Measurement Extraction with Natural Language Processing: A Review

Jan Göpfert[1,2,*] and **Patrick Kuckertz**[1] and **Jann M. Weinand**[1] and **Leander Kotzur**[1] and **Detlef Stolten**[1,2]
[1]Institute of Energy and Climate Research, Techno-economic Systems Analysis (IEK-3), Forschungszentrum Jülich, 52425 Jülich, Germany
[2]Chair for Fuel Cells, RWTH Aachen University, c/o IEK-3, Forschungszentrum Jülich, 52425 Jülich, Germany
*j.goepfert@fz-juelich.de

Abstract

Quantitative data is important in many domains. Information extraction methods draw structured data from documents. However, the extraction of quantities and their contexts has received little attention in the history of information extraction. In this review, an overview of prior work on measurement extraction is presented. We describe different approaches to measurement extraction and outline the challenges posed by this task. The review concludes with an outline of potential future research. Research strains in measurement extraction tend to be isolated and lack a common terminology. Improvements in numerical reasoning, more extensive datasets, and the consideration of wider contexts may lead to significant improvements in measurement extraction.

1 Introduction

Humanity is accumulating more and more knowledge at an ever faster pace. Cast into large amounts of documents, relevant knowledge is no longer graspable by a few individuals. Information extraction (IE) is a task in *natural language processing (NLP)* and assists in managing the amount of information hidden in documents by automatically extracting and organizing information from semi- and unstructured sources (e.g., populating a database from information conveyed in natural language). In the early 1990s, the *Message Understanding Conferences (MUC)* fostered research in IE trough challenges in *template filling* (Grishman, 2019). Later MUCs and the *Automatic Content Extraction (ACE)* program split IE into several sub-challenges, helping *named entity recognition (NER)*, *relation extraction*, *event extraction* and *coreference resolution* to emerge as individual research subjects (Grishman and Sundheim, 1996; Doddington et al., 2004; Weischedel and Boschee, 2018; Grishman, 2019). Today, IE is applied in many domains, such as biomedicine (Wang et al., 2018), chemistry and materials science (Kononova et al., 2021). However, measurements and their contexts have received little attention in the history of IE (Hundman and Mattmann, 2017; Kang and Kayaalp, 2013; Alonso and Sellam, 2018; Lamm et al., 2018a; Roy et al., 2015).

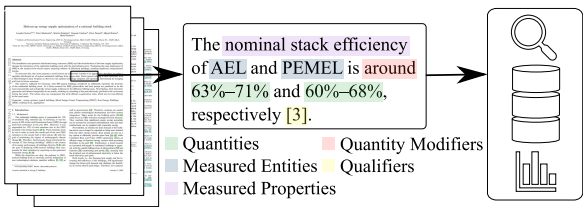

Figure 1: Measurement extraction is the extraction of quantities and related information.

Numbers form a cornerstone of our society, on which science, engineering, trade and much more is built. Numerical reasoning is therefore an essential, albeit underexplored, problem in NLP (Thawani et al., 2021b), the addressing of which seems to even enhance the general literacy of language models (Thawani et al., 2021a). The task of measurement extraction is to identify quantities and related information in texts, tables and figures. In this review, we focus on measurement extraction from text (cf. Figure 1). When specifying measurements, the transition from natural to mathematical language is seamless, making measurement extraction a special task within NLP. The variety of relevant problems to which measurement extraction is applied further highlights its importance. Research on measurement extraction is focused on highly quantitative domains, in particular clinical, biomedical, chemistry, and materials research. Accordingly, most systems are applied to scientific publications and clinical documents. Within the medical domain, more frequent applications include the extraction of eligibility criteria from clinical trials (Hao et al., 2016; Kang et al., 2017), the extraction of lab test information (Kang and

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 2191–2215
December 7-11, 2022

Kayaalp, 2013; Liu et al., 2017), the extraction of measurements from narrative radiology reports (Sevenster et al., 2013, 2015b,a; Bozkurt et al., 2019) and the extraction of the left ventricular ejection fractions of hearts (Kim et al., 2017c; Garvin et al., 2012; Kim et al., 2013; Meystre et al., 2017). Within chemistry and materials science, information on experiments (Hawizy et al., 2011; Deus et al., 2017; Friedrich et al., 2020), materials synthesis (Kim et al., 2017b; Kononova et al., 2019) and nanoscience (Xiao et al., 2013; Jones et al., 2014; Dieb et al., 2012, 2015, 2014) is extracted. Beyond that, application domains include patent analysis (Agatonovic et al., 2008; Aras et al., 2014), automated chart generation (Lamm et al., 2018b), fact-checking (Vlachos and Riedel, 2015), earth science (Hundman and Mattmann, 2017; Petersen et al., 2021), engineering design (Opasjumruskit et al., 2019a; Hsiao et al., 2020), information retrieval (Maiya et al., 2015; Ho et al., 2020; Li et al., 2021), automated compliance checking (Zhang and El-Gohary, 2016), and more. However, research strains tend to be isolated within these domains (see Figure B1), indicating a lack of an overview.

In this paper, we define measurement extraction (Section 2) and survey prior research (Section 3). Subsequently, we highlight special challenges (Section 4) and provide several recommendations for future research (Section 5). Section 6 describes the limitations of this review. To the best of our knowledge, the present review is the first that focuses on measurement extraction.

2 Task definition

The language around measurement extraction lacks standardization (see Section 5). Likewise, the scope of measurement extraction is not well-defined. We define it as follows:

Quantity Extraction is the task of identifying *quantities*. A quantity (e.g., *'1 kg'*) is composed of a *numeric value* and, if applicable, a *unit*. The meaning of a quantity is often altered by *modifiers* such as *'average'*, *'approx.'* or *'above'*. Modifiers adjacent to numeric values are sometimes included in the quantity spans (Friedrich et al., 2020; Harper et al., 2021). A quantity might be given as a range, enumeration, with an uncertainty specification, or all together. Numeric values might be expressed as numeric numbers (e.g., *'27'*), alphabetic numbers (e.g., *'twenty-seven'*), combinations (e.g., *'2 million'*), imprecise quantities (e.g., *'a couple'*; cf. Hanauer et al., 2019) or constants (e.g., *'room temperature'* or *'speed of light'*). Within a quantity span, the unit might be identified. Units are often abbreviated according to their symbol (e.g., *'J'* for Joule). Note that nouns, such as in *'9 family houses'*, are sometimes considered units (Roy et al., 2015). Quantities can be normalized to base SI units. As some unit symbols are ambiguous, the kind of quantity might be identified first (e.g., length for *'1 µm'*). Furthermore, the notions of change (e.g., *'decreased'*) might be extracted for quantities that are given relative to another quantity (e.g., in *"the GDP decreased by 4.6 %"*). The boundary between quantities and equations is fuzzy. Hence, formulaic expressions are considered to differing extents.

Measurement Extraction adds to the identification of quantities by extracting their related *measured properties* and *measured entities* (cf. Figure 1). A measured property might be given implicitly. Measurement extraction can be generic or simplified by only targeting specific measured entities and properties (e.g., if particle sizes should be identified, only length units must be considered). Furthermore, additional *qualifiers* such as constraints, measuring methods or references that qualify a quantitative statement might be extracted. Measured entities, properties, units and relevant context might be disambiguated against a knowledge base (that is, entity linking).

Related tasks that involve numerical reasoning besides measurement extraction from other modalities are, amongst others, product attribute value extraction (Dong et al., 2020), equation parsing (Roy et al., 2016), solving math word problems (Zhang et al., 2020a), quantity entailment (Roy et al., 2015), number sense disambiguation (Chen et al., 2018), numeral attachment (Chen et al., 2019a), and masked measurement prediction (Spokoyny et al., 2022) (see Appendix A). The interested reader is directed to the surveys of Thawani et al. (2021b) and Yoshida and Kita (2021), which provide extensive overviews of various NLP tasks involving numeracy.

3 Prior work on measurement extraction

Various systems for measurement extraction have been proposed. The first research efforts focusing on measurement extraction date back to at least 2006 (Moriceau, 2006). We identified 80 publi-

cations describing one or more systems for measurement extraction. This section summarizes their approaches according to the following subtasks:

- Pre-processing (Section 3.1)
- Identification of quantities, measured entities, properties and qualifiers (Section 3.2)
- Identification of units (Section 3.3)
- Quantity modifier extraction (Section 3.4)
- Relation extraction (Section 3.5)
- Post-processing (Section 3.6)

Tabular overviews of the methods (Table B2 and B3) and scopes of the systems (Table B1), as well as a citation graph of the corresponding publications (Figure B1) are given in the Appendix B.

Varying scopes. Most systems do not cover all subtasks and the respective concept types of the general pipeline depicted above, which fails to reflect the large variations in their scopes. Only a few systems cover the identification of measured entities, properties, further context, and their relations in addition to quantity extraction (see Table B1). Many of those are submissions to MeasEval (task 8 at SemEval 2021; Harper et al., 2021). Frequently, the other systems do not distinguish between measured entities and properties. The rule-based (symbolic) systems tend to have a narrower scope than the learning-based systems; that is, instead of identifying measurement concepts generically, only particular concepts, which are specific to the domain and use case, are identified. Covering only a small set of concepts facilitates normalization and entity linking. In fact, with symbolic approaches, this information is often already evident from the matching patterns. Many systems do not approach the extraction of quantity modifiers and qualifiers. All systems that approach quantity modifier extraction only consider a small set of modifier classes. Only a few systems consider the notions of change (e.g., *'increased'*) for relative quantities (Moriceau, 2006; Roy et al., 2015; Lamm et al., 2018b). In MeasEval, phrases that indicate change are regarded as qualifiers.[1] Only a few articles explicitly state that co-references (Mykowiecka et al., 2009; Roy et al., 2015; Ho et al., 2022) and negations are considered. (Mykowiecka et al., 2009; Yim et al., 2016; Zhang and El-Gohary, 2016; Kang et al., 2017).

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

3.1 Pre-processing

Pre-processing regularly involves optical character recognition, PDF parsing, correcting misspellings and parsing errors, document section and sentence boundary detection, filtering, text normalization, and tokenization. Normalization can include the conversion of alphabetic numbers into numeric numbers and the unification of punctuation, special symbols, digit delimiters, and interchangeably-used characters (Hao et al., 2016; Kang et al., 2017; Swain and Cole, 2016). Karia et al. (2021) found replacing all numerals with '0' to increase quantity identification performance. Madaan et al. (2016) exclude sentences that match change words from a gazetteer. **Custom tokenization rules** can improve the performance, as quantities often include special symbols. For example, numeric values are prevented from being split at their decimal separator (Zhang and El-Gohary, 2016; Lathiff et al., 2021) and separated from adjacent mathematical symbols (Lathiff et al., 2021) and units (Swain and Cole, 2016; Foppiano et al., 2019b,b; Therien et al., 2021). Nevertheless, the subword tokenization of BERT-like encoders will split numbers that are out of vocabulary into multiple tokens (Thawani et al., 2021b; Therien et al., 2021). Therefore, Loukas et al. (2022) detect numbers during pre-processing using regular expressions and experimented with replacing numbers by a [NUM] pseudo-token and special tokens mimicking their shape [X.XX]. Both approaches yield improvements, with the latter being superior to the former. This is possibly because using [NUM] tokens prohibits the models from considering the magnitude when numerical reasoning is required. Finally, some applications require special pre-processing routines such as patient anonymization (Mykowiecka et al., 2009).

3.2 Identification of quantities, measured entities, properties, and qualifiers

Quantity extraction is typically framed as a span identification task, as quantities are rarely given implicitly and the unit is in most cases adjacent to the value. In fact, NER tag sets have long included percentages, monetary expressions (Chinchor, 1998; Grishman and Sundheim, 1995) and quantities (Weischedel et al., 2013). Also, the extraction of measured entities, properties, qualifiers, and units are often framed as span identification tasks.

