SemEval-2021 Task 8: MeasEval – Extracting Counts and Measurements and their Related Contexts

Corey A Harper[1,2], **Jessica Cox**[1], **Curt Kohler**[1], **Antony Scerri**[1], **Ron Daniel, Jr.**[1] and **Paul Groth**[2]

[1]Elsevier Labs, Suite 800, 230 Park Avenue, New York, NY 10169, USA
[1]{*c.harper, j.cox, c.kohler, a.scerri, r.daniel*}*@elsevier.com*
[2]University of Amsterdam, Postbus 94323 / 1090 GH, Amsterdam
[2]{*c.a.harper, p.t.groth*}*@uva.nl*

Abstract

We describe MeasEval, a SemEval task of extracting counts, measurements, and related context from scientific documents, which is of significant importance to the creation of Knowledge Graphs that distill information from the scientific literature. This is a new task in 2021, for which over 75 submissions from 25 participants were received. We expect the data developed for this task and the findings reported to be valuable to the scientific knowledge extraction, metrology, and automated knowledge base construction communities.

1 Introduction

Counts and measurements are an important part of scientific discourse (Rijgersberg et al., 2011). It is relatively easy to find measurements in text (Foppiano et al., 2019a), but a bare measurement like 17mg is not informative without knowing what it is referring to. For example, it is important to know whether a quantity is 17mg of a medicine dosage or 17mg of concrete additive. Only recently have attempts been made to identify the named entity and property being measured (Hundman and Maamann, 2017). Extracting such information is challenging because the way scientists write can be ambiguous and inconsistent. Furthermore, the location of this information relative to the measurement can vary greatly, and might even be in a different sentence.

Being able to extract measurement information automatically can enable the construction of databases of measured properties. Such databases are important in biomedicine (Hao et al., 2016), engineering (Foppiano et al., 2019a), and other scientific disciplines (Bergmann et al., 2017), but the approaches used for populating these databases do not generalize widely. Furthermore, knowledge graphs (Hogan et al., 2021) frequently aggregate quantitative data reported in the literature and are often built through a largely manual curation process. Examples include: LITTERBASE[1] (Bergmann et al., 2017), which aggregates observations of marine litter distribution; NeuroElectro[2] (Tripathy et al., 2014), which collects information on electrophysiological properties of neurons; and various model organism databases like the Zebrafish Information Network[3] (Sprague, 2006), which provide summaries of gene information.

Beyond knowledge graphs and curated databases, clinical health contexts often require extraction of measured values for lab results and patient observations. Moreover, scientific research frequently relies on precise measurements for reproducibility of experimental methods (Kaiser, 2018). Measured property extraction could be of value in many other contexts, such as fact checking and news validation or in statistical analysis for public policy (Einav and Levin, 2014).

Research in information extraction and knowledge graph creation has concentrated on forming triples by extracting entities and relations (Konstantinova, 2014). Little attention has been paid to the extraction of measured properties, entities, and conditions or contexts, yet these elements are needed to place measurements into a database and for their subsequent use in comparison and calculation. Units and measures are an important part of the semantic web, though research has largely been focused on ontology design (Rijgersberg et al., 2013). There is, thus, a need for understanding the state of the art on this important task.

The aim of this paper is to introduce the MeasEval shared task for the extraction of counts, measurements, and related context from English-language scientific documents, as well as to present an analysis of the results of participant systems on

[1]https://litterbase.awi.de/
[2]https://neuroelectro.org/
[3]http://zfin.org/

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 306–316
Bangkok, Thailand (online), August 5–6, 2021.

the task.

The rest of this paper is organized as follows. We begin with a description of related work. This is followed by the description of the task itself (Section 3) and the associated data and annotation procedure (Section 4). The evaluation regime is detailed in Section 5 including baselines. Subsequently, we present an analysis of the results of the systems on the task. Finally, we summarize the various participating systems approaches and conclude.

2 Related Work

There is a substantial body of work discussing units of measurement, ontologies to describe them, systems designed to extract them, as well as related work on knowledge graphs of numerical attributes. Automated extraction of measured quantities, such as 520 +/- 8 items/kg, is straightforward and many tools exist to perform this task (Foppiano et al., 2019b; Deus et al., 2017; Hao et al., 2016). To build a knowledge graph, we must put these measurements in context. We need to determine the properties being measured (e.g. abundances), the entities that exhibit those properties (e.g. the Maowei Sea), and possible qualifying conditions under which measurements are obtained (e.g. the date and depth of the sampling). These properties, entities, and conditions can then be mapped to those that are used in the knowledge graph, so that the measurements can be normalized into a common system.

There are a number of ontologies that cover units of measurement, such as Quantities, Units, Dimensions, and Types Ontologies (QUDT)[4] and the Ontology of Units of Measure and Related Concepts (OM) (Rijgersberg et al., 2013). These and others are discussed in a survey paper by (Steinberg et al., 2017). Most of these ontologies focus on conversion between different systems of measurement, and on classifying types of measurement or domain of application, but do not necessarily address the "thing" being measured. The Joint Committee for Guides in Metrology's (JCGM) International Vocabulary of Metrology covers this in slightly more depth, discussing measurement units and quantity values, then talking about quantities themselves, which it defines as a "property of a phenomenon, body, or substance" (Joint Committee for Guides in Metrology, 2012). We find that this nomenclature, while precise, is likely to be confusing to non-metrologists from both an evaluation and annotation perspective, so to support the data annotation process for this task we use a simplified nomenclature.

