SemEval-2018 Task 5: Counting Events and Participants in the Long Tail

Marten Postma, Filip Ilievski, Piek Vossen
Vrije Universiteit Amsterdam
The Netherlands
{m.c.postma, f.ilievski, piek.vossen}@vu.nl

Abstract

This paper discusses SemEval-2018 Task 5: a referential quantification task of counting events and participants in local, long-tail news documents with high ambiguity. The complexity of this task challenges systems to establish the meaning, reference and identity across documents. The task consists of three subtasks and spans across three domains. We detail the design of this referential quantification task, describe the participating systems, and present additional analysis to gain deeper insight into their performance.

1 Introduction

We present a "referential quantification" task that requires systems to establish the meaning, reference and identity of events[1] and participants in news articles. By "referential quantification", we mean questions concerning the number of incidents of an event type (e.g. *How many killing incidents happened in 2016 in Columbus, MS?*) or participants in roles (e.g. *How many people were killed in 2016 in Columbus, MS?*), as opposed to factoid questions for specific properties of individual events and entities (e.g. *When was 2pac murdered?*). The questions are given with certain constraints on the location, time, participants, and event types, which requires understanding of the meaning of words mentioning these properties (e.g. Word Sense Disambiguation), but also adequately establishing the identity (e.g. reference and coreference) across mentions. The task thus represents both an intrinsic and application-based evaluation, as systems are forced to resolve ambiguity of meaning and reference, as well as variation in reference in order to answer the questions.

Figure 1 shows an overview of our quantification task. We provide the participants with a set of questions and their corresponding news documents.[2] Systems are asked to distill event- and participant-based knowledge from the documents to answer the question. Systems submit both a numeric answer (*3* events in Figure 1), and the corresponding events with their mentions found in the provided texts (e.g., the leftmost incident in Figure 1 is referred to by the coreferring mentions "killed" and "assault" found in two separate documents). Systems are evaluated on both the numeric answers as well as on the sets of coreferring mentions. Mentions are represented by tokens and offsets provided by the organizers.

The incidents and their corresponding news articles are obtained from structured databases, which greatly reduces the need for annotation and mainly requires validation instead. Given this data and using a metric-driven strategy, we created a task that further maximizes ambiguity and variation of the data in relation to the questions. This ambiguity and variation includes a substantial amount of low-frequent, local events and entities, reflecting a large variety of long-tail phenomena. As such, the task is not only highly ambiguous but can also not be tackled by relying on the most frequent and popular (head) interpretations.

We see the following contributions of our task:
1. To the best of our knowledge, we propose the first task that is deliberately designed to address large ambiguity of meaning and reference over a high number of infrequent instances.
2. We introduce a methodology for creating large event-based tasks while avoiding a lot of annotation, since we base the task on structured data. The remaining annotation concerns targeted mentions given the structured data rather than full doc-

[1] By event, we denote a specific instance of an event, e.g. a killing incident happening at a specific location, time, and involving certain participants.

[2] Question parsing is unnecessary, as questions are provided in a structured format.

Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), pages 70–80
New Orleans, Louisiana, June 5–6, 2018. ©2018 Association for Computational Linguistics

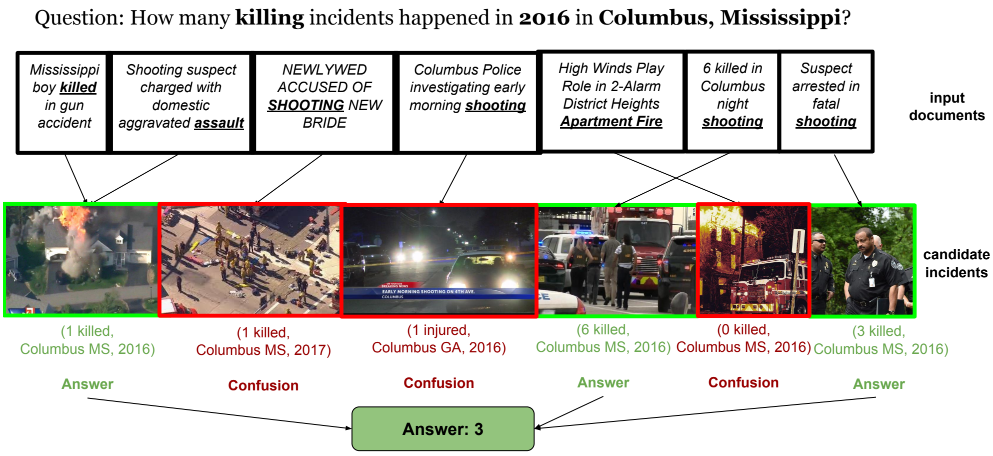
Question: How many **killing** incidents happened in **2016** in **Columbus, Mississippi**?
Mississippi boy killed in gun accident
Shooting suspect charged with domestic aggravated assault
NEWLYWED ACCUSED OF SHOOTING NEW BRIDE
Columbus Police investigating early morning shooting
High Winds Play Role in 2-Alarm District Heights Apartment Fire
6 killed in Columbus night shooting
Suspect arrested in fatal shooting
input documents
candidate incidents
(1 killed, Columbus MS, 2016)
(1 killed, Columbus MS, 2017)
(1 injured, Columbus GA, 2016)
(6 killed, Columbus MS, 2016)
(0 killed, Columbus MS, 2016)
(3 killed, Columbus MS, 2016)
Answer
Confusion
Confusion
Answer
Confusion
Answer
Answer: 3

Figure 1: Task overview. Systems are provided with a question and a set of input documents. Their goal is then to find the documents that fit the question constraints and reason over them to provide an answer.

uments with open-ended interpretations.