3.2.1 Rule-based approaches

Whereas machine learning systems learn to solve a task based on exemplary data, rule-based systems employ the knowledge of domain experts who define patterns and rules to solve a task. As such, rule-based approaches are predominated by combinations of rules, patterns and keyword-, gazetteer-, ontology- or dictionary-matching[2]. Patterns are defined using regular expressions, finite-state automata or grammars in frameworks like GATE (Cunningham et al., 2013). Besides string matching, patterns often involve syntactic rules based on part-of-speech (POS) tags. The extraction of quantities and units is sometimes supported by existing quantity, unit or temporal expression taggers (Liu et al., 2017; Madaan et al., 2016). Analogously, existing NER taggers can support the extraction of measured entities (Hawizy et al., 2011; Madaan et al., 2016). Ontology-based approaches for measurement extraction construct gazetteers from ontology terms rather than to extensively exploit their semantic structure and rules (Xiao et al., 2013; Jones et al., 2014). Combining many of the aforementioned approaches, Maiya et al. (2015), for example, use multiple regular expressions to extract numeric values, including the sign, uncertainty and powers of ten. Units are identified using a unit ontology and rules that support multiples and sub-multiples, as well as derived units. Measured properties are extracted using syntactic rules on POS tags. The POS tag set is extended by an additional tag for mathematical symbols of equivalence and one or two character symbols in order to match, i.a., Greek letters.

3.2.2 Learning-based approaches

For systems targeting scientific publications and diverse web sources, there is a trend towards machine learning systems. In clinical systems, this trend is not observable. It might be reasoned that medical applications require higher levels of traceability and favor precision over recall. Tailored rule-based systems can indeed yield very high levels of precision (cf. Table 6 in Liu et al., 2021b). Patterson et al. (2017), for example, extract heart function measurements from echocardiogram reports using rules and dictionary-matching and reach an average F1 score and precision of 86.4 and 96.2, respectively. Certain components of rule-based systems can be easily applied to machine learning systems making a hybrid approach a potentially effective option (Kang and Kayaalp, 2013). Many of the learning-based systems discussed below are in fact hybrid systems relying on rules for one or more subtasks.

Sequence labeling and extractive question answering. Learning-based approaches mostly cast the span identification tasks as sequence labeling problems using an IOB tagging scheme. In accordance with IE in general, Conditional Random Field (CRF) models (Lafferty et al., 2001), Bidirectional Long Short-Term Memory (BiLSTM) models (Huang et al., 2015), and transformer-based models (Vaswani et al., 2017), in particular BERT-based models (Devlin et al., 2019), have been frequently applied. A popular **CRF**-based system is Grobid-quantities (Foppiano et al., 2019b), which identifies and normalizes physical measurements in scientific and technical documents. It uses multiple CRF taggers: the first model identifies quantity spans and distinguishes them by their type (viz. value, list, base, range, least, and most). Subsequently, the units and values sub-models apply more fine-grained labels. According to the most recent evaluation, the quantity, unit and value model (now using BiLSTM+CRF) yield F1 scores of 88.10, 98.45 and 98.57, respectively[3]. The CRF-only setup achieves almost equal results. The extraction of measured entities and properties (which are not distinguished) is an experimental feature. Grobid-quantities was used in several other works, which extended the system to detect different measurement contexts (Hundman and Mattmann, 2017; Foppiano et al., 2019a; Petersen et al., 2021). For **BiLSTM** and **transformer** models, a CRF layer is often stacked on top, the benefits of which cannot be formulated in general terms (Schweter and Akbik, 2021; Loukas et al., 2022). However, some empirical evidence suggests that, when using subword tokenization, adding a CRF layer improves the performance in measurement extraction-related sequence labeling tasks (Panapitiya et al., 2021; Loukas et al., 2022). Varying scopes and evaluation criteria render a quantitative comparison of different approaches across multiple publications inadequate. However, ablation studies of individual publications suggest that BiLSTM and transformer models outperform CRF models in measurement

[2]For brevity, in the following the term dictionary-matching is sometimes used regardless of the sources of external knowledge (i.e., gazetteers, ontologies, or dictionaries).

[3]https://github.com/kermitt2/grobid-quantities

extraction (Friedrich et al., 2020; Liu et al., 2021b). There is inconsistent evidence as to whether BERT-based models are superior to BiLSTMs in measurement extraction (Friedrich et al., 2020; Loukas et al., 2022). Often, BERT-based models are only compared to each other, with different results as to which is superior (Avram et al., 2021; Panapitiya et al., 2021; Therien et al., 2021; Karia et al., 2021; Gangwar et al., 2021).

Most systems using pre-trained transformer models are submissions to **MeasEval**, that is, task 8 at SemEval 2021 (Harper et al., 2021). Within the given paragraphs, all quantities and units had to be identified. Subsequently, quantities had to be classified into different modifiers[4]. Hereinafter, measured entity, property and qualifier spans had to be identified. Finally, relations between the identified spans had to be extracted[5]. Sharing the same task and evaluation allows for a fair comparison of the systems: all submissions accompanied by a system paper are learning systems, most of which cast quantity span identification as a sequence labeling problem and approach it with a transformer-based model. In addition, many systems utilize a cascaded approach, in which the quantity is identified in a first stage and the other spans and relations are extracted in a second stage. Davletov et al. (2021) cast quantity span extraction as a sequence labeling problem and fine-tune a LUKE NER model (Yamada et al., 2020) on it. A RoBERTa-based model (Liu et al., 2019) extracts all other spans in a question answering style multi-task learning setting without question prefixes. They use a simple data augmentation approach and surround quantity spans with special tokens. Likely limited by the small training set, the system ranked first yielding an overlap F1 score of 51.9 (averaged over all subtasks). Resembling the inter-annotator agreements, the results are significantly better for quantity (86.1 F1) and unit identification (72.2 F1) and much worse for qualifier identification (16.3 F1). Similarly, CONNER (Cao et al., 2021) (ranking 2nd) uses a transformer-based cascaded approach. Quantities are identified with an ensemble of a RoBERTa encoder with a PointerNet (Vinyals et al., 2017) and a CRF layer on top, respectively. For each identified quantity, relation-specific taggers (Wei et al., 2020), which extend the same architecture, identify the other spans. Gangwar et al. (2021) (3rd) and Karia et al. (2021) (6th) formalize quantity span extraction as a sequence labeling problem for which they employ a SciBERT+CRF (Beltagy et al., 2019) and BioBERT (Lee et al., 2020) model, respectively. For each identified quantity, the related measured entities, properties and qualifiers are identified in consecutive sequence labeling passes by surrounding already predicted spans with special symbols using multiple SciBERT+CRF models and a BioBERT model in a multi-task learning setting, respectively. Avram et al. (2021) (5th) identifies quantities using a RoBERTa+CRF model for IOB sequence labeling and extracts related measured entities, properties and qualifiers by treating span extraction as multi-turn question answering (Li et al., 2019) using the same pre-trained language model and relation-specific question templates. Lathiff et al. (2021) (8th) classify token pairs from dependency tree sub-graphs between tokens tagged as cardinal number (CD tag) and other nodes with a deep graph convolution neural network (Zhang and Chen, 2018). Diverging from the cascaded approach, Therien et al. (2021) (4th) extract all span types in a single sequence labeling pass. Although this has the advantage of joint inference across all spans, only one class is assigned to each token, yet the dataset includes instances where, for example, a quantity is a qualifier of another quantity (Harper et al., 2021). As tokens are not distinguished in being inside or at the beginning of a span, adjacent tokens of the same class are merged into a single annotation. Relation extraction is performed based on a distance-based heuristic. Few-shot learning using GPT-3 (Brown et al., 2020) turned out to be an unsuccessful approach (Kohler and Jr, 2021).

Some systems diverge from casting measured property extraction as another span identification problem, but classify the quantity (or given text) into its corresponding measured property (Bakalov et al., 2011; Gruss et al., 2018; Foppiano et al., 2019a) or extract relational triples in which the measured property is a relation between the measured entity and quantity (Hoffmann et al., 2010; Vlachos and Riedel, 2015; Madaan et al., 2016; Saha et al., 2017; Hsiao et al., 2020). Ning et al. (2022) extract the measured property, as well as the spatial and temporal scope of a quantitative statement in a sequence-to-sequence approach using the T5 language model (Raffel et al., 2020).

[4]viz., approx., count, range, list, mean, median, tolerance, mean with standard deviation, mean with tolerance, and range with tolerance

[5]viz. has quantity, has property and qualifies

Extraction of relational triples. For systems extracting relational triples, it is common to use approximate matching when comparing quantities against seed facts or entries in a knowledge base (Hoffmann et al., 2010; Vlachos and Riedel, 2015; Madaan et al., 2016; Saha et al., 2017). LUCHS (Hoffmann et al., 2010) extracts triples using many relation-specific CRF extractors for both numerical and textual attributes. The relation-specific extractors are distantly supervised by matching facts from Wikipedia infoboxes with sentences from the articles they are embedded in. In this way, the system scales to a large number of relations. However, this approach does not generalize well beyond the simplified setting of matching facts within the same article (Vlachos and Riedel, 2015; Madaan et al., 2016). Hence, Vlachos and Riedel (2015) propose an algorithm for extracting numerical triples (e.g., <*Germany, Population, 83 000 000*>) from general text based on facts in a knowledge base. Similarly, Madaan et al. (2016) describe a rule-based system (NumberRule) and a distantly supervised learning-based system (NumberTron) for the extraction of numerical, geopolitical relations. Having a much higher recall and a slightly higher precision than NumberRule, NumberTron achieves an F1 score of 63.78, which is slightly above the F1 score of 61 achieved by LUCHS.

Unlike the aforementioned systems, Saha et al. (2017) approaches measurement extraction in an **Open IE** setting. Hundman and Mattmann (2017) argue that the recall of standard Open IE systems is lower for measurement extraction because such systems are "centered on verb-mediated propositions and measurement context occurs in a variety of other forms such as adverbials".

Template filling and event extraction. Other systems cast measurement extraction as a template filling (Mykowiecka et al., 2009; Zhang and El-Gohary, 2016; Lamm et al., 2018b; Friedrich et al., 2020) or event extraction task (Intxaurrondo et al., 2015). Friedrich et al. (2020) identify entity mentions (that is, material, quantity, device and experiment) and slot-fillers in two consecutive IOB sequence labeling passes. Intxaurrondo et al. (2015) frame the extraction of information about earthquakes from tweets as an event extraction task. Numerical event arguments such as magnitude, depth or deaths are considered. Feature aggregation to better handle ambiguity, as well as approximate matching to cope with inaccuracies when using distant supervision significantly improves performance. Lamm et al. (2018a) define "A Semantic Role-Labeling Schema for Quantitative Facts", which is more generally applicable than the aforementioned templates, and apply it in the identification of analogous and distinct roles of quantitative facts (Lamm et al., 2018b). The task imposes several constraints whose enforcement by solving an integer linear program improves performance.

3.3 Unit span identification

Unit spans, which are typically located within the respective quantity spans, are detected using character-level BiLSTM (Avram et al., 2021; Gangwar et al., 2021; Mehta et al., 2021; Liu et al., 2021b), character-level CRF (Foppiano et al., 2019b) or transformer models (Davletov et al., 2021; Liu et al., 2021a; Kohler and Jr, 2021; Panapitiya et al., 2021). Presumably, character-level methods are more prevalent, because they are better able to represent units given as combinations of one-character-long symbols (e.g., *'k'*, *'m'*, *'/'*, *'s'*) and unit spans are often identified considering only the relatively short quantity strings. Many other systems identify units using rules and dictionaries. In MeasEval, a simple rule-based approach ranked third in unit span identification (Karia et al., 2021). In fact, despite solving other subtasks with machine learning methods, many systems leverage rules and dictionaries to identify units (Lathiff et al., 2021; Cao et al., 2021; Therien et al., 2021).