Metrology research in the Semantic Web community is often focused on ontology alignment for Units of Measurement ontologies. Kaladevi et al. (2016) look at aligning unit ontologies to support merging data across many weather information systems, while Do and Pauwels (2013) more generally look at using MathML for aligning unit ontologies. Efforts around designing linked data models for semantic sensor streams for the Internet of Things also utilize the Units of Measurement ontology for representing measurement information (Barnaghi et al., 2013). None of this work addresses extraction of measurements and their contexts nor building knowledge graphs from such information.

Other research explores creating databases of numeric attributes. Kotnis and Garcıa-Duran (2019) infer new values using linear regression for neighboring entities in a knowledge graph. Gupta et al. (2015) use a logistic regression with distributional vectors. Davidov and Rappoport (2010) use a system of averages and boundary values to infer an estimated numeric attribute value. Rather than imputing new values from related entities, MeasEval starts from a value and puts it into the context of measured entities and measured properties, working toward a knowledge representation of numeric data.

3 Task Description

MeasEval is an entity recognition and semantic relation extraction task focused on finding counts and measurements, attributes of those quantities, and additional information including measured entities, properties, and measurement contexts.

MeasEval is composed of five sub-tasks that cover span extraction, classification, and relation extraction, including cross-sentence relations. Given a paragraph from a scientific text:

- For each paragraph of text, identify all spans containing quantities (e.g. 12 kg). Quantities are treated as strings, and are not converted or normalized.
- For each identified Quantity, identify the Unit of Measurement (e.g. kg), if one exists. For each Quantity classify additional value Modifiers (e.g. count, range, approximate, mean,

[4] http://www.qudt.org/

etc.) that apply to the Quantity.

- For each identified Quantity, identify the Measured Entity (e.g. bed inventory) it applies to (if one exists) and mark its span. If an associated Measured Property (e.g. concentration) also exists, identify it and mark its span.
- Identify and mark the span of any Qualifier (e.g. after incubation) that is needed to record additional related context to either validate or understand each identified Quantity.
- Identify relationships between Quantity, Measured Entity, Measured Property, and Qualifier spans using the HasQuantity, HasProperty, and Qualifies relation types.

More detailed definitions can be found be reviewing the MeasEval Annotation Guidelines.[5] We describe each of the elements to be extracted in more detail in the next section.

4 Annotated Data

4.1 Data Model

As shown in Figure 1, the MeasEval annotation model consists of Quantities, MeasuredEntities, MeasuredProperties, and Qualifiers. A Quantity can be either a count or a measurement, with measurements being composed of a Unit and a Value. Values also can have additional attributes such as "isMean", "isApproximate", or "isRange". Quantities can be directly related to a MeasuredEntity, or can be indirectly related to a MeasuredEntity via a MeasuredProperty. Qualifiers provide additional information that is required to interpret the measurement. These include things like the pressure at which a boiling point was observed, or the depth and location where an ocean sample was taken. Since texts may contain different parts of this information, all relationships are optional. A MeasuredEntity can be related to a MeasuredProperty or a Quantity, a MeasuredProperty can be related a Quantity, and a Qualifier can have a relationship to any span.

4.2 Corpus and Annotations

Annotations are drawn from 110 CC-BY licensed articles that have been made previously available by Elsevier Labs. [6] These articles were the basis of a previous SemEval task for SemEval 2017 (Augenstein et al., 2017). These 110 articles are distributed evenly across 10 subject areas.

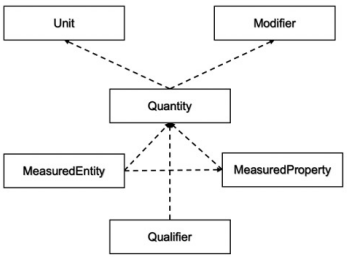

Figure 1: Annotation Model. All relationships are optional.

From these 110 articles, the MeasEval dataset includes 428 paragraphs containing 1663 Quantities. These are split into a training data set of 1164 Quantities (313 paragraphs) and an evaluation set of 499 Quantities (135 paragraphs).

All paragraphs were annotated by at least two annotators, then reviewed and reconciled during an adjudication meeting, often including a third annotator. The MeasEval data release included training data, as well as original annotations from multiple annotators for a 248 Quantity subset of the training data. This was to provide deep information on inter-annotator agreement, and also to allow participants to do their own analysis on how their algorithms perform relative to humans.

The inter-annotator agreement (IAA) shows some variation in interpretation when humans are performing this task. The review process serves to resolve much of the disagreement and to ensure that the data is as consistent as possible given the challenging nature of the task. For this subset of data in this IAA set, Table 1 shows Krippendorff's Alpha values for each class.

Annotation Class	Krippendorff's Alpha
Quantity	0.943
MeasuredProperty	0.641
MeasuredEntity	0.546
Qualifier	0.334
Unit	0.866

Table 1: Krippendorf's Alpha scores for subset of data included in Inter-Annotator Agreement dataset.