3. We made all of our code to create the task available,[3] which may stimulate others to create more tasks and datasets that tackle long-tail phenomena for other aspects of language processing, either within or outside of the SemEval competition.

4. This task provides insights into the strengths and weaknesses of semantic processing systems with respect to various long-tail phenomena. We expect that systems need to innovate by adjusting (deep) learning techniques to capture the referential complexity and knowledge sparseness, or by explicitly modeling aspects of events and entities to establish identity and reference.

2 Motivation & Target Communities

Expressions can have many different meanings and possibly an infinite number of references. At the same time, variation in language is also large, as we can make reference to the same things in many ways. This makes the tasks of Word Sense Disambiguation, Entity Linking, and Event and Nominal Coreference extremely hard. It also makes it very difficult to create a task that represents the problem at its full scale. Any sample of text will reduce the problem to a small set of meanings and references, but also to meanings that are popular at that time excluding many unpopular ones from the distributional long tail. Given this Zipfian distribution, a task that is challenging with respect to ambiguity, reference, and variation, and that is representative for the long tail as well, needs to fit certain constraints.

Our task directly relates to the following communities in semantic processing: 1. disambiguation and reference; 2. reading comprehension and question answering.

2.1 Disambiguation & Reference

Semantic NLP tasks are often limited in terms of the range of concepts and meanings that are covered. This is a necessary consequence of the annotation effort that is needed to create such tasks. Likewise, in Ilievski et al. (2016), we observed that most well-known datasets for semantic tasks have an extremely low ambiguity and variation. Even in datasets that tried to increase the ambiguity and temporal diversity for the disambiguation and reference tasks, we still measured a notable bias with respect to ambiguity, variance, dominance, and time. Overall, tasks and their datasets show a strong semantic overfitting to the head of the distribution (the most popular part of the world) and are not representative for the diversity of the long tail.

Our task differs from existing ones in that: 1. we deliberately created a task with a high number of event instances per event, many of which with similar properties, leading to high confusability 2. we present an application-based task which requires to perform on a combination of intrinsic tasks such as reference, disambiguation, and spatial-temporal reasoning, that are usually tested separately in existing tasks.

2.2 Reading Comprehension & Question Answering

In several recent tasks, systems are asked to answer entity-based questions, typically by point-

[3] https://github.com/cltl/LongTailQATask

ing to the correct segment or coreference chain in text, or by composing an answer by abstracting over multiple paragraphs/text pieces. These tasks are based on Wikipedia (SQuAD (Rajpurkar et al., 2016), WikiQA (Yang et al., 2015), QASent (Wang et al., 2007), WIKIREADING (Hewlett et al., 2016)) or on annotated individual documents (MARCO (Nguyen et al., 2016), CNN and DailyMail datasets (Hermann et al., 2015)).

Weston et al. (2015) outlined 20 skill sets, such as causality, resolving time and location, and reasoning over world knowledge, that are needed to build an intelligent QA system. These have been partially captured by the datasets MCTest (Richardson et al., 2013) and QuizBowl (Iyyer et al., 2014)), as well as the SemEval task on *Answer Selection in Community Question Answering* (Nakov et al., 2015, 2016).[4]

However, all these datasets avoid representing real-world referential ambiguity to its full extent by mainly asking questions that require knowledge about popular Wikipedia entities and/or text understanding of a single document.[5] Unlike existing work, our task deliberately addresses the referential ambiguity of the world beyond Wikipedia, by asking questions about long-tail events described in multiple documents. By doing so, we require deep processing of text and establishing identity and reference across single documents.

3 Task Requirements

Our quantification task consists of questions like *How many killing incidents happened in 2016 in Columbus, MS?* on a dataset that maximizes confusability of meaning, reference and identity. To guide the creation of such task, we defined five requirements that apply to the data for a single event type, e.g. *killing* (Postma et al., 2016).

Each event type should contain:

R1 Multiple event instances per event type, e.g. *the killing of Joe Doe* and *the killing of Joe Roe*.

R2 Multiple event mentions per event instance within the same document.

R3 Multiple documents with varying creation times that describe the same event.