3.4 Quantity modifier extraction

Quantity modifier extraction is also approached via rules and keywords (Roy et al., 2015; Liu et al., 2021b; Karia et al., 2021), CRF (Foppiano et al., 2019b), char-level BiLSTM (Avram et al., 2021) and BERT-based models (Gangwar et al., 2021; Therien et al., 2021; Lathiff et al., 2021; Liu et al., 2021a; Cao et al., 2021; Davletov et al., 2021). In MeasEval, quantity modifier extraction is framed as a quantity span classification. Interestingly, CONNER (Cao et al., 2021) predicts the quantity modifier based on only the quantity span, as additional context proved to be detrimental. For rule-based systems, the unit span and type of quantity is often inherent to the pattern that matches the quantity (Xiao et al., 2013; Jones et al., 2014; Patterson et al., 2017).

3.5 Relation extraction

For both rule- and learning-based systems, relation extraction or the grouping of identified spans is often already **inherent** to the approaches for span identification. It is either implicit in the span extraction patterns, relation-specific tagging, or to modeling measurement extraction as a template filling task. **Relation-specific tagging** anchored at already identified quantities appears advantageous in MeasEval compared to the more traditional approach of performing span identification for all concept types, followed by a pairwise relation classification (Harper et al., 2021). Since they are relatively easy to identify, most sequential approaches start with identifying quantities. In such a multi-stage approach, errors in the first stage propagate to all other subtasks, rendering them sensitive to quantity span extraction (Avram et al., 2021; Karia et al., 2021). However, when adopting relation-specific tagging, span identification and relation extraction are jointly performed for the remaining concepts, which shortens the error cascade. For example, when answering the relation-specific question *"Which property is quantified by 150 W?"* in an extractive question answering pass, the respective span and its relation to the quantity are jointly extracted. In addition, fusing the input texts with predictions of earlier stages provides additional valuable information in later stages. A sequential approach starting with quantity span identification also proves valuable for rule-based systems; Zhang and El-Gohary (2016) compared a sequential approach with a concurrent one for the rule-based extraction of quantitative information and found the sequential approach to both require fewer patterns and yield better results. In LaTeX-Numeric (Mehta et al., 2021), the B and I labels for quantities are attribute-specific (e.g., B-WEIGHT). Hence, quantities are identified and assigned to a measured property in a single step. Especially in rule-based systems, it is common to relate concepts to each other via **proximity heuristics**, that is, to assume all concepts within a sentence, paragraph, character window or those that are closest to each other belong together. Other systems rely on **dependency tree analyses** (Nanba et al., 2007; Madaan et al., 2016; Kim et al., 2017b,a; Kononova et al., 2019), whilst **pairwise classification** on the identified spans (Yim et al., 2016; Kang et al., 2017) is rarely performed.

3.6 Post-processing

In post-processing, candidates might be normalized and filtered according to different criteria. Ill-formed intervals, quantities outside a viable range, quantities possessing inappropriate units or that do not contain digits and strings like *'two'* or *'teen'* are dropped (Tetko et al., 2016; Hao et al., 2016; Wu et al., 2018; Liu et al., 2021a). Based on viable ranges, missing units can be inferred (Cai et al., 2019). Implicitly stated values might be replaced by known numeric values (e.g., *'room temperature'* $\rightarrow 21\,^{\circ}\mathrm{C}$; Roy et al., 2015; Kuniyoshi et al., 2021), absolute values might be calculated for relative values (Mykowiecka et al., 2009) and non-scientific units might be replaced with WordNet synsets (Roy et al., 2015). Furthermore, task-specific constraints might be enforced (Sevenster et al., 2015b; Lamm et al., 2018b). Depending on the use case, additional tasks and post-processing steps might be performed, such as the pairing of measurements with prior measurements (Sevenster et al., 2013, 2015b,a), determining whether a lab test is normal or abnormal (Jiang et al., 2020) or calculating balanced chemical equations from the extracted quantitative data (Kononova et al., 2019).

4 Challenges

Quantities are easy to identify in text, both numbers and units, which facilitates anchoring semantic role labeling schemata (Lamm et al., 2018a). In addition, many numerical relations are accompanied by only a few keywords (Madaan et al., 2016) and values of numerical attributes "can be estimated even if they are not explicitly mentioned in the text" (Davidov and Rappoport, 2010). Nonetheless, measurement extraction poses various challenges:

Measurements are diversely expressed. Quantities can be expressed in a myriad of different surface forms, yet alone by different levels of rounding and combinations of units. In addition, different writing styles for decimal and thousands separators exist. Also, complex patterns involving multiple quantities such as *"group 1, 2 and 3 were given 4, 5 and 6* $\frac{\mu g}{mL}$*, respectively"* are common (Deus et al., 2017) and measurement extraction might include parsing formulaic expressions (e.g., "$t(29) = -1.85, p = 0.074$"; Epp et al., 2021).

Modifiers have a great impact on meaning. A subtle change of its modifiers can dramatically alter the meaning of a quantity (e.g., consider the differ-

ence in *'above'* instead of *'well below'* 1.5 °C). Thus, quantity modifiers must be correctly extracted. The same applies to change words like *'increase'* (Madaan et al., 2016). Additionally, modifiers concerning measured entities or properties can subtly alter the scope of a quantitative statement. There is a huge semantic difference in *'India'* and *'rural India'*, or *'cell efficiency'* and *'system efficiency'* (Madaan et al., 2016). Even the semantics of bare numerals are still being analyzed in the linguistic literature (Bylinina and Nouwen, 2020).

Qualifiers are difficult to identify. Quantities are precise and, as such, are only valid under specific constraints. Thus, the constraints for which the quantity holds true must also be precisely defined. However, even humans struggle to agree on what is deemed a qualifier; in Harper et al. (2021) the inter-annotator agreement for identifying qualifiers was worse than for all other concept types. In addition, relevant context is often distant. IE is often performed sentence by sentence. Yet, the context given by a single sentence is often much too narrow for understanding measurement contexts (Weikum, 2020).

The document genres that measurement extraction is applied to are often written in domain-specific and complex languages. Clinical reports and notes, for example, include various quantitative information like ages, laboratory test results, dates, severity, odds ratios, and more (Hanauer et al., 2019). However, clinical reports pose various challenges for NLP, i.a., misspellings, temporality, hedge phrases and negation (Hanauer et al., 2019; Nadkarni et al., 2011; Edinger et al., 2012; Mykowiecka et al., 2009). Some abbreviations and acronyms are ambiguous or equal stopwords (e.g., 'OR' for operating room; Hanauer et al., 2019). Medical texts are often written in a complex and informal manner that is sometimes even confusing for humans (Patterson et al., 2017), rendering POS tags and syntactic features less effective (Liu et al., 2021b). Other document genres like product data sheets make heavy use of tables and technical drawings to communicate information (Opasjumruskit et al., 2019a). Additionally, many systems start from PDF documents as input. Parsing PDF documents into machine-readable formats creates noise. The situation is worse for measurement extraction, as mathematical formatting is likely to be lost and special characters are inconsistently converted (e.g., '$10^3\,m^2$' → '103 m2' and '$€_{2015}$' → 'V2015') (Maiya et al., 2015; Foppiano et al., 2019a). In the context of measurement extraction, the wrongly parsed tokens are often only one or a few characters long, which makes their correct recovery more difficult. For example, it is harder to recover '€' from 'V' than *'photovoltaic'* from *'photo□oltaic'*.

Common sense and domain knowledge is required for understanding quantitative statements when information is omitted due to brevity, when dealing with constants like *'speed of light'*, to infer whether an interval includes or excludes its endpoints, or in cases of quantities given relative to a standard (e.g., *"1.15 times the upper limit of normal"*; Hao et al., 2016). Implicit assumptions and world knowledge are common when describing physical processes (Kuehne and Forbus, 2004). Also, gapping and unit ellipsis are common phenomena (Lamm et al., 2018b). Furthermore, measurements must be distinguished from irrelevant quantifications such as *"he had two priorities"* (Alonso and Sellam, 2018) and from references to chemical entities (Hawizy et al., 2011), figures, tables and cited literature (Agatonovic et al., 2008; Aras et al., 2014).

Numeracy has received little attention in NLP until recently (Thawani et al., 2021b). Yet, the relevance of numerical reasoning for natural language understanding is evident from a simple example: *"The battery of the hybrid Toyota Prius lasts well over 100,000 miles."* (Weikum, 2020). Considering the order of magnitude, most humans will infer that this statement refers to the battery lifetime and not the driving range possible with a single charge. For language models to do so, a good representation of numbers is required. However, common models in NLP, such as BERT, suffer from sub-optimal number representations (Wallace et al., 2019; Zhang et al., 2020b; Thawani et al., 2021a). This limits them in tasks that require numerical reasoning and possibly even beyond (Dua et al., 2019; Thawani et al., 2021a).

Weaker distant supervision. Distant supervision is based on a simple heuristic: If a sentence includes a pair of entities for which a relation in a knowledge base exists, there is a high chance that this sentence expresses that relation (Mintz et al., 2009). "However, since quantities can appear in far more contexts than typical entities, dis-

tantly supervised training data becomes much more noisy", especially "for small whole numbers that appear unit-less or with popular units" (Madaan et al., 2016). Furthermore, many quantities change over time (e.g., consider the rising CO_2 concentration in the atmosphere). In addition, even the same quantity in different documents might be expressed with different numbers of decimal places or with different units. Thus, normalization and partial matching (that is, approximate rather than exact matching) is required (Madaan et al., 2016; Vlachos and Riedel, 2015; Intxaurrondo et al., 2015). This also illustrates why keyword-search is inappropriate for quantities (Agatonovic et al., 2008) and why it is difficult to generate numerical answers in question answering (Liu et al., 2016).

5 A vision for the future

Having arranged and summarized the prior work in measurement extraction, we now provide several recommendations that might positively shape future research. These go beyond addressing the aforementioned challenges, which must inevitably be dealt with, in that they are concrete recommendations for action.

A common terminology is what language around quantitative information extraction is lacking. Although standardization efforts exist (Hao et al., 2017, 2018), different terms are used for the same concept and the same terms are used for different concepts. For example, the terms measurement entity (Yim et al., 2016), numeric property (Aras et al., 2014), and value (Friedrich et al., 2020) are all used for referring to a quantity. Adding to the confusion, in Lamm et al. (2018a) a quantity denotes a measured property. Aiming to end this confusion, we propose to adopt the terminology of MeasEval, which defines the terms quantity, measured entity, measured property, quantity modifiers, and qualifiers (Harper et al., 2021). In line with the unit ontology QUDT (Ray, 2011), a quantity is composed of a numeric value and a unit.

More extensive datasets that cover quantities as well as their contexts could greatly improve results in measurement extraction. The dataset used in MeasEval, for example, consists of only 428 paragraphs (Harper et al., 2021), limiting the performance of the learning-based methods (Lathiff et al., 2021). More generic annotations could render datasets for measurement extraction more sustainable. We argue that the reuse of datasets is hindered by incompatible annotations. For example, the sets of quantity modifier classes in the datasets of MeasEval and Grobid-quantities do not match. Also, the sets of modifiers are selective, not considering all occurring modifiers and combinations. Therefore, we propose annotating the quantity spans with pseudo-mathematical representations that can be parsed into classes depending on the task or directly used for sequence-to-sequence approaches.

Improving the numerical reasoning capabilities of the models may well improve the performance of measurement extraction systems. Character-level embeddings, for example, outperform word- and subword-level methods (Wallace et al., 2019). Altering the **surface form** of all numbers during pre-processing can improve model performance (Wallace et al., 2019; Zhang et al., 2020b; Nogueira et al., 2021). Furthermore, extending language models with special **representations** of numbers improves numerical reasoning capabilities (Thawani et al., 2021a). Andor et al. (2019) propose the extension of language models with a set of executable programs for **symbolic reasoning**. Still, recent advances in numerical reasoning have been barely considered in the literature on measurement extraction.