4.3 Data Formats

To increase the usability of the data, multiple formats are provided. The MeasEval data includes a

[5] https://github.com/harperco/MeasEval/tree/main/annotationGuidelines

[6] https://github.com/elsevierlabs/OA-STM-Corpus

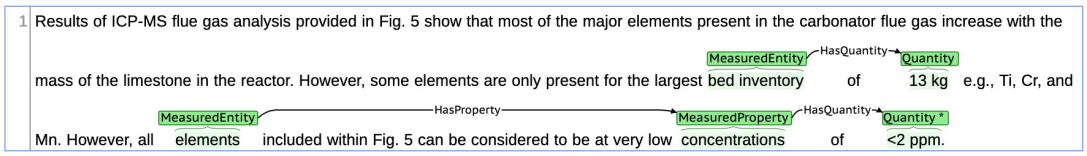

Figure 2: BRAT Example of a Quantity with related annotations.

text file and a set of annotations for each paragraph of scientific text. Annotations are provided in a tab-separated value (.tsv) file format, and in the BRAT annotation format. The BRAT format is for the purpose of visualization and review, but the official data format for the task is the .tsv, which is used for submissions and evaluation. For .tsv and .txt files, there is one file per paragraph of annotated text, and the .tsv file contains all annotations. For the BRAT files, there are one .ann and one .txt file per annotated Quantity.

For example, given the BRAT annotations illustrated in Figure 2, the data will have a raw text file (S0016236113008041-3153.txt), a BRAT annotation file *per Quantity* (S0016236113008041-3153-1.ann, and S0016236113008041-3153-2.ann), and a tab-separated file containing all annotations from each Quantity (S0016236113008041-3153.tsv).

More detail on each of these formats, including examples, as well as all MeasEval training and evaluation data, inter-annotator agreement annotations, and annotation guidelines can be found on the MeasEval Github repository. [7]

5 Evaluation

Evaluation is scored by providing a single SQuAD-style (Rajpurkar et al., 2016) F1 (Overlap) score for each submission, averaged across all nine components of the five subtasks. The 9 components are the Quantity, MeasuredProperty, MeasuredEntity, and Qualifier spans; the Modifier and Unit extensions to Quantity, and the HasQuantity, HasProperty, and Qualifies relationships. The evaluation script also provides a number of other metrics, described below.

In order to effectively evaluate all 9 components of the sub-tasks, it is necessary to first pin all Quantities in a submission to the corresponding Quantities in the gold data. As an example, consider the sentence "The dog weighed 25 pounds, while the average weight of the cats was 9 lbs." We want to avoid crediting correct MeasuredEntities if associated with the wrong Quantity. For example, if a submission listed "dog" as the MeasuredEntity associated with the average weight of 9 lbs, this would be incorrect.

The first pass matches each submission "annotSet" ID to a corresponding Gold Set annotationId, and propagates this matched identifier across all of the data.

From there, the script calculates Precision, Recall, F-measure, and an Exact Match and SQuAD-style F1 (overlap) score. Exact Match and F1 are averaged across the entire submission. Exact Match is a binary value of 0 or 1, while F1 is a token level overlap ratio of submission to gold spans, where tokenization is done using simple white space delimiters. For components that do not include a span, Exact Match and F1 scores are the same. Relations are also scored with a binary Exact Match and F1 score if the relation types match and both endpoints match either exactly or with some overlap.

Any span, unit, modifier, or relationship found in the gold data, but not the submission, or found in the submission, but not the gold data is included as a "penalty row" with a score of 0 in order to sufficiently penalize both false positives and false negatives when averaging scores. This calculation leads to very fine-grained differences in the distribution of scores in the results tables.

Although not used for calculating leaderboard rankings, the evaluation code can also provide all the same scores micro-averaged by scoring component, by subject area, or by paragraph for further analysis of error. Additional documentation as well as the evaluation code itself can be found on the MeasEval GitHub repository. [8]

5.1 Baseline Models

MeasEval also includes two very similar baseline models. Baseline 1 is the best-performing of these, and scores an overall F1 (Overlap) of 0.239 in the evaluation as reported in Tables 2 and 3. Base-

[7] https://github.com/harperco/MeasEval

[8] https://github.com/harperco/MeasEval/tree/main/eval

Team Name	Overall	Quantity	Unit	Modifier	MeasuredEntity	MeasuredProperty	Qualifier
LIORI*	**0.519**	**0.861**	0.722	**0.642**	**0.437**	**0.467**	**0.163**
jarvis@tencent*	0.473	0.855	0.719	0.523	*0.398*	0.437	0.000
zyy_77	*0.448*	0.842	0.697	0.507	0.383	*0.385*	0.000
zz362	0.433	0.821	0.720	0.498	0.344	0.365	0.000
Counts@IITK*	0.432	**0.861**	0.406	0.245	*0.077*	**0.804**	0.614
yorkey	0.399	0.745	0.661	0.314	0.344	0.365	0.000
XMSHI	0.392	0.736	0.624	0.313	0.348	0.353	0.000
CLaC-BP*	0.389	0.855	0.677	*0.546*	0.251	0.318	0.107
clockwise9*	0.369	*0.850*	0.618	0.000	0.327	0.350	0.000
UPB*	0.369	0.742	0.533	0.277	0.331	0.374	0.040
Baseline	0.239	0.827	0.561	0.000	0.053	0.064	0.005
KGP*	0.278	0.787	*0.748*	0.309	0.113	0.012	0.005
Stanford MLab*	0.272	0.818	0.760	0.408	0.000	0.000	0.000
BuckschJ	0.263	0.825	0.695	0.375	0.000	0.000	0.000
CLaC-np*	0.241	0.756	0.495	0.408	0.056	0.006	0.000
FabianW	0.238	0.826	0.624	0.438	0.060	0.045	0.006
ugeijtsv	0.229	0.759	0.582	0.210	0.000	0.000	0.000
Jo	0.212	0.754	0.377	0.291	0.000	0.000	0.000
joe_o123	0.185	0.376	0.383	0.242	0.000	0.000	0.000
SU-NLP	0.001	0.007	0.002	0.000	0.000	0.000	0.000