R4 Event confusability by combining one or multiple confusion factors:

a) ambiguity of event mentions, e.g. *John Doe fires a gun*, and *John Doe fires a worker*.

b) variance of event mentions, e.g. *John Doe kills Joe Roe*, and *John Doe murders Joe Roe*.

c) time, e.g. *killing A that happened in January 2013*, and *killing B in October 2016*.

d) participants, e.g. *killing A committed by John Doe*, and *killing B committed by Joe Roe*.

e) location, e.g. *killing A that happened in Columbus, MS*, and *killing B in Houston, TX*.

R5 Representation of non-dominant events and entities, i.e. instances that receive little media coverage. Hence, the entities would not be restricted to celebrities and the events are not widely discussed such as general elections.

4 Data & Resources

In this Section, we present our data sources and an example document. We also discuss considerations of licensing and availability.

4.1 Structured data

The majority of the source texts in this task are sampled from structured databases that contain supportive news sources about gun violence incidents. While these texts already contain enough confusability with respect to the aspects defined in Section 3, we add confusion through leveraging structured data from two other domains: fire incidents and business.

As a direct consequence of using these databases and our exploitation strategy, we are able to satisfy all requirements we set in Section 3. These databases contain many event instances per event type (R1), multiple event mentions in the same document per event instance (R2), cover a wide spread of publishing times per event instance (R3), represent non-dominant events and entities (R5), and contain rich annotation of event properties that allows us to create high confusability (R4, see Section 5.3 for our methodology).

For a large portion of the information in the structured databases, we manually validated that this information could be found in the supportive news sources, and excluded the documents for which this was not the case. For the remaining documents, we performed automatic tests to filter out low-quality entries.

4.1.1 Gun Violence

The gun violence data is collected from the standard reports provided by the *Gun Violence Archive*

[4] The 2017 run can be found at `http://alt.qcri.org/semeval2017/task3/`.

[5] e.g. the Quiz Bowl dataset deliberately focuses on domains with much training data and frequent answers, thus avoiding the long tail problem in reference.

(GVA) website.[6] Each incident contains information about: 1. its **location** 2. its **time** 3. how many people were **killed** 4. how many people were **injured** 5. its **participants**. Participant information includes: (a) the **role**, i.e. victim or suspect (b) the **name** (c) the **age** 6. the **news articles** describing this incident. Table 1 provides a more detailed overview of the information available in the GVA.

Event Property	Granularity	Example value
Location	Address	Central Avenue
	City	Waynesboro
	State	Mississippi
Incident time	Day	14-3-2017
	Month	3-2017
	Year	2017
Participant	First name	John
	Last name	Smith
	Full name	John Smith

Table 1: Overview of the GVA incident properties of location, time, and participant.

To prevent systems from cheating (by using the structured data directly), the set of incidents and news articles is extended with news articles from the Signal-1M Dataset (Corney et al., 2016) and from the Web, that also stem from the gun violence domain, but are not found in the GVA.

4.1.2 Other domains

For the fire incidents domain, we make use of the *FireRescue1* reports,[7] which describe the following information about 417 incidents: 1. their **location** as a surface form 2. their **reporting time** 3. one **free text summary** describing the incidents. 4. **no** information about **participants**. Based on this information, we manually annotated the incident time and mapped the location to its representation in Wikipedia.

We further carefully selected a small amount of news articles from the business domain from The Signal-1M Dataset. Since these documents were not semantically annotated with respect to event information, we manually annotated this data with the same kind of information as the other databases: incident location, time, and information on the affected participants.

4.2 Example document

For each document, we provide its **title**, **content (tokenized)**, and **creation time**, e.g.:
Title: *$70K reward in deadly shooting near N. Philadelphia school*
Content: *A $70,000 reward is being offered for information in a quadruple shooting near a Roman Catholic school ...*
DCT: *2017-4-5*

4.3 Licensing & Availability

The news documents in our task are published on a very diverse set of (commercial) websites. Due to this diversity, there is no easy mechanism to check their licenses individually. Instead, we overcome potential licensing issues by distributing the data under the Fair Use policy.[8] [9]

During the SemEval-2018 period, but also afterwards, systems can easily test their submissions via our competition on Codalab.[10]

5 Task Design

For every incident in the task, we have fine-grained structured data with respect to its event type, location, time, and participants, and unstructured data in the form of the news sources that report on it. In this Section, we explain how we exploited this data in order to create the task. We present our three subtasks and the question template after which we outline the question creation. Finally, we explain how we divided the data into trial and test sets and provide some statistics about the data. For detailed information about the task, e.g. about the question and answer representation, we refer to the CodaLab website of the task.

5.1 Subtasks

The task contains two event-based subtasks and one entity-based subtask.

Subtask 1 (S1): Find the single event that answers the question e.g. *Which killing incident happened in Wilmington, CA in June 2014?* The main challenge is not to determine how many incidents satisfy the question, but to identify the documents that describe the single answer incident.