Document understanding remains an ambitious objective. Systems for measurement extraction that consider document context and incorporate information from other modalities are rare (Swain and Cole, 2016; Mavračić et al., 2021; Hsiao et al., 2020). In fact, many systems operate on a sentence-level or truncate the processed text after a fixed token limit. We argue that both context from other modalities (e.g., joint inference from text and tables) and distant context should be considered.

6 Limitations

Relevant literature has been iteratively identified using different academic search engines, foremost Semantic Scholar[6], and by tracing the references in already identified publications. Publications disclosing systems whose scope is too narrow or offset, that target figures or tables, or that lack detailed information are dropped. Related work that was not deemed relevant is listed in Appendix A. That said, many systems extract quantities, amongst other

[6] https://www.semanticscholar.org/

concepts, but do not elaborate on it. It is likely that additional systems exist that identify quantities and their contexts, but which are not included in this review. It was decided against a quantitative assessment of the systems' performance, as both their scopes and evaluations differ from each other, making a fair comparison difficult.

Acknowledgements

The authors would like to thank the German Federal Government, the German state governments, and the Joint Science Conference (GWK) for their funding and support as part of the NFDI4Ing consortium. Funded by the German Research Foundation (DFG) – project number: 442146713. Furthermore, this work was supported by the Helmholtz Association under the program "Energy System Design".

References

Milan Agatonovic, Niraj Aswani, Kalina Bontcheva, Hamish Cunningham, Thomas Heitz, Yaoyong Li, Ian Roberts, and Valentin Tablan. 2008. Large-scale, parallel automatic patent annotation. In *Proceedings of the 1st ACM Workshop on Patent Information Retrieval*, PaIR '08, pages 1–8, New York, NY, USA. Association for Computing Machinery.

Omar Alonso and Thibault Sellam. 2018. Quantitative Information Extraction From Social Data. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pages 1005–1008, New York, NY, USA. Association for Computing Machinery.

Daniel Andor, Luheng He, Kenton Lee, and Emily Pitler. 2019. Giving BERT a Calculator: Finding Operations and Arguments with Reading Comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5947–5952, Hong Kong, China. Association for Computational Linguistics.

Hidir Aras, René Hackl-Sommer, Michael Schwantner, and Mustafa Sofean. 2014. Applications and Challenges of Text Mining with Patents. In *Proceedings of the First International Workshop on Patent Mining and Its Applications (IPaMin 2014) Co-Located with Konvens 2014*, volume Vol-1292, Hildesheim, Germany.

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 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 534–540, Online. Association for Computational Linguistics.

Ali Ayadi, Mélanie Auffan, and Jérôme Rose. 2020. Ontology-based NLP information extraction to enrich nanomaterial environmental exposure database. *Procedia Computer Science*, 176:360–369.

Anton Bakalov, Ariel Fuxman, Partha Pratim Talukdar, and Soumen Chakrabarti. 2011. SCAD: Collective discovery of attribute values. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 447–456, New York, NY, USA. Association for Computing Machinery.

Kai Barkschat. 2014. Semantic Information Extraction on Domain Specific Data Sheets. In *The Semantic Web: Trends and Challenges*, Lecture Notes in Computer Science, pages 864–873, Cham. Springer International Publishing.

Božo Bekavac, Željko Agić, Krešimir Šojat, and Marko Tadić. 2009. Detecting Measurement Expressions using NooJ. *Finite State Language Engineering: NooJ 2009 International Conference and Workshop*, page 121.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3613–3618, Hong Kong, China. Association for Computational Linguistics.

Selen Bozkurt, Emel Alkim, Imon Banerjee, and Daniel L. Rubin. 2019. Automated Detection of Measurements and Their Descriptors in Radiology Reports Using a Hybrid Natural Language Processing Algorithm. *Journal of Digital Imaging*, 32(4):544–553.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Lisa Bylinina and Rick Nouwen. 2020. Numeral semantics. *Language and Linguistics Compass*, 14(8):e12390.

Tianrun Cai, Luwan Zhang, Nicole Yang, Kanako K. Kumamaru, Frank J. Rybicki, Tianxi Cai, and Katherine P. Liao. 2019. EXTraction of EMR numerical

data: An efficient and generalizable tool to EXTEND clinical research. *BMC medical informatics and decision making*, 19(1):226.

Jiarun Cao, Yuejia Xiang, Yunyan Zhang, Zhiyuan Qi, Xi Chen, and Yefeng Zheng. 2021. CONNER: A Cascade Count and Measurement Extraction Tool for Scientific Discourse. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1239–1244, Online. Association for Computational Linguistics.

Arun Chaganty and Percy Liang. 2016. How Much is 131 Million Dollars? Putting Numbers in Perspective with Compositional Descriptions. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 578–587, Berlin, Germany. Association for Computational Linguistics.

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2019a. Numeral Attachment with Auxiliary Tasks. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, pages 1161–1164, New York, NY, USA. Association for Computing Machinery.

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020. NumClaim: Investor's Fine-grained Claim Detection. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1973–1976, Virtual Event Ireland. ACM.

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021. Numerals in Financial Narratives. In Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen, editors, *From Opinion Mining to Financial Argument Mining*, SpringerBriefs in Computer Science, pages 55–71. Springer, Singapore.

Chung-Chi Chen, Hen-Hsen Huang, Yow-Ting Shiue, and Hsin-Hsi Chen. 2018. Numeral Understanding in Financial Tweets for Fine-grained Crowd-based Forecasting. *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 136–143.

Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2019b. Numeracy-600K: Learning Numeracy for Detecting Exaggerated Information in Market Comments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6307–6313, Florence, Italy. Association for Computational Linguistics.

Chung-Chi Chen, Hen-Hsen Huang, Chia-Wen Tsai, and Hsin-Hsi Chen. 2019c. CrowdPT: Summarizing Crowd Opinions as Professional Analyst. In *The World Wide Web Conference*, WWW '19, pages 3498–3502, New York, NY, USA. Association for Computing Machinery.

Nancy A. Chinchor. 1998. Overview of MUC-7. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*.

Laura Chiticariu, Yunyao Li, and Frederick R. Reiss. 2013. Rule-Based Information Extraction is Dead! Long Live Rule-Based Information Extraction Systems! In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 827–832, Seattle, Washington, USA. Association for Computational Linguistics.

Callum J. Court and Jacqueline M. Cole. 2018. Auto-generated materials database of Curie and Néel temperatures via semi-supervised relationship extraction. *Scientific Data*, 5(1):180111.

Hamish Cunningham, Valentin Tablan, Angus Roberts, and Kalina Bontcheva. 2013. Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. *PLOS Computational Biology*, 9(2):e1002854.

Dmitry Davidov and Ari Rappoport. 2010. Extraction and Approximation of Numerical Attributes from the Web. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1308–1317, Uppsala, Sweden. Association for Computational Linguistics.

Adis Davletov, Denis Gordeev, Nikolay Arefyev, and Emil Davletov. 2021. LIORI at SemEval-2021 Task 8: Ask Transformer for measurements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1249–1254, Online. Association for Computational Linguistics.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Thaer M. Dieb, Masaharu Yoshioka, and Shinjiro Hara. 2012. Automatic Information Extraction of Experiments from Nanodevices Development Papers. In *2012 IIAI International Conference on Advanced Applied Informatics*, pages 42–47.

Thaer M. Dieb, Masaharu Yoshioka, Shinjiro Hara, and Marcus C. Newton. 2015. Framework for automatic information extraction from research papers on nanocrystal devices. *Beilstein Journal of Nanotechnology*, 6:1872–1882.

Thaer M. Dieb, Masaharu Yoshioka, Shinjiroh Hara, and Marcus C. Newton. 2014. Automatic Annotation of Parameters from Nanodevice Development

Research Papers. In *Proceedings of the 4th International Workshop on Computational Terminology (Computerm)*, pages 77–85, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The Automatic Content Extraction (ACE) Program – Tasks, Data, and Evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Xin Luna Dong, Xiang He, Andrey Kan, Xian Li, Yan Liang, Jun Ma, Yifan Ethan Xu, Chenwei Zhang, Tong Zhao, Gabriel Blanco Saldana, Saurabh Deshpande, Alexandre Michetti Manduca, Jay Ren, Surender Pal Singh, Fan Xiao, Haw-Shiuan Chang, Giannis Karamanolakis, Yuning Mao, Yaqing Wang, Christos Faloutsos, Andrew McCallum, and Jiawei Han. 2020. AutoKnow: Self-Driving Knowledge Collection for Products of Thousands of Types. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2724–2734.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Tracy Edinger, Aaron M. Cohen, Steven Bedrick, Kyle Ambert, and William Hersh. 2012. Barriers to Retrieving Patient Information from Electronic Health Record Data: Failure Analysis from the TREC Medical Records Track. *AMIA Annual Symposium Proceedings*, 2012:180–188.

Steffen Epp, Marcel Hoffmann, Nicolas Lell, Michael Mohr, and Ansgar Scherp. 2021. A Machine Learning Pipeline for Automatic Extraction of Statistic Reports and Experimental Conditions from Scientific Papers.

Luca Foppiano, Thaer M. Dieb, Akira Suzuki, and Masashi Ishii. 2019a. Proposal for Automatic Extraction Framework of Superconductors Related Information from Scientific Literature. page 6.

Luca Foppiano, Laurent Romary, Masashi Ishii, and Mikiko Tanifuji. 2019b. Automatic Identification and Normalisation of Physical Measurements in Scientific Literature. In *Proceedings of the ACM Symposium on Document Engineering 2019*, pages 1–4, Berlin Germany. ACM.

Annemarie Friedrich, Heike Adel, Federico Tomazic, Johannes Hingerl, Renou Benteau, Anika Marusczyk, and Lukas Lange. 2020. The SOFC-Exp Corpus and Neural Approaches to Information Extraction in the Materials Science Domain. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1255–1268, Online. Association for Computational Linguistics.

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 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1232–1238, Online. Association for Computational Linguistics.

Jennifer H Garvin, Scott L DuVall, Brett R South, Bruce E Bray, Daniel Bolton, Julia Heavirland, Steve Pickard, Paul Heidenreich, Shuying Shen, Charlene Weir, Matthew Samore, and Mary K Goldstein. 2012. Automated extraction of ejection fraction for quality measurement using regular expressions in Unstructured Information Management Architecture (UIMA) for heart failure. *Journal of the American Medical Informatics Association*, 19(5):859–866.

Ralph Grishman. 2019. Twenty-five years of information extraction. *Natural Language Engineering*, 25(6):677–692.

Ralph Grishman and Beth Sundheim. 1995. Design of the MUC-6 Evaluation. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference- 6: A Brief History. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Richard Gruss, Alan S. Abrahams, Weiguo Fan, and G. Alan Wang. 2018. By the numbers: The magic of numerical intelligence in text analytic systems. *Decision Support Systems*, 113:86–98.

David A. Hanauer, Qiaozhu Mei, V. G. Vinod Vydiswaran, Karandeep Singh, Zach Landis-Lewis, and Chunhua Weng. 2019. Complexities, variations, and errors of numbering within clinical notes: The potential impact on information extraction and cohort-identification. *BMC Medical Informatics and Decision Making*, 19(3):75.

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(03):266–275.

Tianyong Hao, Haotai Wang, Xinyu Cao, and Kiyong Lee. 2018. Annotating Measurable Quantitative Informationin Language: For an ISO Standard. In *Proceedings 14th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, pages 69–75, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Tianyong Hao, Yunyan We, Jiaqi Qiang, Haitao Wang, and Kiyong Lee. 2017. The representation and extraction of qunatitative information. In *Proceedings of the 13th Joint ISO-ACL Workshop on Interoperable Semantic Annotation (ISA-13)*.

Corey Harper, Jessica Cox, Curt Kohler, Antony Scerri, Ron Daniel Jr., and Paul Groth. 2021. SemEval-2021 Task 8: MeasEval – Extracting Counts and Measurements and their Related Contexts. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 306–316, Online. Association for Computational Linguistics.