Table 2: Top result for each team/user, ordered by Overall F1 along with micro-averages for each annotation span, for units, and for modifiers. Team Names marked with * have submitted system information for further analysis and discussion. Top, second, and third place scores per category represented by **bold**, underline, and *italics* respectively.

line 1 use spaCy Named Entity Recognition (NER) models for each of the four classes independently. Unfortunately, some training examples need to be thrown away because spaCy's NER functionality does not support overlapping spans in the same model. Since there is frequently an overlap between MeasEval spans of different types, this necessitates training each annotation type separately, and stripping out edge cases where multiple annotations of the same type intersect.

Baseline 1 generates a deduplicated list of all units in the training data, and checks each Quantity against this list. If there are one or more matches in this comparison, the system returns the "longest last matching" unit, ensuring that cm would be preferred to m in "22 cm" and that s would be preferred to m in "approximately 22 s". The baseline does not attempt the Modifier component, though could be augmented with a set of regular expressions that search the Quantity string for key phrases and symbols, including "approximately", "between", ">", and "∼".

Once the NER models and unit matching are completed, baseline 1 matches Quantities to MeasuredEntities, MeasuredProperties, and Qualifiers using a knockout match algorithm based on proximity. So each MeasuredProperty matches the nearest Quantity, each MeasuredEntity matches the nearest MeasuredProperty or Quantity, and each Qualifier matches the nearest span of any type. Baseline 2 is a variant that does much simpler matching, taking each span in the order they appear in the data. Baseline 2 does not appear in the results tables, but scores an overall F1 (Overlap) of 0.223. The code for both baselines is available in a Jupyter notebook on the MeasEval Github repository. [9]

6 Results and Discussion

During the 21-day evaluation period (January 10 through 31, 2021), 26 CodaLab users submitted a total of 89 submissions, of which 77 passed validation and were successfully scored by the evaluation script. Given the complexity of the task, we opted to allow for five submissions total during the evaluation, although some collaboration between users meant that some teams were able to effectively submit more than five times. We note that submissions did not calculate scores on sub-tasks, thus making it difficult to overly optimize models using just the overall score. The relatively generous submission allowance does not seem to have presented too much of an over-fitting problem, as scores remain relatively low on all tasks, although the collaboration could have given some participants a slight advantage in the rankings.

Table 2 shows the top submission from each of

[9] https://github.com/harperco/MeasEval/blob/main/baselines/first-baseline.ipynb

the 19 teams that submitted successfully, as well as the top performing baseline. 10 of the 19 exceed the benchmark of the baseline spaCy model. In addition to the overall F1 scores, Table 2 shows micro-averaged F1 across the four annotation spans as well as Units and Modifiers. Table 3 provides this same breakdown for each of the three relationship types. Team names marked with an asterisk (*) represent teams which have either submitted system description papers or responded to a request for system information.

The overall top-performing model was at least tied for top performance in five out of 6 of the component scores in Table 2, but interestingly, the second best and third best performing models varied across scoring component. Models that did particularly well at Quantities, Units, or Modifiers, may have had their overall performance reduced by lower performance at the MeasuredEntity and MeasuredProperty spans.

Table 3 shows the scores for the Relation Extraction subtasks: HasQuantity, HasProperty, and Qualifies. These largely align with the annotation span components of the scoring which they are dependent on. In both Table 2 and Table 3 it is worth noting that only 7 teams attempted extraction of Qualifiers and the Qualifies relation, as these were the most difficult aspects of the task.

Team Name	HasQuantity	HasProperty	Qualifies
LIORI*	**0.482**	**0.318**	**0.092**
jarvis@tencent*	0.424	0.257	0.000
zyy_77	*0.387*	0.229	0.000
zz362	0.375	0.203	0.000
Counts@IITK*	0.311	0.183	0.064
yorkey	0.375	0.203	0.000
XMSHI	0.373	0.199	0.000
CLaC-BP*	0.308	0.147	*0.058*
clockwise9*	0.366	0.167	0.000
UPB*	0.350	*0.242*	0.019
Baseline	0.075	0.007	0.000
KGP*	0.076	0.006	0.000
Stanford MLab*	0.000	0.000	0.000
BuckschJ	0.000	0.000	0.000
CLaC-np*	0.028	0.000	0.000
FabianW	0.037	0.007	0.000
ugeijtsv	0.000	0.000	0.000
Jo	0.000	0.000	0.000
joe_o123	0.000	0.000	0.000
SU-NLP	0.000	0.000	0.000

Table 3: Top result for each team/user for the Relation Extraction components of the score. Team Names marked with * have submitted system information for further analysis and discussion. Top, second, and third place scores per category represented by **bold**, underline, and *italics* respectively.