Subtask 2 (S2): Find all events (if any) that answer the question. This subtask differs from

[6] http://gunviolencearchive.org/reports/

[7] https://www.firerescue1.com/incident-reports/

[8] Fair use policy in USA: https://goo.gl/hXiEKL

[9] Fair use policy in EU: https://goo.gl/s8V5Zs

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

S1 in that the system now also has to determine the number of answer incidents, which makes this subtask harder. To make it more realistic, we also include questions with zero as an answer.

Subtask 3 (S3): Find all participant-role relations that answer the question e.g. *How many people were killed in Wilmington, CA with the last name Smith?* The goal is to determine the number of entities that satisfy the question. The system not only needs to identify the relevant incidents, but also to reason over the participant roles.

5.2 Question Template

Questions in each subtask consist of an event type and two event properties.

Event type We consider four event types in this task described through their representation in WordNet (Fellbaum, 1998) and FrameNet (F. Baker et al., 1998). Each question is constrained by exactly one event type.

event type	description	meanings
killing	at least one person is killed	wn30:killing.n.02 wn30:kill.v.01 fn17:Killing
injuring	at least one person is injured	wn30:injure.v.01 wn30:injured.a.01 fn17:Cause_harm fn17:Experience_bodily_harm
fire_burning	the event of something burning	wn30:fire.n.01 fn17:Fire_burning
job_firing	terminated employment	wn30:displace.v.03 fn17:Firing

Table 2: Description of the event types. The meanings column lists meanings that best describe the event type. It contains both FrameNet 1.7 frames (prefixed by *fn17*) and WordNet 3.0 synsets (prefixed by *wn30*).

Event properties For each event property in our task (time, location, participants), we distinguish between three levels of granularity (see Table 1). In addition, we make a distinction between the surface form and the meaning of an event property value. For example, the surface form *Wilmington* can denote several meanings: the Wilmington cities in the states of California, North Carolina, and Delaware. When composing questions, for time and location we take the semantic (meaning) level, while for participants we use the surface form of their names. This is because the vast majority of the participants in our task are long tail instances which have no semantic representation in a structured knowledge base.

5.3 Question Creation

Our question creation strategy consists of three consecutive phases: question composition, generation of answer and confusion sets, and question scoring. These steps are common for both the event-based subtasks (S1 and S2) and the entity-based subtask S3.

1. Question composition We compose questions based on the template described in Section 5.2. This entails: 1. choice of a subtask 2. choice of an event type, e.g. *killing* 3. choice of two event properties (e.g. *time and location*) with their corresponding granularities (e.g. *month and city*) and concrete values (e.g. *June 2014 and Wilmington, CA*). This step generates a vast amount of potential questions (hundreds of thousands) in a data-driven way, i.e. we select the event type and properties per question purely based on the combinations we find in our data. Example questions are:

Which killing event happened in June 2014 in Wilmington, CA? (subtask S1)

How many killing events happened in June 2014 in Wilmington, CA? (subtask S2)

How many people were killed in June 2014 in Wilmington, CA? (subtask S3)

2. Answer and confusion sets generation For each generated question, we define a set of answer and confusion incidents with their corresponding documents. **Answer** incidents are the ones which entirely fit the question parameters, e.g. *all killing incidents that occur in June 2014 and in the city of Wilmington, CA*. **Confusion** incidents fit some, but not all, values of the question parameters , i.e. they differ with respect to an event type or property (e.g. *all fire incidents in June 2014 in Wilmington, CA*; or *all killings in June 2014, but not in Wilmington, CA*; or *all killings in Wilmington, CA, but not in June 2014*).

3. Question scoring The generated questions with their corresponding answers and confusion are next scored with respect to several metrics that measure their complexity. The per-question scores allow us to detect and remove the “easy” ones, and keep those that: 1. have a high number of answer incidents (only applicable to S2 and S3) 2. have a high number of confusion incidents 3. have a high average number of answer and confusion documents, i.e. news sources describing the answer and the confusion incidents correspondingly 4. have a high temporal spread with respect to the publishing dates reporting on each incident from the answer and confusion incidents 5. have a high ambiguity with respect to the surface forms of an event property value in a granularity level (e.g. we would favor *Wilmington*, since it is a city in at least three US states in our task data).

5.4 Data Partitioning

We divided the overall task data into two partitions: **trial** and **test** data. In practice, we separated these two data partitions by reserving one year of news documents (2017) from our task for the trial data, while using all the other data as test data.

The trial data stems from the gun violence domain, whereas the test data also contains data from the fire incidents and business domain. A subset of the trial and test data has been annotated for event coreference. Table 3 presents the most important statistics of the trial and test data.

	S	#Qs	Avg answer	Avg # answer docs
trial	1	424	1.00	1.68
	2	469	4.22	7.68
	3	585	5.48	5.47
test	1	1032	1.00	1.60
	2	997	3.79	6.64
	3	2456	3.66	3.74

Table 3: General statistics about trial and test data. For each subtask (*S*), we show the number of questions (*#Qs*), the average answer (*Avg answer*), and the average number of answer documents (*Avg # answer docs*).