Lezan Hawizy, David M. Jessop, Nico Adams, and Peter Murray-Rust. 2011. ChemicalTagger: A tool for semantic text-mining in chemistry. *Journal of Cheminformatics*, 3(1):17.

Yury Hetsevich and Alena Skopinava. 2014. Processing of Quantitative Expressions with Measurement Units in the Nominative, Genitive, and Accusative Cases for Belarusian and Russian. In Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala, editors, *Text, Speech and Dialogue*, volume 8655, pages 101–107. Springer International Publishing, Cham.

Vinh Thinh Ho, Yusra Ibrahim, Koninika Pal, Klaus Berberich, and Gerhard Weikum. 2019. Qsearch: Answering Quantity Queries from Text. In *The Semantic Web – ISWC 2019*, Lecture Notes in Computer Science, pages 237–257, Cham. Springer International Publishing.

Vinh Thinh Ho, Koninika Pal, Niko Kleer, Klaus Berberich, and Gerhard Weikum. 2020. Entities with Quantities: Extraction, Search, and Ranking. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 833–836, New York, NY, USA. Association for Computing Machinery.

Vinh Thinh Ho, Daria Stepanova, Dragan Milchevski, Jannik Strötgen, and Gerhard Weikum. 2022. Enhancing Knowledge Bases with Quantity Facts. In *Proceedings of the ACM Web Conference 2022*, WWW '22, pages 893–901, New York, NY, USA. Association for Computing Machinery.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. 2011. Knowledge-Based Weak Supervision for Information Extraction of Overlapping Relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 541–550, Portland, Oregon, USA. Association for Computational Linguistics.

Raphael Hoffmann, Congle Zhang, and Daniel S. Weld. 2010. Learning 5000 Relational Extractors. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 286–295, Uppsala, Sweden. Association for Computational Linguistics.

Luke Hsiao, Sen Wu, Nicholas Chiang, Christopher Ré, and Philip Levis. 2020. Creating Hardware Component Knowledge Bases with Training Data Generation and Multi-task Learning. *ACM Transactions on Embedded Computing Systems*, 19(6):42:1–42:26.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging.

Kyle Hundman and Chris A. Mattmann. 2017. Measurement Context Extraction from Text: Discovering Opportunities and Gaps in Earth Science. *CoRR*.

Ander Intxaurrondo, Eneko Agirre, Oier Lopez de Lacalle, and Mihai Surdeanu. 2015. Diamonds in the Rough: Event Extraction from Imperfect Microblog Data. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 641–650, Denver, Colorado. Association for Computational Linguistics.

Zach Jensen, Edward Kim, Soonhyoung Kwon, Terry Z. H. Gani, Yuriy Román-Leshkov, Manuel Moliner, Avelino Corma, and Elsa Olivetti. 2019. A Machine Learning Approach to Zeolite Synthesis Enabled by Automatic Literature Data Extraction. *ACS Central Science*, 5(5):892–899.

Kun Jiang, Tao Yang, Chunyan Wu, Luming Chen, Longfei Mao, Yongyou Wu, Lizong Deng, and Taijiao Jiang. 2020. LATTE: A knowledge-based method to normalize various expressions of laboratory test results in free text of Chinese electronic health records. *Journal of Biomedical Informatics*, 102:103372.

David E. Jones, Sean Igo, John Hurdle, and Julio C. Facelli. 2014. Automatic Extraction of Nanoparticle Properties Using Natural Language Processing: NanoSifter an Application to Acquire PAMAM Dendrimer Properties. *PLOS ONE*, 9(1):e83932.

Tian Kang, Shaodian Zhang, Youlan Tang, Gregory W Hruby, Alexander Rusanov, Noémie Elhadad, and Chunhua Weng. 2017. EliIE: An open-source information extraction system for clinical trial eligibility criteria. *Journal of the American Medical Informatics Association*, 24(6):1062–1071.

Yanna Shen Kang and Mehmet Kayaalp. 2013. Extracting laboratory test information from biomedical text. *Journal of Pathology Informatics*, 4(1):23.

Neel Karia, Ayush Kaushal, and Faraaz Mallick. 2021. KGP at SemEval-2021 Task 8: Leveraging Multi-Staged Language Models for Extracting Measurements, their Attributes and Relations. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 387–396, Online. Association for Computational Linguistics.

Edward Kim, Kevin Huang, Adam Saunders, Andrew McCallum, Gerbrand Ceder, and Elsa Olivetti. 2017a. Materials Synthesis Insights from Scientific Literature via Text Extraction and Machine Learning. *Chemistry of Materials*, 29(21):9436–9444.

Edward Kim, Kevin Huang, Alex Tomala, Sara Matthews, Emma Strubell, Adam Saunders, Andrew McCallum, and Elsa Olivetti. 2017b. Machine-learned and codified synthesis parameters of oxide materials. *Scientific Data*, 4(1):170127.

Edward Kim, Zach Jensen, Alexander van Grootel, Kevin Huang, Matthew Staib, Sheshera Mysore, Haw-Shiuan Chang, Emma Strubell, Andrew McCallum, Stefanie Jegelka, and Elsa Olivetti. 2020. Inorganic Materials Synthesis Planning with Literature-Trained Neural Networks. *Journal of Chemical Information and Modeling*, 60(3):1194–1201.

Youngjun Kim, Jennifer Garvin, Julia Heavirland, and Stéphane M. Meystre. 2013. Improving heart failure information extraction by domain adaptation. *Studies in Health Technology and Informatics*, 192:185–189.

Youngjun Kim, Jennifer H. Garvin, Mary K. Goldstein, Tammy S. Hwang, Andrew Redd, Dan Bolton, Paul A. Heidenreich, and Stéphane M. Meystre. 2017c. Extraction of left ventricular ejection fraction information from various types of clinical reports. *Journal of Biomedical Informatics*, 67:42–48.

Curt Kohler and Ron Daniel Jr. 2021. What's in a Measurement? Using GPT-3 on SemEval 2021 Task 8 - MeasEval. *CoRR*, page 11.

Olga Kononova, Tanjin He, Haoyan Huo, Amalie Trewartha, Elsa A. Olivetti, and Gerbrand Ceder. 2021. Opportunities and challenges of text mining in materials research. *iScience*, 24(3):102155.

Olga Kononova, Haoyan Huo, Tanjin He, Ziqin Rong, Tiago Botari, Wenhao Sun, Vahe Tshitoyan, and Gerbrand Ceder. 2019. Text-mined dataset of inorganic materials synthesis recipes. *Scientific Data*, 6(1):203.

Taku Kudo and Yuji Matsumoto. 2003. Fast Methods for Kernel-Based Text Analysis. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 24–31, Sapporo, Japan. Association for Computational Linguistics.

Sven E Kuehne and Kenneth D Forbus. 2004. Capturing QP-relevant Information from Natural Language Text. In *Proceedings of the 18th International Qualitative Reasoning Workshop*, page 8.

Fusataka Kuniyoshi, Jun Ozawa, and Makoto Miwa. 2021. Analyzing Research Trends in Inorganic Materials Literature Using NLP. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part V*, pages 319–334, Berlin, Heidelberg. Springer-Verlag.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Departmental Papers (CIS)*.

Matthew Lamm, Arun Chaganty, Dan Jurafsky, Christopher D Manning, and Percy Liang. 2018a. QSRL: A Semantic Role-Labeling Schema for Quantitative Facts. page 8.

Matthew Lamm, Arun Chaganty, Christopher D. Manning, Dan Jurafsky, and Percy Liang. 2018b. Textual Analogy Parsing: What's Shared and What's Compared among Analogous Facts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 82–92, Brussels, Belgium. Association for Computational Linguistics.

Nihatha Lathiff, Pavel PK Khloponin, and Sabine Bergler. 2021. CLaC-np at SemEval-2021 Task 8: Dependency DGCNN. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 404–409, Online. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Tongliang Li, Lei Fang, Jian-Guang Lou, Zhoujun Li, and Dongmei Zhang. 2021. AnaSearch: Extract, Retrieve and Visualize Structured Results from Unstructured Text for Analytical Queries. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21, pages 906–909, New York, NY, USA. Association for Computing Machinery.

Xiaoya Li, Fan Yin, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, and Jiwei Li. 2019. Entity-Relation Extraction as Multi-Turn Question Answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1340–1350, Florence, Italy. Association for Computational Linguistics.

Patrick Liu, Niveditha Iyer, Erik Rozi, and Ethan A. Chi. 2021a. Stanford MLab at SemEval-2021 Task 8: 48 Hours Is All You Need. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1245–1248, Online. Association for Computational Linguistics.

Shanshan Liu, Wenjie Nie, Dongfa Gao, Hao Yang, Jun Yan, and Tianyong Hao. 2021b. Clinical quantitative information recognition and entity-quantity association from Chinese electronic medical records. *International Journal of Machine Learning and Cybernetics*, 12(1):117–130.

Shanshan Liu, Xiaoyi Pan, Boyu Chen, Dongfa Gao, and Tianyong Hao. 2018. An Automated Approach for Clinical Quantitative Information Extraction from Chinese Electronic Medical Records. In *Health Information Science*, Lecture Notes in Computer Science, pages 98–109, Cham. Springer International Publishing.

Sijia Liu, Liwei Wang, Donna Ihrke, Vipin Chaudhary, Cui Tao, Chunhua Weng, and Hongfang Liu. 2017. Correlating Lab Test Results in Clinical Notes with Structured Lab Data: A Case Study in HbA1c and Glucose. *AMIA Summits on Translational Science Proceedings*, 2017:221–228.

Yaqing Liu, Lidong Wang, Rong Chen, Yingjie Song, and Yalin Cai. 2016. A PUT-Based Approach to Automatically Extracting Quantities and Generating Final Answers for Numerical Attributes. *Entropy*, 18(6):235.

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.

Patrice Lopez. 2009. GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications. In *Research and Advanced Technology for Digital Libraries*, Lecture Notes in Computer Science, pages 473–474, Berlin, Heidelberg. Springer.

Lefteris Loukas, Manos Fergadiotis, Ilias Chalkidis, Eirini Spyropoulou, Prodromos Malakasiotis, Ion Androutsopoulos, and Georgios Paliouras. 2022. FiNER: Financial Numeric Entity Recognition for XBRL Tagging. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4419–4431, Dublin, Ireland. Association for Computational Linguistics.

Daniel M. Lowe and Roger A. Sayle. 2015. LeadMine: A grammar and dictionary driven approach to entity recognition. *Journal of Cheminformatics*, 7(1):S5.

Aman Madaan, Ashish Mittal, Mausam, Ganesh Ramakrishnan, and Sunita Sarawagi. 2016. Numerical relation extraction with minimal supervision. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 2764–2771, Phoenix, Arizona. AAAI Press.

Arun S. Maiya, Dale Visser, and Andrew Wan. 2015. Mining Measured Information from Text. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 899–902, New York, NY, USA. Association for Computing Machinery.

Juraj Mavračić, Callum J. Court, Taketomo Isazawa, Stephen R. Elliott, and Jacqueline M. Cole. 2021. ChemDataExtractor 2.0: Autopopulated Ontologies for Materials Science. *Journal of Chemical Information and Modeling*, 61(9):4280–4289.

Kartik Mehta, Ioana Oprea, and Nikhil Rasiwasia. 2021. LATEX-Numeric: Language Agnostic Text Attribute Extraction for Numeric Attributes. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 272–279, Online. Association for Computational Linguistics.

Stéphane M Meystre, Youngjun Kim, Glenn T Gobbel, Michael E Matheny, Andrew Redd, Bruce E Bray, and Jennifer H Garvin. 2017. Congestive heart failure information extraction framework for automated treatment performance measures assessment. *Journal of the American Medical Informatics Association*, 24(e1):e40–e46.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics.

Paramita Mirza, Simon Razniewski, Fariz Darari, and Gerhard Weikum. 2017. Cardinal Virtues: Extracting Relation Cardinalities from Text. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 347–351, Vancouver, Canada. Association for Computational Linguistics.