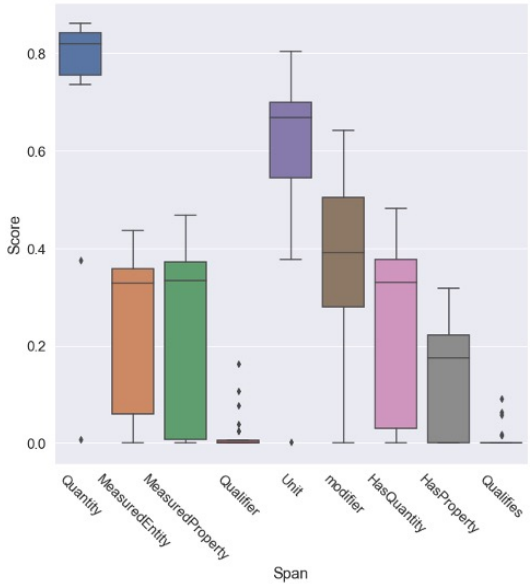

Figure 3: Visualization of average scores for each scoring component across top score for all participants.

Figure 3 provides a visualization of the distribution of scores for each scoring component from Tables 2 and 3. From this, it is clear that Quantity and Unit are the easiest aspects of the shared task, which makes intuitive sense. The relatively high scores for Modifier is also of interest, as these are the components of the extraction that capture uncertainty and variance in value, which is an important part of the task and not one that we expected to see handled as well as it was. This clearly demonstrates that the various Quantity contextualization subtasks are far more difficult and more work is needed in how best to handle the extraction of related MeasuredEntities, MeasuredProperties, and Qualifiers.

Figure 4 provides a visualization of the distribution of scores for 9 of the 10 subject areas in the corpus. The mathematics subject area has been omitted from this graphic due to under-representation in both the training and evaluation datasets.

6.1 Impact of Duplicates

As noted previously, the MeasEval evaluation algorithm was designed toward lenience, and as a result sometimes inflates scores if multiple submission annotations match a single entry in the gold data. This was done to allow submissions to get credit for submitting multiple Quantity annotations that partially matched a single gold data span as well as to generally not penalize systems that might make multiple predictions pinned to the same numeric

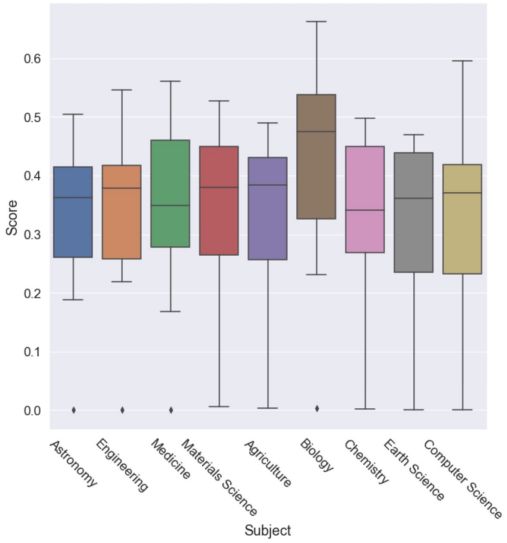

Figure 4: Visualization of average scores for each subject area across top score for all participants.

value.

However, allowing duplicates can in some cases result in inflated scores. This is especially evident in cases where submissions contained entries that duplicated entire annotations sets. For a small number of submissions that exhibited annotation set level duplicates, a post-processing routine removed all set level duplicates before final evaluation, resulting in the scores in Table 2 and Table 3.

Additionally, Quantity-level duplicates can also potentially inflate Quantity scores, but should have a neutral impact on other components of scoring. For example, if a system identified the same Quantity two times, but found a different MeasuredEntity for each occurrence, the submission will score extra points associated with the Quantity, and potentially the Unit and Modifiers if those are also correct, but will only get points for the correct MeasuredEntity while being penalized for the incorrect MeasuredEntity. An ablation analysis was performed for the eight submissions covered by system papers, assessing the impact of these duplicates on the Overall F1 (Overlap) metric.

Table 4 gives the extent of duplication for these submissions, the initial overall score from Table 3, the overall score with exact quantity duplicates removed, and the overall score with both exact and overlapping duplicates removed. For example, if the gold data includes the Quantity "approximately 23 mm", and a submissions included annotation sets with both "23 mm" and "approximately 23 mm", the exact match duplicate removal would not drop either score, whereas the overlapping match score would drop whichever occurred last in the submission, whether or not it is the correct answer.

Team Name	F1	Count Exact / Overlap	F1 w/out Exact	F1 w/out Any Duplicates
LIORI	0.519	125 / 32	0.499	0.487
jarvis@tencent	0.473	0 / 11	0.473	0.470
Counts@IITK	0.432	0 / 0	0.432	0.432
CLaC-BP	0.389	0 / 0	0.389	0.389
UPB	0.369	0 / 1	0.369	0.369
KGP	0.278	0 / 0	0.278	0.278
Stanford MLab	0.272	0 / 0	0.272	0.272
CLaC-np	0.241	55 / 0	0.231	0.231

Table 4: Ablation analysis of duplicates and Overall F1 (Overlap) score for each of the eight Teams with System Papers.

The general downward trajectory seen while removing duplicates that are not at the set level is informative. Partly this is due to declining Quantity score from duplicate removal, but some effect is attributable to the possibility that deduplication deletes a correct MeasuredProperty or MeasuredEntity and leaves an incorrect one, given that they may include different values. The ablation analysis simply removes all but the first occurrence, so there is no control over whether removed values are a closer match to the gold data.