We made an effort to make the trial data representative for the test data with respect to the main aspects of our task: its referential complexity, high confusability, and long-tail instances. Despite the fact that the trial data contains less questions than the test data, Table 3 shows that it is similar to the test data with respect to the core properties, meaning that the trial data can be used as training data.

6 Evaluation

This Section describes the evaluation criteria in this task and the baselines we compare against.

6.1 Criteria

Evaluation is performed on three levels: incident-level, document-level, and mention-level.

The incident-level evaluation compares the numeric answer provided by the system to the gold answer for each of the questions. The comparison is done twofold: by exact matching and by Root Mean Square Error (RMSE) for difference scoring. The scores per subtask are then averaged over all questions to compute a single incident-level evaluation score.

The document-level evaluation compares the set of answer documents between the system and the gold standard, resulting in a value for the customary metrics of Precision, Recall, and F1 per question. The scores per subtask are then averaged over all questions to compute a single document-level evaluation score.

The mention-level evaluation is a cross-document event coreference evaluation. Mention-level evaluation is only done for questions with the event types *killing* or *injuring*. We apply the customary metrics to score the event coreference: BCUB (Bagga and Baldwin, 1998), BLANC (Recasens and Hovy, 2011), entity-based CEAF (CEAF_E) and mention-based CEAF (CEAF_M) (Luo, 2005), and MUC (Vilain et al., 1995). The final F1-score is the average of the F1-scores of the individual metrics. The set of mentions to annotate should conform to the schema defined in the task annotation guidelines.[11]

6.2 Baselines

To stimulate participation in general and to stimulate approaches beyond surface form or majority class strategies, we implemented one baseline to infer incidents per subtask and one baseline for mention annotation.[12]

Incident inference baseline This baseline uses surface forms based on the question components to find the answer documents. We only consider documents that contain the label of the event type or at least one of its WordNet synonyms. The labels of locations and participants are queried directly in the document (e.g. if the location requested is the *US state of Texas*, then we only consider documents that contain the surface form *Texas*, and similarly for participants such as *John*). The temporal constraint is handled differently: we only consider documents whose publishing date falls within the time requested in the question.

For subtask 1, this baseline assumes that all documents that fit the created constraints are referring

[11] Link to the guidelines: https://goo.gl/8JpwCE.

[12] The code of the baselines can be found here: https://goo.gl/MwSqBj.

to the same incident. If there is no such document, then the baseline does not answer the question (because S1 always has at least one supporting document). For subtask 2, we assume that none of the documents are coreferential. Hence, if 10 documents match the constraints, we infer that there are also 10 corresponding incidents. No baseline was implemented for subtask 3.

Mention annotation baseline We annotate mentions of events of type *killing* and *injuring*, when these surface forms or their synonyms in WordNet are found as tokens in a document. We assume that all mentions of the same event type within a document are coreferential, whereas all mentions found in different documents are not.

7 Participants

In this Section, we describe the systems that took part in SemEval-2018 task 5. We refer to the individual system papers for further information.

NewsReader (Vossen, 2018) consists of three steps: 1. the event mentions in the input documents are represented as Event-Centric Knowledge Graphs (ECKGs). 2. the ECKGs of all documents are compared to each other to decide which documents refer to the same incident, resulting in an incident-document index. 3. the constraints of each question (its event type, time, participant names, and location) are matched with the stored ECKGs, resulting in a number of incidents and source documents for each question.

NAI-SEA (Liu and Li, 2018) consists of three components: 1. extraction of basic information on time, location, and participants with regular expressions, named entity recognition, and term matching; 2. event classification with an SVM classifier; 3. document similarity by applying a classifier to detect similar documents. In terms of resources, NAI-SEA combines the training data with data on American cities, counties, and states.

Team **FEUP** (Abreu and Oliveira, 2018) developed an experimental system to extract entities from news articles for the sake of Question & Answering. For this main task, the team proposed a supervised learning approach to enable the recognition of two different types of entities: Locations (e.g. *Birmingham*) and Participants (e.g. *John List*). They have also studied the use of distance-based algorithms (using Levenshtein distance and Q-grams) for the detection of documents' closeness based on entities extracted.

Team **ID-DE** (Mirza et al., 2018) created KOI (Knowledge of Incidents), a system that builds a knowledge graph of incidents, given news articles as input. The required steps include: 1. Document preprocessing using various semantic NLP tasks such as Word Sense Disambiguation, Named-Entity Recognition, Temporal expression recognition, and Semantic Role Labeling. 2. Incident extraction and document clustering based on the output of step 1. 3. Ontology construction to capture the knowledge model from incidents and documents which makes it possible to run SPARQL queries on the ontology to answer the questions.