Ines Montani, Matthew Honnibal, Matthew Honnibal, Sofie Van Landeghem, Adriane Boyd, Henning Peters, Paul O'Leary McCann, Maxim Samsonov, Jim Geovedi, Jim O'Regan, Duygu Altinok, György Orosz, Søren Lind Kristiansen, Daniël de Kok, Lj Miranda, Roman, Explosion Bot, Leander Fiedler, Grégory Howard, Edward, Wannaphong Phatthiyaphaibun, Richard Hudson, Yohei Tamura, Sam Bozek, murat, Ryn Daniels, Peter Baumgartner, Mark Amery, and Björn Böing. 2022. Explosion/spaCy: New Span Ruler component, JSON (de)serialization of Doc, span analyzer and more. Zenodo.

Véronique Moriceau. 2006. Numerical Data Integration for Cooperative Question-Answering. In *Proceedings of the Workshop KRAQ'06: Knowledge and Reasoning for Language Processing*.

Agnieszka Mykowiecka, Małgorzata Marciniak, and Anna Kupść. 2009. Rule-based information extraction from patients' clinical data. *Journal of Biomedical Informatics*, 42(5):923–936.

Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. 2011. Natural language processing: An introduction. *Journal of the American Medical Informatics Association : JAMIA*, 18(5):544–551.

Hidetsugu Nanba, Nao Okuda, and Manabu Okumura. 2007. Extraction and Visualization of Trend Information from Newspaper Articles and Blogs. In *Proceedings of the 6th {NTCIR} Workshop Meeting on*

Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access, page 6, National Center of Sciences, Tokyo, Japan. National Institute of Informatics (NII).

Qiang Ning, Ben Zhou, Hao Wu, Haoruo Peng, Chuchu Fan, and Matt Gardner. 2022. A Meta-framework for Spatiotemporal Quantity Extraction from Text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2736–2749, Dublin, Ireland. Association for Computational Linguistics.

Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. 2021. Investigating the Limitations of Transformers with Simple Arithmetic Tasks. *arXiv:2102.13019 [cs]*.

Kobkaew Opasjumruskit, Diana Peters, and Sirko Schindler. 2019a. ConTrOn: Continuously Trained Ontology based on Technical Data Sheets and Wikidata. *arXiv:1906.06752 [cs]*.

Kobkaew Opasjumruskit, Sirko Schindler, Laura Thiele, and Philipp Matthias Schafer. 2019b. Towards Learning from User Feedback for Ontology-based Information Extraction. page 9.

Gihan Panapitiya, Fred Parks, Jonathan Sepulveda, and Emily Saldanha. 2021. Extracting Material Property Measurement Data from Scientific Articles. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5393–5402, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Olga V. Patterson, Matthew S. Freiberg, Melissa Skanderson, Samah J. Fodeh, Cynthia A. Brandt, and Scott L. DuVall. 2017. Unlocking echocardiogram measurements for heart disease research through natural language processing. *BMC Cardiovascular Disorders*, 17(1):151.

Heiko Paulheim. 2017. A Robust Number Parser Based on Conditional Random Fields. In *KI 2017: Advances in Artificial Intelligence*, Lecture Notes in Computer Science, pages 337–343, Cham. Springer International Publishing.

Thorge Petersen, Muhammad Asif Suryani, Christian Beth, Hardik Patel, Klaus Wallmann, and Matthias Renz. 2021. Geo-Quantities: A Framework for Automatic Extraction of Measurements and Spatial Context from Scientific Documents. In *17th International Symposium on Spatial and Temporal Databases*, SSTD '21, pages 166–169, New York, NY, USA. Association for Computing Machinery.

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 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 397–403, Online. Association for Computational Linguistics.

Disheng Qiu, Luciano Barbosa, Xin Luna Dong, Yanyan Shen, and Divesh Srivastava. 2015. Dexter: Large-scale discovery and extraction of product specifications on the web. *Proceedings of the VLDB Endowment*, 8(13):2194–2205.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Steven Ray. 2011. Quantities, Units, Dimensions and Types.

Martin Rezk, Laura Alonso Alemany, Lasguido Nio, and Ted Zhang. 2019. Accurate Product Attribute Extraction on the Field. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 1862–1873, Macao, Macao. IEEE.

Subhro Roy, Shyam Upadhyay, and Dan Roth. 2016. Equation Parsing : Mapping Sentences to Grounded Equations. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1088–1097, Austin, Texas. Association for Computational Linguistics.

Subhro Roy, Tim Vieira, and Dan Roth. 2015. Reasoning about Quantities in Natural Language. *Transactions of the Association for Computational Linguistics*, 3:1–13.

Swarnadeep Saha and Mausam. 2018. Open Information Extraction from Conjunctive Sentences. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2288–2299, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Swarnadeep Saha, Harinder Pal, and Mausam. 2017. Bootstrapping for Numerical Open IE. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 317–323, Vancouver, Canada. Association for Computational Linguistics.

Sunita Sarawagi and Soumen Chakrabarti. 2014. Open-domain quantity queries on web tables: Annotation, response, and consensus models. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 711–720, New York, NY, USA. Association for Computing Machinery.

Jodi Schneider, Michele Avissar-Whiting, Caitlin Bakker, Hannah Heckner, Sylvain Massip, Randy Townsend, and Nathan D. Woods. 2021. Addressing disorder in scholarly communication: Strategies from NISO 2021. *Information Services & Use*, Preprint(Preprint):1–15.

Stefan Schweter and Alan Akbik. 2021. FLERT: Document-Level Features for Named Entity Recognition. *arXiv:2011.06993 [cs]*.

M. Sevenster, J. Buurman, P. Liu, J. F. Peters, and P. J. Chang. 2015a. Natural Language Processing Techniques for Extracting and Categorizing Finding Measurements in Narrative Radiology Reports. *Applied Clinical Informatics*, 06(03):600–610.

Merlijn Sevenster, Jeffrey Bozeman, Andrea Cowhy, and William Trost. 2013. Automatically Pairing Measured Findings across Narrative Abdomen CT Reports. *AMIA Annual Symposium Proceedings*, 2013:1262–1271.

Merlijn Sevenster, Jeffrey Bozeman, Andrea Cowhy, and William Trost. 2015b. A natural language processing pipeline for pairing measurements uniquely across free-text CT reports. *Journal of Biomedical Informatics*, 53:36–48.

Basel Shbita, Arunkumar Rajendran, Jay Pujara, and Craig A Knoblock. 2019. Parsing, Representing and Transforming Units of Measure. *Modeling the World's Systems*, page 7.

Alena Skopinava and Yury Hetsevich. 2013. Identification of Expressions with Units of Measurement in Scientific, Technical & Legal Texts in Belarusian and Russian. In *Proceedings of the Workshop on Integrating IR Technologies for Professional Search*, Moscow, Russian Federation.

Daniel Spokoyny, Ivan Lee, Zhao Jin, and Taylor Berg-Kirkpatrick. 2022. Masked Measurement Prediction: Learning to Jointly Predict Quantities and Units from Textual Context. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 17–29, Seattle, United States. Association for Computational Linguistics.

Julien Subercaze. 2017. Chaudron: Extending DBpedia with Measurement. In *The Semantic Web*, Lecture Notes in Computer Science, pages 434–448, Cham. Springer International Publishing.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D. Manning. 2012. Multi-instance Multi-label Learning for Relation Extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 455–465, Jeju Island, Korea. Association for Computational Linguistics.

Matthew C. Swain and Jacqueline M. Cole. 2016. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. *Journal of Chemical Information and Modeling*, 56(10):1894–1904.

Igor V. Tetko, Daniel M. Lowe, and Antony J. Williams. 2016. The development of models to predict melting and pyrolysis point data associated with several hundred thousand compounds mined from PATENTS. *Journal of Cheminformatics*, 8(1):2.

Avijit Thawani, Jay Pujara, and Filip Ilievski. 2021a. Numeracy enhances the Literacy of Language Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6960–6967, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Avijit Thawani, Jay Pujara, Filip Ilievski, and Pedro Szekely. 2021b. Representing Numbers in NLP: A Survey and a Vision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–656, Online. Association for Computational Linguistics.

Benjamin Therien, Parsa Bagherzadeh, and Sabine Bergler. 2021. CLaC-BP at SemEval-2021 Task 8: SciBERT Plus Rules for MeasEval. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 410–415, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2017. Pointer Networks.

Andreas Vlachos and Sebastian Riedel. 2015. Identification and Verification of Simple Claims about Statistical Properties. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2596–2601, Lisbon, Portugal. Association for Computational Linguistics.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do NLP Models Know Numbers? Probing Numeracy in Embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5307–5315, Hong Kong, China. Association for Computational Linguistics.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, and Hongfang Liu. 2018. Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*, 77:34–49.

Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. 2020. A Novel Cascade Binary Tagging Framework for Relational Triple Extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1476–1488, Online. Association for Computational Linguistics.

Gerhard Weikum. 2020. Entities with Quantities. *Bulletin of the Technical Committee on Data Engineering*, 43(1):4–8.

Ralph Weischedel and Elizabeth Boschee. 2018. Last Words: What Can Be Accomplished with the State of the Art in Information Extraction? A Personal View. *Computational Linguistics*, 44(4):651–658.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, El-Bachouti, Mohammed, Robert Belvin, and Ann Houston. 2013. OntoNotes Release 5.0.

Minji Wu and Amélie Marian. 2007. Corroborating Answers from Multiple Web Sources. In *Tenth International Workshop on the Web and Databases*, page 6, Beijing, China.

Sen Wu, Luke Hsiao, Xiao Cheng, Braden Hancock, Theodoros Rekatsinas, Philip Levis, and Christopher Ré. 2018. Fonduer: Knowledge Base Construction from Richly Formatted Data. *Proceedings of the 2018 International Conference on Management of Data*, pages 1301–1316.

Lemin Xiao, Kaizhi Tang, Xiong Liu, Hui Yang, Zheng Chen, and Roger Xu. 2013. Information extraction from nanotoxicity related publications. In *2013 IEEE International Conference on Bioinformatics and Biomedicine*, pages 25–30.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.

Engy Yehia, Hussein Boshnak, Sayed AbdelGaber, Amany Abdo, and Doaa S. Elzanfaly. 2019. Ontology-based clinical information extraction from physician's free-text notes. *Journal of Biomedical Informatics*, 98:103276.

Wen-wai Yim, Tyler Denman, Sharon W. Kwan, and Meliha Yetisgen. 2016. Tumor information extraction in radiology reports for hepatocellular carcinoma patients. *AMIA Summits on Translational Science Proceedings*, 2016:455–464.

Minoru Yoshida and Kenji Kita. 2021. *Mining Numbers in Text: A Survey*. IntechOpen.

Dongxiang Zhang, Lei Wang, Luming Zhang, Bing Tian Dai, and Heng Tao Shen. 2020a. The Gap of Semantic Parsing: A Survey on Automatic Math Word Problem Solvers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(9):2287–2305.

Jiansong Zhang and Nora M. El-Gohary. 2016. Semantic NLP-Based Information Extraction from Construction Regulatory Documents for Automated Compliance Checking. *Journal of Computing in Civil Engineering*, 30(2):04015014.

Muhan Zhang and Yixin Chen. 2018. Link Prediction Based on Graph Neural Networks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Xikun Zhang, Deepak Ramachandran, Ian Tenney, Yanai Elazar, and Dan Roth. 2020b. Do Language Embeddings capture Scales? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4889–4896, Online. Association for Computational Linguistics.

Guineng Zheng, Subhabrata Mukherjee, Xin Luna Dong, and Feifei Li. 2018. OpenTag: Open Attribute Value Extraction from Product Profiles. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pages 1049–1058, New York, NY, USA. Association for Computing Machinery.