6.2 Multiple Hypotheses Hypothesis

The results of de-duplication analysis, the relatively low inter-annotator agreement scores, and deeper consideration of the annotation guidelines present an interesting hypothesis. It could be that the different interpretations of the context of a measurement are not automatically right or wrong. It could be that different interpretations are useful in different downstream applications. While it is out of scope in for this task description, future work may look more closely at categorization of the areas where annotators disagreed and systems produced multiple interpretations, to see if there is alignment in the differences and whether there are patterns to the variance.

7 Summary of Participating Systems

The MeasEval track at SemEval-2021 received nine system description paper submissions, eight of which are represented in the analysis in Table 4. A ninth paper formulated a new task from the MeasEval dataset focusing on just the relation extraction

part of the problem. One system only attempted the Quantity, Unit, and Modifier parts of the task, while another did not submit any Qualifier spans or Qualifies relationships. Four of the nine systems have released code or models.

There are several points of similarity between the eight main submissions. All but one of the systems are based on the BERT architecture (Devlin et al., 2018) or a derivative, such as SciBERT (Beltagy et al., 2019), BioBERT (Lee et al., 2019), or RoBERTa (Liu et al., 2019). All but one used a pipeline architecture, starting with Quantity extraction. All but one used sequence tagging with a BIO encoding scheme, and four followed the sequence tagger by a Conditional Random Fields (CRF) model to assemble tokens into spans and improve accuracy over simple token-level classifiers. Unit and Modifier extraction was either done using a character-level BiLSTMs, another BERT model, or a rule-based approach. Finally, it was common to see MeasuredEntity, MeasuredProperty, and Qualifier, and sometimes the relation extraction components, stacked together in a multi-task sequence tagging model as a final stage, taking both the original sentences and Quantity spans as input. One system diverged from the sequence tagging tagging approach and used templated question answering techniques to handle the relation extraction along with related spans. Table 5 provides a high-level summary of frequency of architectures and techniques in use by more than one system.

Technique / Model	Submission Count
BERT	3
BioBERT	1
SciBERT	3
RoBERTa	1
CRF	4
BiLSTM Units / Mods	3
Rule-based Units / Mods	3
Dict-based Units / Mods	2
Question Answering	1
Sequence Tagging	7

Table 5: Summary of techniques and architectures used in MeasEval System Description Submissions.

7.1 System Specifics

Davletov et al. (2021) (LIORI), achieve their state-of-the-art performance using pre-trained models RoBERTa (Liu et al., 2019) and LUKE (Yamada et al., 2020). They use LUKE to fine-tune an NER model for Quantity extraction, and a RoBERTa-based multi-task model for all other spans. Modifiers are predicted as Quantity-types. All other spans, including units, are extracted using Question Answering style sequence tagging (start/end logits) without question prompts for each annotation type queried for each extracted Quantity.

This sequential ensemble approach of Quantity-detection followed by either "all-in-one-multi-task" extraction or a staged approach to one or more of the other subtasks proved very common among the top-performing systems.

Cao et al. (2021) (jarvis@tencent), do initial Quantity extraction with an ensemble of a Pointer Net (Vinyals et al., 2015) and a CRF. They use a BERT-based classifier for Modifier tagging and a rule-based system for Units, and then use relation-specific taggers with the same architecgure as the Quantity-tagger for all other task components.

Gangwar et al. (2021) (Counts@IITK) similarly tag Quantities witha SciBERT sequence tagger and a CRF model and SciBERT for Modifiers, but use a Character based bidirectional LSTM for Unit tagging. They then encode the Quantity into SciBERT input when tagging MeasuredEntity and HasQuantity, and iteratively mark new spans in the input when tagging then next sub-task, using a rule-base for assembling the necessary relationships. Their performance on Quantity, Unit, and Modifier was near the top performing, but they struggled with MeasuredProperty and HasProperty.

Therien et al. (2021) (CLaC-BP) use SciBERT in a token-level multi-class classifier across all span classes. This is an interesting approach, given the opportunity for joint inference between the various types of spans. However, it penalizes them in that each token in their model can only be one class, while there are cases when a Quantity and MeasuredEntity from one set may be part of, e.g. a Qualifier in another. Quantity spans are then fed to another SciBERT model for Modifier typing, and rule-based systems are used for Units and for the Relations between spans.

Avram et al. (2021) (UPB) use RoBERTa along with a CRF for Quantity extraction. They also tested SciBERT and BERT. They achieved their best results on their dev subset with SciBERT, but their best results on evaluation set came from RoBERTa. They use a BiLSTM to extract Units and classifier Modifiers, and then use a templated Question Answering system as a joint entity and relation extraction system for the remaining subtasks. Unlike LIORI, who did not use prefixes or

suffixes in their question templates, UPB asks more natural language questions of the form "What is the measured property of the quantity ___?"

Karia et al. (2021) (KGP) also use BioBERT after testing various BERT-based pre-trained models, but depart from many of the other submissions by using a binary classifier rather than BIO tags and CRF layers for Quantity sequence tagging. Modifiers and Units are handled using keywords and dictionary matching, while they use a multi-task BERT model for the remaining components, first finding MeasuredEntity based on the Quantity predictions, then fusing these results for the remaining spans and relations. They also continued refining their approach into the post-evaluation phase, and reported improving their score from an Overall F1 (Overlap) of 0.278 to 0.456.