8 Results

R	Team	s2_inc_acc norm	s2_inc_acc (% of Qs answered)	s2_inc rmse
1	FEUP	**26.38**	26.38 (100.0%)	6.13
2	*NewsReader	21.87	21.87 (100.0%)	43.96
3	Baseline	18.25	18.25 (100.0%)	8.50
4	NAI-SEA	17.35	17.35 (100.0%)	20.59
5	ID-DE	13.74	20.36 (67.5%)	6.15

Table 4: For subtask 2, we report the normalized incident-level accuracy (*s2_inc_acc norm*), the accuracy on the answered questions only (*s2_inc_acc*), and the RMSE value (*s2_inc rmse*). Systems are ordered by their rank (*R*).

R	Team	s3_inc_acc norm	s3_inc_acc (% of Qs answered)	s3_inc rmse
1	FEUP	**30.42**	30.42 (100.0%)	478.71
2	*NewsReader	21.05	21.05 (100.0%)	296.45
3	NAI-SEA	20.20	20.2 (100.0%)	13.45
4	ID-DE	12.87	19.32 (66.61%)	7.87

Table 5: For subtask 3, we report the normalized incident-level accuracy (*s3_inc_acc norm*), the accuracy on the answered questions only (*s3_inc_acc*), and the RMSE value (*s3_inc rmse*). Systems are ordered by their rank (*R*).

Before we report the system results, we introduce a few clarifications regarding the result tables:

1. For the incident- and document-level evaluation, we report both the performance with respect to the subset of questions answered and a **normalized score**, which indicates the performance on all questions of a subtask. If a submission provides answers for all questions, the normalized score will be the same as the non-normalized score.

2. Contrary to the other metrics, a lower **RMSE** value indicates better system performance. In addition, the RMSE scores have not been normalized since it is not reasonable to set a default value for non-answered questions.

3. **The mention-level evaluation** was the same across all three subtasks. For this reason, results are only reported once (see Section 8.3).

4. The teams whose member **co-organized SemEval-2018 task 5** are marked explicitly with an asterisk in the results.

8.1 Incident-level evaluation

The incident-level evaluation assesses whether the system provided the right numeric answer to a question. The results of this evaluation are given in the Tables 4 and 5, for the subtasks 2 and 3 correspondingly.[13] On both subtasks, the order of the participating systems is identical, team *FEUP* having the highest score.

These tables also show the RMSE values, which measure the proximity between the system and the gold answer, punishing cases where the absolute difference between them is large. While for subtask 2 the system with the lowest error rate corresponds to the system with the highest accuracy, this is different for subtask 3. *NAI-SEA*, ranked third in terms of accuracy, has the lowest RMSE. This means that although their answers were not exactly correct, they were on average much closer to the correct answer than those of the other systems. This is more notable in subtask 3 since here the range of answers is larger than in subtask 2 (the maximum answer in subtask 3 is 171).

We performed additional analysis to compare the performance of systems per subtype and per numeric answer class. Table 6 shows that the system *FEUP* is not only superior in terms of incident-level accuracy overall, but this is also mirrored for most of the event types, especially those corresponding to the gun violence domain. On the other hand, Figure 2 shows the accuracy distribution of each system per answer class. Notably, for most systems the accuracy is highest for the questions with answer 0 or 1, and gradually declines for higher answers, forming a Zipfian-like distribution. The exception here is the team *ID-DE*, whose accuracy is almost uniformly spread across the various answer classes.

8.2 Document-level evaluation

The intent behind document-level evaluation is to assess the ability of systems to distinguish between answer and non-answer documents. The tables 9, 10, and 11 present the F1-scores for the subtasks 1, 2, and 3, respectively. Curiously, the system ranking is very different and almost opposite compared to the incident-level rankings, with the system *NAI-SEA* being the one with the highest F1-score. This can be explained by the multifaceted nature of this task, in which different systems may optimize for different goals.

Next, we investigated the F1-scores of systems per event property pair. As shown in Table 7, the best-performing system consistently has the highest performance over all pairs of event properties.

R	Team	s1_doc_f1 norm	s1_doc_f1 (% of Qs answered)
1	NAI-SEA	**78.33**	78.33 (100.0%)
2	ID-DE	36.67	82.99 (44.19%)
3	FEUP	24.65	24.65 (100.0%)
4	*NewsReader	23.82	46.2 (51.55%)
5	Baseline	11.09	67.33 (16.47%)

Table 9: For subtask 1, we report the normalized document-level F1 (*s1_doc_f1 norm*) and the accuracy on the answered questions only (*s1_doc_f1*). Systems are ordered by their rank (*R*).

R	Team	s2_doc_f1 norm	s2_doc_f1 (% of Qs answered)
1	NAI-SEA	**50.52**	50.52 (100.0%)
2	ID-DE	37.24	55.16 (67.5%)
3	*NewsReader	36.91	36.91 (100.0%)
4	FEUP	30.51	30.51 (100.0%)
5	Baseline	26.38	26.38 (100.0%)

Table 10: For subtask 2, we report the normalized document-level F1 (*s2_doc_f1 norm*) and the accuracy on the answered questions only (*s2_doc_f1*). Systems are ordered by their rank (*R*).