Qunzhi Zhou, Zhe Wu, Jon Degenhardt, Ethan Hart, Petar Ristoski, Aritra Mandal, Julie Netzloff, and Anu Mandalam. 2021. Leveraging Knowledge Graph and DeepNER to Improve UoM Handling in Search. In *Proceedings of the {ISWC} 2021 Posters, Demos and Industry Tracks: From Novel Ideas to Industrial Practice Co-Located with 20th International Semantic Web Conference*, volume CEUR Workshop Proceedings of *2980*, page 2, Virtual Conference. CEUR-WS.org.

A Related work that is not considered

Systems whose scope is too narrow or offset are not considered in this review. The pure identification of units is not considered measurement extraction. Shbita et al. (2019) parses unit strings into a structured semantic representation using the QUDT ontology (Ray, 2011) and thereby allows the transformation of (compound) units. Zhou et al. (2021) perform entity linking of units in text against a knowledge graph. They also extract numbers but do not elaborate on it. Furthermore, the identification and parsing of numerals (e.g., Paulheim, 2017; Chen et al., 2019b) is not considered. Not all numerals are part of measurements, for example, ordinal numbers (e.g.,. *'Fig. 1'*), or nominal numbers (e.g., postal codes). Pouran Ben Veyseh et al. (2021) approaches only the relation extraction subtask of MeasEval and is therefore not considered. Moreover, systems targeting other modalities than text are not considered. Subercaze (2017), for example, extracts measurements from Wikipedia infoboxes.

Furthermore, systems that, alongside other information, extract quantities, but do not elaborate on it, are not covered. For example, GATE (Cunningham et al., 2013) has a plugin for tagging measurements, but further information is missing. Ayadi et al. (2020) extract information, including quantities, but without addressing these specifically. Wu and Marian (2007) aggregate numeric results for web search queries without sharing details on their IE system. Quantalyze[7] is a product from max.recall information systems GmbH. It seemed to have poor recall (Hundman and Mattmann, 2017) and supported only a small set of units (Aras et al., 2014). Quantulum[8] is a Python library for the extraction of quantities from text, which is able to perform unit disambiguation based on Wikipedia and GloVe vector representations. Other IE systems extract quantitative information from data sheets but do not elaborate on it (Barkschat, 2014; Hsiao et al., 2020). Many systems have been proposed that extract attribute values from product profile pages, but do so without specifically discussing numeric values (Qiu et al., 2015; Zheng et al., 2018; Rezk et al., 2019; Dong et al., 2020).

Lastly, we did not consider systems that target related but distinct tasks from measurement extraction such as the identification of text fragments that contain measurements and their contexts (Alonso and Sellam, 2018), the extraction of molar ratios of material compositions (Jensen et al., 2019), the prediction if a numeral in a sentence is a claim or fact (Chen et al., 2020), the classification of numerals in financial tweets into different categories (Chen et al., 2018, 2019c), financial numeral attachment (Chen et al., 2019a, 2021), extracting relation cardinalities (e.g., the cardinality for <Obama, hasChildren> is two, as he has two children; Mirza et al., 2017) and generating descriptions of quantities that put them into relation with other quantities (e.g., *"about twice the median income for a year"* given a sentence that contains the quantity *'100 000 $'*; Chaganty and Liang, 2016). Spokoyny et al. (2022) argue that language models should jointly reason about numbers and units to learn good representations of measurements and propose the task of masked measurement prediction.

B Overview of the considered systems

This review covers 80 publications that disclose systems for measurement extraction. To provide an overview, Figure B1 depicts these publications in a citation graph and Table B1 summarizes the scopes of the corresponding systems. Table B2 and B3 give an overview of the methods employed in the rule-based and machine learning systems, respectively. The system characterizations are based on the authors' interpretation of the respective scientific publications accompanying the systems. It should be noted that we also include systems that perform measurement extraction but have a different primary purpose (e.g., automated compliance checking). We do not distinguish between hybrid and machine-learning systems, as rules are often employed at some stage of a learning-based system and authors tend to under-report them (Chiticariu et al., 2013). Furthermore, we do not regard an otherwise rule-based system as a learning-based one if for a subtask an existing learning-based model is used without updating its weights (e.g., for POS or NER tagging). The category of diverse web sources includes, i.a., newspaper, tweets and Wikipedia articles. The category of regulatory documents includes decision summaries by the U.S. Food and Drug Administration, construction regulatory documents and financial business reports.

[7] https://www.quantalyze.com/

[8] https://github.com/nielstron/quantulum3

Citation graph. Figure B1 arranges the publications that describe systems for measurement extraction in a citation graph. We used Grobid[9] (Lopez, 2009) to detect the references within PDF files and queried bibliographic APIs (Semantic Scholar[10] and OpenCitations[11]) for citation data. Subsequently, the citation network was created by aggregating the information from all sources. The code for generating the citation graph is published under an open-source license at https://github.com/FZJ-IEK3-VSA/citation-graph-builder.

Scope definitions. In Table B1, the respective scopes are set to fully fulfilled if potentially all quantities, measured entities, etc. are considered by a system or the number of classes is very high. The scope is deemed partially fulfilled if only a small set of quantities (e.g., only scalar values), measured properties (e.g., only left ventricular ejection fraction) and so forth is considered. Quantity normalization is considered fully fulfilled if the identified quantities are converted into a canonical form (e.g., into the respective SI units). Quantity normalization is considered partially fulfilled if the unit and value are obvious from the quantity identification patterns or if operations, such as creating a chart, are performed on the quantities. If quantity extraction is performed via patterns, unit extraction is assumed to be in scope, as the required information is already inherent to the patterns. Similarly, if only measured entities or properties of a small set of concepts are considered, entity linking is assumed to be within the scope, as the concepts are known beforehand.

[9] https://github.com/kermitt2/grobid
[10] https://www.semanticscholar.org/product/api
[11] https://opencitations.net/querying

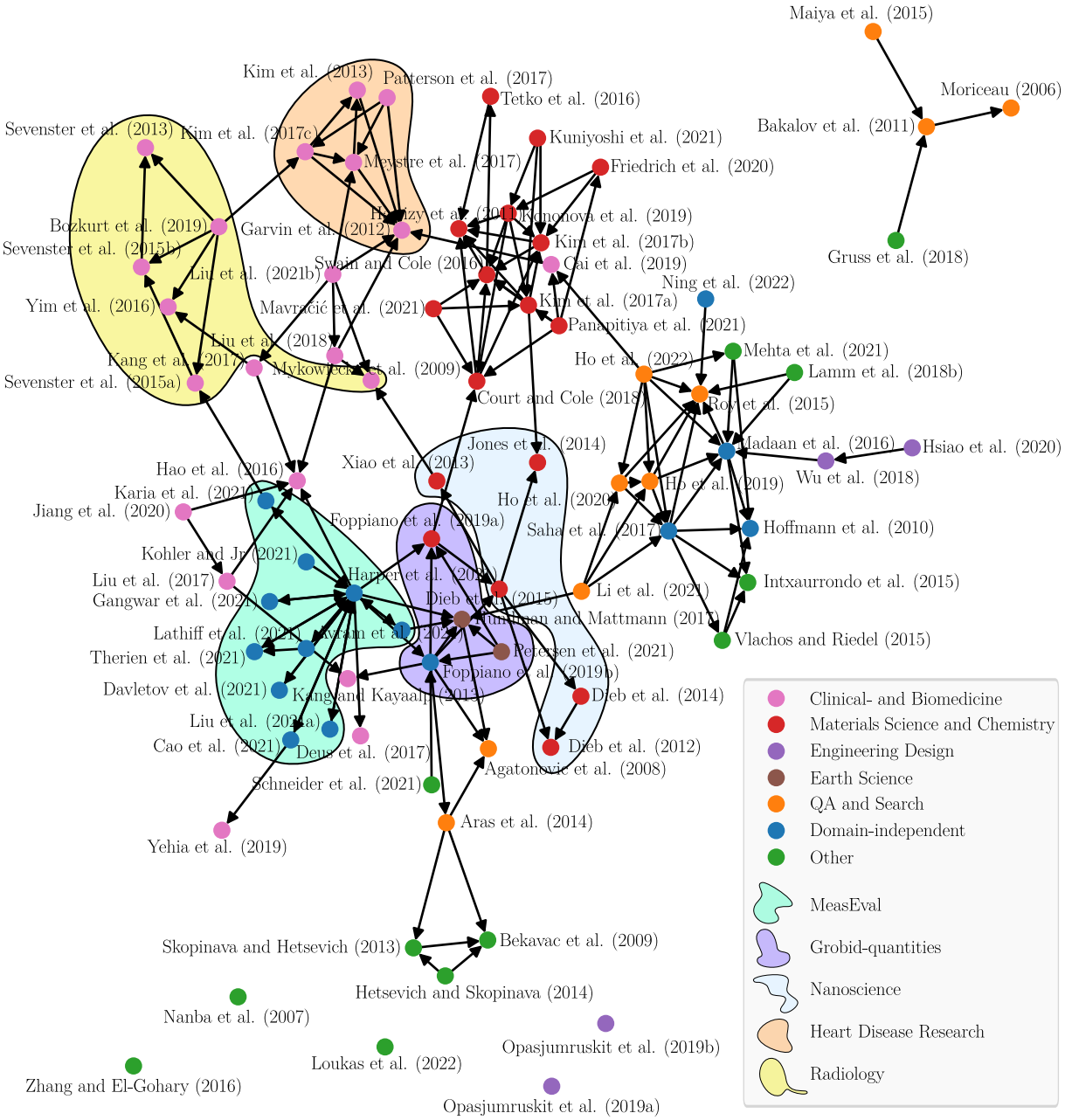

Figure B1: A citation graph of publications describing systems for measurement extraction. Each node represents a publication and the directed edges represent citations to other publications. Note that only citations within the considered set of publications are shown. The application domains envisioned are represented by the color of the node. The allocation to subdomains, Grobid-quantities and MeasEval is highlighted by the colored areal clusters.

