. Besides annotating verbal and nominal expressions, we also annotated mentions of: other parts of speech (including adjectives and adverbs), idioms, multiword units, and collocations. In principle, we annotated the minimal span of a mention, usually the head, unless this would result in a meaningless annotation, e.g. we would annotate *critical condition* as a multiword unit instead of just the head *condition*.

Additional specification of the annotation decisions, such as: how we handled negation, the irrealis, ellipsis, phrasal verbs, and various cases of implicit event mentions, can be found in the full guidelines of GVC.[13]

The annotation of events has similarities to the Entities Relations Events (ERE) (Song et al., 2015) and the Rich Event Description (RED) (O'Gorman et al., 2016), but also differences. The Bag of Event level annotation corresponds to the event hopper annotation in ERE, but we differentiate specific subevent and subset relations as well, as is done in RED. Our annotation is more restricted to the specific events of this database only. However, it can be extended to other domains by defining a different event schema.

4.2. Annotation environment

To the best of our knowledge, there is no tool that starts from structured event data to annotate event mentions and event coreference relations. We therefore built our own environment for annotating events in reference texts that are related to structured data on an incident.

The goal of the tool is to allow annotators to find evidence in the reference texts for the event properties in the structured data. To support this goal, the tool reads the structured event data and presents the event properties, e.g. time, location, and participants, in a table. Annotators mark the event mentions, select the participants involved and select the type of event. The annotators only need to annotate the mentions of the predefined schema and not all other types of events.

By applying this strategy to all mentions within and across Reference Texts of an incident, we establish coreference and identity across the mentions. Notably, it is not needed to annotate coreference explicitly. Instead, the coreference chains are inferred by the annotation environment, based on the combination of two factors of the individual mention

[13] The full guidelines are available at `https://goo.gl/Yj1Hra`

annotations: event type and participants.
In addition, we have built in lexical support for the annotators, based on the set of already annotated event mentions. Reference text mentions which have been frequently annotated in other texts but not in the current one, are visually suggested to be also annotated. The annotators can then decide whether to accept this suggestion.
Figure 4 provides a screenshot of the mention annotation environment when the incident *108112* is loaded by the user *piek*. The incident selection menu is marked with (1) in the Figure. The selected incident is supported by two reference texts, rendered in the middle of the screen (marked with (2)). Annotators can select one or multiple mentions from this area for annotation. The top panel contains the structured data about the current incident (marked with (3)), followed by menus and a table for annotations of properties for the selected mention (4). Mentions in colors have already been annotated by this user, and the event type is signaled by the color. The color scheme is explained in detail in the legend (5). Moreover, inversely colored mentions (e.g. "funeral" and "autopsy" in Figure 4) are the ones proposed by the tool to be annotated additionally. Annotators can also discard individual documents with the 'Mark non-relevant' button (6). Finally, the area on the right displays the coreferential chains that the tool has inferred so far about the current incident (marked with (7) in the Figure).
The source code of the annotation software is publicly available on Github.[14]

4.3. Annotation process

Two linguistic students were hired to perform the annotations. After completing the training phase, which resulted in some simplifications of the guidelines, the students started with the annotation of the Gun Violence Corpus. In six weeks, the students annotated the 510 documents that are part of the corpus. In addition, 25 documents were selected in order to compute the inter-annotator-agreement (IAA). The first annotator annotated 432 event mentions in this set, whereas the second one annotated 457 event mentions. The annotators provided the same annotation in 350 cases, resulting in a Cohen's kappa coefficient (Cohen, 1960) of 0.72. According to Landis and Koch (1977), a score between 0.61 and 0.80 is considered *substantial*, from which we conclude that there was high agreement between the annotators. For comparison, ECB+ reported a Cohen's kappa coefficient of 0.68 for a similar size and agreement analysis to ours. ECB+ annotators only had to consider 2 incidents per topic with about 10 articles per incident and 1.8 sentences on average per article, whereas in our case, 510 documents need to be annotated for a few hundred incidents. In terms of speed, one annotator averaged 5 minutes per document, whereas the other took 4 minutes to annotate one document on average.
Unlike in T2D, our method scales only linearly instead of exponentially. Namely, to include documents that report on a new incident, one does not need to compare their mentions to all other incidents, since the structured data already guarantees they are not coreferential. In Table 1, we report statistics on the size of our corpus. Although our corpus annotated with mentions is currently smaller than existing datasets, the speed and the linear scalability of our method provide a promise that its size can increase up to the limit posed by the original structured data sources.