R	Team	s3_doc_f1 norm	s3_doc_f1 (% of Qs answered)
1	NAI-SEA	**63.59**	63.59 (100.0%)
2	ID-DE	46.33	69.56 (66.61%)
3	*NewsReader	26.84	26.84 (100.0%)
4	FEUP	26.79	26.79 (100.0%)

Table 11: For subtask 3, we report the normalized document-level F1 (*s3_doc_f1 norm*) and the accuracy on the answered questions only (*s3_doc_f1*). Systems are ordered by their rank (*R*).

8.3 Mention-level evaluation

Table 8 shows the event coreference results for the participating systems: *ID-DE* and *NewsReader*, as well as our baseline. The columns present the F1-score for the metrics BCUB, BLANC, CEAF_E. CEAF_M, and MUC. The final column indicates

[13] Incident-level evaluation was not performed for subtask 1, because per definition, its answer is always 1.

Event type	Subtask	#Qs	FEUP	ID-DE	NAI-SEA	*NewsReader	Baseline
fire_burning	S2	79	40.51	-	31.65	39.24	**49.37**
	S3	0	-	-	-	-	-
injuring	S2	543	**21.92**	^13.44	14.36	21.73	17.68
	S3	1502	**30.49**	^8.39	16.78	23.17	-
job_firing	S2	4	0.0	-	25.0	25.0	**50.0**
	S3	26	**30.77**	-	26.92	15.38	-
killing	S2	371	**30.19**	^17.25	18.6	18.33	12.13
	S3	928	**30.28**	^20.47	25.54	17.78	-

Table 6: For subtask 2 (*S2*) and subtask 3 (*S3*), we report the incident-level accuracy and the number of questions (*#Qs*) per event type. The best result per event type for a subtask is marked in bold. ‘^’ indicates that the accuracy is normalized for the number of answered questions, in cases where a system answered a subset of all questions.

.

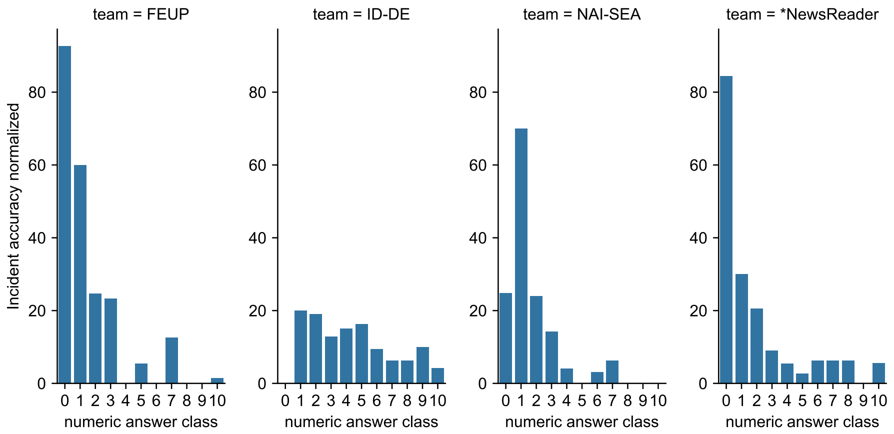

Figure 2: Incident-level accuracy of all systems per numeric answer class for subtask 2. The class *10* represents all answers of 10 or higher.

Event properties	Subtask	#Qs	FEUP	ID-DE	NAI-SEA	*NewsReader	Baseline
location&time	S1	594	23.06	^26.64	**82.91**	^26.22	^8.71
	S2	680	30.95	^41.81	**49.99**	39.22	28.61
	S3	1335	26.4	^41.55	**63.27**	36.15	-
participant&location	S1	140	13.48	^43.86	**70.22**	^11.83	^9.76
	S2	49	14.66	^21.26	**50.41**	13.53	10.02
	S3	301	14.2	^44.28	**62.38**	6.65	-
participant&time	S1	298	33.06	^53.28	**73.01**	^24.65	^16.47
	S2	268	32.27	^28.55	**51.87**	35.34	23.71
	S3	820	32.06	^54.88	**64.56**	19.09	-

Table 7: Document-level F1-score and number of questions (*#Qs*) for each subtask (*S1, S2, and S3*) and event property pair as given in the task questions. The best result per property pair for a subtask is marked in bold. ‘^’ indicates that the F1-score is normalized for the number of answered questions, in cases where a system answered a subset of all questions.

R	Team	BCUB	BLANC	CEAF_E	CEAF_M	MUC	AVG
1	ID-DE	**44.61%**	**31.59%**	37.45%	**47.23%**	**53.12%**	**42.8%**
2	*NewsReader	37.28%	28.11%	**42.15%**	46.16%	46.29%	40.0%
3	Baseline	6.14%	0.89%	13.3%	8.45%	3.59%	6.47%

Table 8: Results for mention-level evaluation, scored with the customary event coreference metrics: BCUB (Bagga and Baldwin, 1998), BLANC (Recasens and Hovy, 2011), entity-based CEAF (CEAF_E) and mention-based CEAF (CEAF_M) (Luo, 2005), and MUC (Vilain et al., 1995). The individual scores are averaged in a single number (*AVG*), which is used to rank (*R*) the systems.

the mean F1-score over these five metrics, which is used to rank the participants. The Table shows that the system *ID-DE* has a slightly better event coreference score on average over all metrics than the second-ranked system, *NewsReader*.