Liu et al. (2021) (Stanford MLab) only tackle the Quantity, Unit, and Modifier subtasks. Notably, they report building their system for these components from inception to submission in under 48 hours. They use BERT-large for IOB sequence tagging for Quantities, use a similar IO sequence tagging scheme on Quantities to tag Units, and a multi-class classifier to classify Quantities to the appropriate Modifiers. Their system performs well on all subtasks they attempted, even scoring second place overall for Units.

Lathiff et al. (2021) (CLaC-np) diverge from other submissions in their approach. They preprocess their text using GATE and ANNIE, and use custom rules to further clean up tokenization. They treated Stanford Core Dependency Parse trees as graphs to extract subgraphs starting each path query from the CD tokens to identify MeasureEntities, MeasuredProperties and Qualifiers with the use of Graph CNN. They relied on the models from CLaC-BP to map from their tokens to annotation spans for each type in assembling their final submission.

Finally, not shown in Table 1 is Veyseh et al. (2021). They formulated their own task based on the MeasEval data. Although they did not submit a solution during the evaluation period, they have submitted a system description paper describing a novel approach to relation extraction, which they have evaluated on MeasEval sub-task 5. Using our Gold Quantity, MeasuredEntity, MeasuredProperty, and Qualifier spans as input (without annotation sets), they compared their approach to two other baseline models. They encode contextual embeddings, positional embeddings, and entity types for each annotation span, and perform dependency path reasoning along with an "Information Bottleneck" regularization technique to complete their Relation Extraction task.

8 Conclusion

In this paper, we present the design, the data, the evaluation the process, the results, and the systems for MeasEval at SemEval 2021. The shared task is challenging, partly due to the relatively small training data, and partly due to the inter-relationships between many different components of the task. Quantity and Unit identification, and to a lesser extent Modifier typing, appear to be the simplest parts of the task based on average performance, with one participating system building their end-to-end pipeline for these components in under 48 hours. The contextual elements MeasuredEntity, MeasuredProperty, and Qualifier, and their relationships, are far more difficult, which is not surprising given that these are subject to more human annotator disagreement. The challenge of context is especially pronounced in the Qualifier span and Qualifies relationship.

Common components shown in Table 5 include the BERT family of pre-trained neural language models, CRF models, BiLSTMs, and rule-based components. In general, the task does not appear to require whole new novel models and architectures, but rather pipelines and cascading ensembles stitching together various existing methods. There is still room for improvement on this task, and whether progress will come from novel models or creative applications of existing techniques remains to be seen. Work also remains to be done in applying the entities and relationships extracted for this task to the larger end goal of scientific knowledge graph construction and related downstream applications. Future work could be done to further analyze areas of error and disagreement in these annotations, and to investigate entity linking across Quantity, MeasuredEntity, and MeasuredProperty annotation spans to various measurement ontologies and to domain-specific entity and property ontologies.

Acknowledgments

The authors would like to thank Darin McBeath and Pierre-Yves Vandenbussche for help with annotations and annotation rules. We also thank Elsevier's Discovery Lab team for their feedback on this work.

References

Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. 2017. SemEval 2017 task 10: ScienceIE - extracting keyphrases and relations from scientific publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 546–555, Vancouver, Canada. Association for Computational Linguistics.

Andrei-marius Avram, George-eduard Zaharia, Dumitru-clementin Cercel, and Mihai Dascalu. 2021. UPB at SemEval-2021 Task 8 : Extracting Semantic Information on Measurements as Multi-Turn Question Answering. In *Proceedings of the Fifteenth Workshop on Semantic Evaluation (SemEval-2021)*, 2.

Payam Barnaghi, Wei Wang, Lijun Dong, and Chonggang Wang. 2013. A linked-data model for semantic sensor streams. *Proceedings - 2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing, GreenCom-iThings-CPSCom 2013*, pages 468–475.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. pages 3613–3618.

Melanie Bergmann, Mine B Tekman, and L. Gutow. 2017. *LITTERBASE: An Online Portal for Marine Litter and Microplastics and Their Implications for Marine Life*, pages 106–107.

Jiarun Cao, Yuejia Xiang, Yunyan Zhang, Zhiyuan Qi, and Xi Chen. 2021. CONNER : A Cascade Count and Measurement Extraction Tool for Scientific Discourse. In *Proceedings of the Fifteenth Workshop on Semantic Evaluation (SemEval-2021)*.

Dmitry Davidov and Ari Rappoport. 2010. Extraction and approximation of numerical attributes from the Web. *ACL 2010 - 48th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, (July):1308–1317.

Adis Davletov, Denis Gordeev, Nikolay Arefyev, and Emil Davletov. 2021. LIORI at SemEval-2021 Task 8: Ask Transformer for measurements. In *Proceedings of the Fifteenth Workshop on Semantic Evaluation (SemEval-2021)*.

Helena F. Deus, Corey A. Harper, Darin McBeath, and Ron Daniel. 2017. Combining pattern matching with word embeddings for the extraction of experimental variables from scientific literature. *2017 IEEE International Conference on Big Data (Big Data)*, pages 4287–4292.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Chau Do and Eric J. Pauwels. 2013. Using MathML to represent units of measurement for improved ontology alignment. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7961 LNAI:310–325.