4.4. Corpus description

The GVC[15] contains 7,298 mentions, referring to 241 incidents. In total, 510 documents contain at least one mention. Table 3 presents the annotation frequency for each event type.

event type	annotation frequency
Death	2,206
Firing a gun	1,622
Hitting	1,122
Bag of events	755
Injuring	726
Other	596
Generic	270
Missing	2

Table 3: Mention frequency of each event type.

Most mentions in our Gun Violence Corpus refer to the event types *Death* and *Firing a gun*, respectively. In addition, about 4% of all mentions (i.e. 270 mentions), refer to generic uses of shooting and killings. Finally, it is not uncommon that the text refers to other incidents than the main incident of the article, which happens in about 8% of all mentions (i.e. 596). This means that systems can not fully rely on a one-incident-per-document heuristic to detect coreference chains.
Table 4 presents the most used expressions for each event type.

event type	most common expressions
Death	dead (305) died (285) killed (283)
Firing a gun	shooting (680) gunshot (247) went off (72)
Hitting	shot (801) shooting (83) struck (46)
Bag of events	shooting (247) incident (164) it (88)
Injuring	wound (175) injured (75) injuries (68)
Other	shot (105) shooting (70) killed (47)
Generic	accident (57) shooting (13) tragedy (11)
Missing	surgery (1) missed (1)

Table 4: Most common expressions used for event types

As presented in this Table, the most common expressions that are used to refer to event types are covered well in resources such as WordNet. For example, the most common expressions for the event type *Death* can be detected by correctly identifying the WordNet synsets *kill.v.01* (cause to die; put to death, usually intentionally or knowingly) and *killing.n.02* (the act of terminating a life). However, this

[14] https://github.com/cltl/LongTailAnnotation

[15] The corpus can be downloaded at: https://github.com/cltl/GunViolenceCorpus

is not the case for all expressions in the GVC. For example, expressions like *mourn* and *autopsy* that refer to the event type *Death* show that manual and automatic annotators can not fully rely on resources to detect all event types correctly, but that additional reasoning is needed. We analyze the referential potential of this corpus further in Vossen et al. (2018).

5. Conclusions

We discussed the problems in collecting large scale and high-quality text corpora for event extraction, and specifically with respect to identity and reference. We concluded that most data has been created through a text-to-data method, which faces the obstacles of data size and scalability. To circumvent these, we propose a scalable data-to-text method to create far more data with high quality, ambiguity, and variation. Following this method, we created the Gun Violence Corpus, whose development is reported in this paper. We present the specification and the guidelines of the annotation, as well as our annotation environment which was purposefully developed to support mention annotation for data-to-text use cases. Finally, we show that we achieve high agreement and annotation speed, and report statistics of the resulting corpus. For future works, we aim to compare our annotation process to traditional annotation using text-to-data tools such as CAT (Lenzi et al., 2012) to annotate the same documents used in this study.

6. Acknowledgements

The work presented in this paper was funded by the Netherlands Organization for Scientific Research (NWO) via the Spinoza grant, awarded to Piek Vossen in the project "Understanding Language by Machines".

7. Bibliographical References

Bejan, C. A. and Harabagiu, S. (2010). Unsupervised event coreference resolution with rich linguistic features. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1412–1422. Association for Computational Linguistics.

Chen, Z. and Ji, H. (2009). Graph-based event coreference resolution. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*, pages 54–57. Association for Computational Linguistics.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Cybulska, A. and Vossen, P. (2014). Using a sledgehammer to crack a nut? Lexical diversity and event coreference resolution. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 4545–4552.

Elbassuoni, S., Ramanath, M., Schenkel, R., and Weikum, G. (2010). Searching RDF graphs with SPARQL and keywords.

Guarino, N. (1999). The role of identity conditions in ontology design. In *International Conference on Spatial Information Theory*, pages 221–234. Springer.

Hewlett, D., Lacoste, A., Jones, L., Polosukhin, I., Fandrianto, A., Han, J., Kelcey, M., and Berthelot, D. (2016). Wikireading: A novel large-scale language understanding task over wikipedia. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1545. Association for Computational Linguistics.

Hovy, E., Mitamura, T., Verdejo, F., Araki, J., and Philpot, A. (2013). Events are not simple: Identity, non-identity, and quasi-identity. In *Workshop on Events: Definition, Detection, Coreference, and Representation*, pages 21–28.

Ilievski, F., Postma, M., and Vossen, P. (2016). Semantic overfitting: what 'world' do we consider when evaluating disambiguation of text? In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1180–1191.

Knuth, M., Lehmann, J., Kontokostas, D., Steiner, T., and Sack, H. (2015). The DBpedia Events Dataset. In *International Semantic Web Conference (Posters & Demos)*.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Lee, H., Recasens, M., Chang, A., Surdeanu, M., and Jurafsky, D. (2012). Joint entity and event coreference resolution across documents. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 489–500. Association for Computational Linguistics.