9 Conclusions

In this paper we have introduced SemEval-2018 Task 5, a referential quantification task of counting events and participants in local news articles with high ambiguity. The complexity of this task challenges systems to establish the meaning, reference, and identity across documents. SemEval-2018 Task 5 consists of two subtasks of counting events, and one subtask of counting event participants in their corresponding roles. We evaluated system performance with a set of metrics, on three levels: incident-, document-, and mention-level.

We described the approaches and presented the results of four participating systems, as well as two baseline algorithms. All four teams submitted a result for all three subtasks, and two teams participated in the mention-level evaluation. We observed that the ranking of systems differs dramatically per evaluation level. Given the multifaceted nature of this task, it is not surprising that different systems optimized for different goals. Although the systems are able to retrieve many of the answer documents, the highest accuracy of counting events or participants is 30%. This suggests that further research is necessary in order to develop complete and robust models that can natively deal with the challenge of counting referential units within sparse and ambiguous textual data.

Out-of-competition participation is enabled by the Codalab platform, where this task was hosted.

References

Carla Abreu and Eugénio Oliveira. 2018. FEUP at SemEval-2018 Task 5: An Experimental Study of a Question Answering System. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*. Association for Computational Linguistics.

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*, volume 1, pages 563–566. Granada.

David Corney, Dyaa Albakour, Miguel Martinez, and Samir Moussa. 2016. What do a million news articles look like? In *Proceedings of the First International Workshop on Recent Trends in News Information Retrieval co-located with 38th European Conference on Information Retrieval (ECIR 2016), Padua, Italy, March 20, 2016.*, pages 42–47.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database* . The MIT Press, Cambridge, MA ; London.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701.

Daniel Hewlett, Alexandre Lacoste, Llion Jones, Illia Polosukhin, Andrew Fandrianto, Jay Han, Matthew Kelcey, and David Berthelot. 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.

Filip Ilievski, Marten Postma, and Piek Vossen. 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. The COLING 2016 Organizing Committee.

Mohit Iyyer, Jordan L Boyd-Graber, Leonardo Max Batista Claudino, Richard Socher, and Hal Daumé III. 2014. A Neural Network for Factoid Question Answering over Paragraphs. In *EMNLP*, pages 633–644.

Yingchi Liu and Quanzhi Li. 2018. NAI-SEA at SemEval-2018 Task 5: An Event Search System. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*. Association for Computational Linguistics.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 25–32. Association for Computational Linguistics.

Paramita Mirza, Fariz Darari, and Rahmad Mahendra. 2018. KOI at SemEval-2018 Task 5: Building Knowledge Graph of Incidents. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*. Association for Computational Linguistics.

Preslav Nakov, Lluís Màrquez, Walid Magdy, Alessandro Moschitti, Jim Glass, and Bilal Randeree. 2015. Semeval-2015 task 3: Answer selection in community question answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 269–281. Association for Computational Linguistics.

Preslav Nakov, Lluís Màrquez, Alessandro Moschitti, Walid Magdy, Hamdy Mubarak, abed Alhakim Freihat, Jim Glass, and Bilal Randeree. 2016. Semeval-2016 task 3: Community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 525–545. Association for Computational Linguistics.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *CoRR*, abs/1611.09268.

Marten Postma, Filip Ilievski, Piek Vossen, and Marieke van Erp. 2016. Moving away from semantic overfitting in disambiguation datasets. In *Proceedings of the Workshop on Uphill Battles in Language Processing: Scaling Early Achievements to Robust Methods*, pages 17–21, Austin, TX. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100, 000+ Questions for Machine Comprehension of Text. *CoRR*, abs/1606.05250.

Marta Recasens and Eduard Hovy. 2011. BLANC: Implementing the Rand index for coreference evaluation. *Natural Language Engineering*, 17(4):485–510.

Matthew Richardson, Christopher JC Burges, and Erin Renshaw. 2013. MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text. In *EMNLP*, volume 3, page 4.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th conference on Message understanding*, pages 45–52. Association for Computational Linguistics.

Piek Vossen. 2018. NewsReader at SemEval-2018 Task 5: Counting events by reasoning over event-centric-knowledge-graphs. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*. Association for Computational Linguistics.

Mengqiu Wang, Noah A Smith, and Teruko Mitamura. 2007. What is the Jeopardy Model? A Quasi-Synchronous Grammar for QA. In *EMNLP-CoNLL*, volume 7, pages 22–32.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards AI-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A Challenge Dataset for Open-Domain Question Answering. In *Proceedings of EMNLP*, pages 2013–2018. Citeseer.