Liran Einav and Jonathan Levin. 2014. The data revolution and economic analysis. *Innovation Policy and the Economy*, 14:1–24.

Luca Foppiano, Thaer M Dieb, Akira Suzuki, and Masashi Ishii. 2019a. Proposal for Automatic Extraction Framework of Superconductors Related Information from Scientific Literature Proposal for Automatic Extraction Framework of Superconductors Related Information from Scientific Literature. (June):0–5.

Luca Foppiano, Laurent Romary, Masashi Ishii, and Mikiko Tanifuji. 2019b. Automatic identification and normalisation of physical measurements in scientific literature. *Proceedings of the ACM Symposium on Document Engineering, DocEng 2019*, pages 0–4.

Akash Gangwar, Sabhay Jain, Shubham Sourav, and Ashutosh Modi. 2021. Counts @ IITK at SemEval-2021 Task 8 : SciBERT Based Entity And Semantic Relation Extraction For Scientific Data. In *Proceedings of the Fifteenth Workshop on Semantic Evaluation (SemEval-2021)*.

Abhijeet Gupta, Gemma Boleda, Marco Baroni, and Sebastian Padó. 2015. Distributional vectors encode referential attributes. *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, (September):12–21.

Tianyong Hao, Hongfang Liu, and Chunhua Weng. 2016. Valx: A system for extracting and structuring numeric lab test comparison statements from text. *Methods of information in medicine*, 55.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, José Emilio Labra Gayo, Sabrina Kirrane, Sebastian Neumaier, Axel Polleres, Roberto Navigli, Axel-Cyrille Ngonga Ngomo, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. 2021. Knowledge graphs.

Kyle Hundman and Chris A Maamann. 2017. Measurement Context Ex-traction from Text: Discovering Opportunities and Gaps in Earth Science. 8:pages.

Joint Committee for Guides in Metrology. 2012. *International vocabulary of metrology. Basic and general concepts and associated terms.*, 3rd ed edition.

Jocelyn Kaiser. 2018. Plan to replicate 50 high-impact cancer papers shrinks to just 18. *Science*.

Ramar Kaladevi, Gurunathan Geetha, and P. Narayanasamy. 2016. Ontological based interoperability and integration framework for heterogeneous weather systems. *Revista Tecnica de la Facultad de Ingenieria Universidad del Zulia*, 39(1):185–192.

Neel Karia, Ayush Kaushal, and Faraaz Rahman Mallick. 2021. KGP at SemEval-2021 Task 8 : Leveraging Multi-Staged Language Models for Extracting Measurements, their Attributes and Relations. In *Proceedings of the Fifteenth Workshop on Semantic Evaluation (SemEval-2021)*.

Natalia Konstantinova. 2014. Review of relation extraction methods: What is new out there? *Communications in Computer and Information Science*, 436:15–28.

Bhushan Kotnis and Alberto Garcıa-Duran. 2019. Learning Numerical Attributes in Knowledge Bases. *Automated Knowledge Base Construction (2019)*.

Nihatha Lathiff, Pavel Khloponin, and Sabine Bergler. 2021. CLaC-np at SemEval-2021 Task 8 : Dependency DGCNN. In *Proceedings of the Fifteenth Workshop on Semantic Evaluation (SemEval-2021)*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. pages 1–8.

Patrick Liu, Niveditha S Iyer, Erik Rozi, and Ethan A Chi. 2021. Stanford MLab at SemEval-2021 Task 8 : 48 Hours Is All You Need. In *Proceedings of the Fifteenth Workshop on Semantic Evaluation (SemEval-2021)*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv*, (1).

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. (ii).

H. Rijgersberg, M. Wigham, and J. L. Top. 2011. How semantics can improve engineering processes: A case of units of measure and quantities. *Advanced Engineering Informatics*, 25(2):276–287.

Hajo Rijgersberg, Mark Van Assem, and Jan Top. 2013. Ontology of units of measure and related concepts. *Semantic Web*, 4(1):3–13.

J. Sprague. 2006. The Zebrafish Information Network: the zebrafish model organism database. *Nucleic Acids Research*, 34(90001):D581–D585.

Markus D. Steinberg, Sirko Schindler, and Jan Martin Keil. 2017. Use cases and suitability metrics for unit ontologies. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10161 LNCS:40–54.

Benjamin Therien, Parsa Bagherzadeh, and Sabine Bergler. 2021. CLaC-BP at SemEval-2021 Task 8 : SciBERT Plus Rules for MeasEval. In *Proceedings of the Fifteenth Workshop on Semantic Evaluation (SemEval-2021)*.

Shreejoy J. Tripathy, Judith Savitskaya, Shawn D. Burton, Nathaniel N. Urban, and Richard C. Gerkin. 2014. NeuroElectro: A window to the world's neuron electrophysiology data. *Frontiers in Neuroinformatics*, 8(APR):1–11.

Amir Pouran Ben Veyseh, Franck Dernoncourt, and Thien Huu Nguyen. 2021. DPR at SemEval-2021 Task 8: Dynamic Path Reasoning for Measurement Relation Extraction. In *Proceedings of the Fifteenth Workshop on Semantic Evaluation (SemEval-2021)*.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. *Advances in Neural Information Processing Systems*, 2015-January:2692–2700.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. *arXiv*, pages 6442–6454.