Lenzi, V. B., Moretti, G., and Sprugnoli, R. (2012). CAT: the CELCT Annotation Tool. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC) 2012*. European Language Resources Association (ELRA).

Liu, Z., Araki, J., Hovy, E. H., and Mitamura, T. (2014). Supervised within-document event coreference using information propagation. In *LREC*, pages 4539–4544.

Lu, J. and Ng, V. (2016). Event Coreference Resolution with Multi-Pass Sieves. In *LREC 2016*, pages 3996–4003.

Peng, H., Song, Y., and Roth, D. (2016). Event detection and co-reference with minimal supervision. In *EMNLP*, pages 392–402.

Vossen, P. and Cybulska, A. (2016). Identity and granularity of events in text. *Cicling 2016. Konya, Turkey*.

Vossen, P., Postma, M., and Ilievski, F. (2018). ReferenceNet: a semantic-pragmatic network for capturing reference relations. In *Global Wordnet Conference 2018, Singapore*.

8. Language Resource References

Baker, C. F., Fillmore, C. J., and Cronin, B. (2003). The structure of the framenet database. *International Journal of Lexicography*, 16(3):281–296.

Bejan, C. A. and Harabagiu, S. M. (2008). A Linguistic Resource for Discovering Event Structures and Resolving Event Coreference. In *LREC*.

Bethard, S., Derczynski, L., Savova, G., Pustejovsky, J.,

and Verhagen, M. (2015). SemEval-2015 task 6: Clinical TempEval. In *SemEval@ NAACL-HLT*, pages 806–814.

Boguraev, B., Pustejovsky, J., Ando, R., and Verhagen, M. (2007). TimeBank evolution as a community resource for timeml parsing. *Language Resources and Evaluation*, 41(1):91–115.

Caselli, T. and Morante, R. (2016). Vuacltl at semeval 2016 task 12: A crf pipeline to clinical tempeval. *Proceedings of SemEval*, pages 1241–1247.

Cybulska, A. and Vossen, P. (2014). Using a sledgehammer to crack a nut? Lexical diversity and event coreference resolution. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 4545–4552.

Hong, Y., Zhang, T., O'Gorman, T., Horowit-Hendler, S., Ji, H., and Palmer, M. (2016). Building a cross-document event-event relation corpus. *LAW X*.

Hovy, E., Mitamura, T., Verdejo, F., Araki, J., and Philpot, A. (2013). Events are not simple: Identity, non-identity, and quasi-identity. In *Workshop on Events: Definition, Detection, Coreference, and Representation*, pages 21–28.

Kingsbury, P. and Palmer, M. (2002). From TreeBank to PropBank. In *LREC*, pages 1989–1993. Citeseer.

Lee, H., Recasens, M., Chang, A., Surdeanu, M., and Jurafsky, D. (2012). Joint entity and event coreference resolution across documents. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 489–500. Association for Computational Linguistics.

Meyers, A., Reeves, R., Macleod, C., Szekely, R., Zielinska, V., Young, B., and Grishman, R. (2004). The NomBank project: An interim report. In *Proceedings of the Workshop Frontiers in Corpus Annotation at HLT-NAACL 2004*.

Minard, A.-L., Speranza, M., Urizar, R., Altuna, B., van Erp, M., Schoen, A., and van Son, C. (2016). MEANTIME, the NewsReader Multilingual Event and Time Corpus. In *Proceedings of LREC 2016*.

Mitamura, T., Yamakawa, Y., Holm, S., Song, Z., Bies, A., Kulick, S., and Strassel, S. (2015). Event nugget annotation: Processes and issues. In *Proceedings of the 3rd Workshop on EVENTS at the NAACL-HLT*, pages 66–76.

O'Gorman, T., Wright-Bettner, K., and Palmer, M. (2016). Richer Event Description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pages 47–56.

Peng, H., Song, Y., and Roth, D. (2016). Event detection and co-reference with minimal supervision. In *EMNLP*, pages 392–402.

Pradhan, S. S., Loper, E., Dligach, D., and Palmer, M. (2007). SemEval-2007 task 17: English lexical sample, SRL and all words. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 87–92. Association for Computational Linguistics.

Pustejovsky, J. and Verhagen, M. (2009). SemEval-2010 task 13: evaluating events, time expressions, and temporal relations (TempEval-2). In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 112–116. Association for Computational Linguistics.

Song, Z., Bies, A., Strassel, S., Riese, T., Mott, J., Ellis, J., Wright, J., Kulick, S., Ryant, N., and Ma, X. (2015). From Light to Rich ERE: Annotation of Entities, Relations, and Events. In *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 89–98. Association for Computational Linguistics.