Approach	Doc. Genre	Paper	System Name	Quantity Extraction: Scope	Quantity Extraction: Normalization	Meas. Entity Extraction: Scope	Meas. Entity Extraction: Entity Linking	Meas. Property Extraction: Scope	Meas. Property Extraction: Entity Linking	Qualifier/ Context Extr.: Scope	Unit Extr.: Scope	Quantity Modifier Extr.: Scope
Learning-based	Clinical documents	Liu et al. (2021b)		●	○	●	○	○	○	○	●	◐
		Bozkurt et al. (2019)		◐	◐	◐	○	○	○	◐	◐	○
		Liu et al. (2018)		●	○	◐	○	◁	◁	○	●	◐
		Kang et al. (2017)	EliIE	●	○	◐	●	○	○	◐	○	○
		Kim et al. (2017c)	TUCP, TUCP+Pred.	◐	○	○	○	◐	●	○	○	○
		Meystre et al. (2017)	CHIEF ADAHF	◐	○	○	○	◐	●	○	○	○
		Yim et al. (2016)		◐	◐	◐	○	○	○	◐	◐	○
		Kim et al. (2013)	CHIEF EF	◐	○	○	○	◐	●	○	○	○
	Diverse web sources	Ho et al. (2022)	QL	●	●	●	●	◐	N/A	○	●	○
		Ning et al. (2022)		◐	○	◐	●	◁	◁	◐	○	○
		Li et al. (2021)	AnaSearch	N/A	◐	●	●	◁	◁	◐	N/A	◐
		Ho et al. (2019, 2020)	Qsearch	●	●	●	●	▷	○	●	●	◁
		Gruss et al. (2018)		●	●	○	○	◐	●	○	◐	◐
		Lamm et al. (2018b)		●	◐	●	○	◁	◁	●	●	◐
		Saha et al. (2017)	BONIE	●	●	●	○[t]	●	○[t]	○	●	◐
		Madaan et al. (2016)	NumberTron	N/A	●	●	N/A	◐	●	○	●	○
		Intxaurrondo et al. (2015)		◐	◐	◐	●	◐	●	◐	◐	○
		Roy et al. (2015)	Illinois Quantifier	●	●	○	○	○	○	○	●	◐
		Vlachos and Riedel (2015)		◐	◐	◐	●	◐	●	○	○	○
		Hoffmann et al. (2010)	LUCHS	●	○	●	○	●	○	○	○	○
	Reg. doc.	Loukas et al. (2022)		◐	○	○	○	◁	●	○	○	○
	Product descriptions	Mehta et al. (2021)	LaTeX-Numeric	◐	○	○	○	◐	○	○	◐	○
		Opasjumruskit et al. (2019a,b)	ConTrOn	●	N/A	○	○	◐	●	○	N/A	N/A
		Wu et al. (2018); Hsiao et al. (2020)		◐	○	◐	○	◐	●	○	○	○
		Bakalov et al. (2011)	SCAD	◐	●	N/A	N/A	◐	●	○	◐	○
	Scientific publications	Avram et al. (2021)[M]	UPB	●	○	●	○	●	○	●	●	◐
		Cao et al. (2021)[M]	CONNER	●	○	●	○	●	○	●	●	◐
		Davletov et al. (2021)[M]	LIORI	●	○	●	○	●	○	●	●	◐
		Gangwar et al. (2021)[M]	Counts@IITK	●	○	●	○	●	○	●	●	◐
		Harper et al. (2021)[M]	MeasEval Baseline 1	●	○	●	○	●	○	●	●	○
		Karia et al. (2021)[M]	KGP	●	○	●	○	●	○	●	●	◐
		Kohler and Jr (2021)[M]	GPT-3	●	○	●	○	●	○	○	●	○
		Kuniyoshi et al. (2021)		●	●	○	○	◐	●	◐	◐	◐
		Lathiff et al. (2021)[M]	CLaC-np	●	○	●	○	●	○	●	●	◐
		Liu et al. (2021a)[M]	Stanford MLab	●	○	○	○	○	○	○	●	◐
		Mavračić et al. (2021)	ChemDataExtr... 2.0	●	●	◐	N/A	◐	●	◐	◐	◐
		Panapitiya et al. (2021)		◐	○	◐	○	◐	●	○	◐	○
		Petersen et al. (2021)[G]	Geo-Quantities	●	●	●	○	○	○	◐	●	◐
		Therien et al. (2021)[M]	CLaC-BP	●	○	●	○	●	○	●	●	◐
		Friedrich et al. (2020)		◐	○	◐	○	◐	●	◐	○	○
		Foppiano et al. (2019a)[G]		◐	○	◐	N/A	◐	●	○	●	◐
		Foppiano et al. (2019b)[G,*]	Grobid-quantities	●	●	●[a]	○	◁	◁	○	●	◐
		Kononova et al. (2019)		◐	◐	N/A	N/A	◐	●	◐	N/A	◐
		Hundman and Mattmann (2017)[G]	Marve	●	●	●	○	◁	◁	◁	●	◐
		Kim et al. (2017b,a)		◐	◐	◐	N/A	◐	N/A	◐	◐	○
		Swain and Cole (2016)	ChemDataExtractor	◐	◐	◐	N/A	◐	●	◐	◐	○
		Dieb et al. (2014, 2015)	NaDevEx	◐	○	○	○	◐	○	◐	○	○
		Dieb et al. (2012)		◐	○	○	○	◐	○	◐	○	○
Rule-based	Clinical documents	Jiang et al. (2020)	LATTE	◐	●	◐	●	◐	●	◐	◐	○
		Cai et al. (2019)	EXTEND	◐	◐	◐	●	◁	◁	○	◐	◐
		Yehia et al. (2019)		◐	○	◐	●	◁	◁	◐	○	○
		Liu et al. (2017)		◐	◐	◐	●	◁	◁	◐	N/A	◐
		Patterson et al. (2017)	EchoExtractor	◐	◐	◐	●	◁	◁	○	◐	◐
		Hao et al. (2016)	Valx	◐	●	○	○	◐	●	○	●	◐
		Sevenster et al. (2013, 2015b,a)		◐	●	○	○	○	○	◐	◐	○
		Garvin et al. (2012)	CUIMANDREef	◐	○	○	○	◐	●	◐	◐	◐
		Mykowiecka et al. (2009)		◐	N/A	◐	●	◐	●	◐	◐	◐
	Diverse web sources	Madaan et al. (2016)	NumberRule	N/A	○	●	○	○	●	○	●	○
		Skopinava and Hetsevich (2013)		●	◐	○	○	○	○	○	◐	○
		Nanba et al. (2007)		◐	◐	◐	○	◁	◁	◐	◐[i]	○
		Moriceau (2006)	QRISTAL	●	●	◐	N/A	◁	◁	●	●	◐
		Bekavac et al. (2009)		◐	○	○	○	○	○	○	◐	◐
	Patents	Tetko et al. (2016)		◐	●	◐	○	◐	●	◐	◐	◐
		Aras et al. (2014)		●	●	○	○	○	○	○	●	◐
		Agatonovic et al. (2008)		◐	○	○	○	○	○	○	◐	◐
	Regulatory documents	Zhang and El-Gohary (2016)		◐	○	◐	●	◐	N/A	◐	◐	◐
		Kang and Kayaalp (2013)		◐	○	◐	N/A	◐	N/A	○	◐	○
	Scientific publications	Schneider et al. (2021)		◐	○	◐	●	○	○	○	○	○
		Deus et al. (2017)		●	●	○	○	○	○	○	◐	○
		Maiya et al. (2015)[*]	MQSearch	◐	○	○	○	●	○	○	●	○
		Jones et al. (2014)	NanoSifter	◐	◐	○	○	◐	●	○	◐	◐
		Xiao et al. (2013)		◐	◐	◐	●	◐	●	◐	◐	◐
		Hawizy et al. (2011)	ChemicalTagger	◐	◐	◐	●	◐	N/A	◐	◐	○

● = Fully fulfilled; ◐ = Partially fulfilled; ○ = Not fulfilled; ▷(◁) = Mixed with subtask on the right (left); N/A = Aspect not evident to the authors;
[†]Ensemble; [*]Also targeted at technical documents; [M]Part of MeasEval; [G]Related to Grobid-quantities; [a]Experimental feature;
[t]Matching against a KB only during seed fact generation; [i]The unit and measured entity are input to the system

Table B1: The scopes of systems for measurement extraction with regard to different subtasks.

		Quantity Extraction (QE)	Meas. Entity Extraction (MEE)	Meas. Property Extr. (MPE)	Qualifier or Context Extraction	Unit of Measurement Extraction	Quantity Modifier Extraction	Grouping or Relation Extraction
Clinical documents	Jiang et al. (2020) (LATTE)	REGEX	Rules & dict.-matching	Dict.-matching	Keywords & rules	Dict.-matching	–	Prox. heuristic
	Cai et al. (2019) (EXTEND)	Patterns	Patterns & dict.-matching	Mixed with MEE	–	Rules/patterns	Patterns	Prox. heuristic
	Yehia et al. (2019)	Dict.-matching of noun phrases	Dict.-matching of noun phrases	Mixed with MEE	Dict.-matching of noun phrases	–	–	Prox. heuristic, sentence clf., ontology
	Liu et al. (2017)	REGEX & temporal expression tagger	REGEX & domain knowl. from Valx	Mixed with MEE	Temporal expression tagger	Implicit by rules for QE	REGEX	Prox. heuristic
	Patterson et al. (2017) (EchoExtractor)	REGEX & semantic patterns	Dict.-matching, REGEX	Mixed with MEE	–	Implicit by rules for QE	Implicit by rules for QE	Patterns
	Hao et al. (2016) (Valx)	REGEX	–	Patterns & dicts	–	Dict.-matching & rules	–	Patterns
	Sevenster et al. (2013, 2015b,a)	REGEX	–	–	Max. entropy classifier	Implicit by rules for QE	–	–
	Garvin et al. (2012) (CUIMANDREef)	Rules/patterns	–	Rules/patterns	Rules/patterns	Implicit by rules for QE	Rules/patterns	Rules/patterns
	Mykowiecka et al. (2009)	Rules/patterns (ontology, dict, morph.)	Rules/patterns (ontology, dict, morph.)	Rules/patterns (ontology, dict, morph.)	Rules/patterns (ontology, dict, morph.)	Rules/patterns (ontology, dict, morph.)	Rules/patterns (ontology, dict, morph.)	Rules & prox. heuristic
Diverse web sources	Madaan et al. (2016) (NumberRule)	UnitTagger[u]	NER tagger	Keywords & filter rules	–	UnitTagger[u]	–	Dep. tree analysis & prox. heuristic
	Skopinava and Hetsevich (2013); Hetsevich and Skopinava (2014)	FSA-based grammar	–	–	–	Inherent to QE patterns	–	Inherent to patterns
	Bekavac et al. (2009)	REGEX & lexical resources	–	–	–	REGEX & lexical resources	REGEX & lexical resources	Inherent to patterns
	Nanba et al. (2007)	Patterns	Keyword-matching	Mixed with MEE	Statistical syntactic parser (CaboCha)	Patterns	–	Dep. tree analysis & cue phrases & prox. heuristic
	Moriceau (2006) (QRISTAL)	Rules/patterns	Rules/patterns	Mixed with MEE	Rules/patterns	Rules/patterns	Rules/patterns	Probably inherent to patterns & prox. heuristic
Patents	Tetko et al. (2016)	LeadMine[f] (based on FSA & dict) & additional dict	LeadMine[f] (based on FSA & dict)	LeadMine[f] (based on FSA & dict)	LeadMine[f] (based on FSA & dict)	Probably based on LeadMine[f] grammars	LeadMine[f] (based on FSA & dict)	ChemicalTagger & LeadMine[f] (both based on grammars)
	Aras et al. (2014)	FSA-based grammar	–	–	–	Probably dict.-matching	Implicit by rules for QE	–
	Agatonovic et al. (2008)	FSA-based grammar	–	–	–	Dict.-matching	Implicit by rules for QE	–
Regulatory doc.	Zhang and El-Gohary (2016)	Rules/patterns (ontology, dict, grammar & POS)	Rules/patterns (ontology, dict, grammar & POS)	Rules/patterns (ontology, dict, grammar & POS)	Rules/patterns (ontology, dict, grammar & POS)	Rules/patterns (ontology, dict, grammar & POS)	Rules/patterns (ontology, dict, grammar & POS)	Prox. heuristic & patterns
	Kang and Kayaalp (2013)	REGEX, trigger events	Dict.-matching, lexical rules, trigger events	Dict.-matching, lexical rules, trigger events	–	Dict.-matching, lexical rules, trigger events	–	Inherent to patterns
Scientific publications	Schneider et al. (2021)	Rules	Dict.-matching	–	–	–	–	Prox. heuristic
	Deus et al. (2017)	Patterns (POS)	–	–	–	Patterns (POS)	–	–
	Maiya et al. (2015)[*] (MQSearch)	REGEX	–	Patterns (POS, custom tags)	–	Unit ontology & rules	–	Inherent to patterns
	Jones et al. (2014) (NanoSifter)	REGEX	–	Dict.-matching	–	Inherent to QE patterns	Inherent to QE patterns	Prox. heuristic
	Xiao et al. (2013)	REGEX	Dict.-matching	Keyword-matching	Dict.-matching	Inherent to QE patterns	Inherent to QE patterns	Prox. heuristic & REGEX
	Hawizy et al. (2011) (ChemicalTagger)	Phrase parsing (POS, REGEX tags)	Chemical entity tagger & REGEX	Probably inherent to QE patterns	Chemical entity tagger & REGEX	REGEX	–	Rule-based creation of parse tree

= Rules or patterns; = Dictionary-, gazetteer-, keyword- or ontology-matching; = Constituency or dependency parse tree analysis; = External model;
= Concepts are related based on proximity heuristics; = Not distinguished between subtasks; = Subtask is indirectly fulfilled;
N/A = Aspect not evident to the authors; [u]UnitTagger (Sarawagi and Chakrabarti, 2014); [f]LeadMine (Lowe and Sayle, 2015);
[w]Word2Vec (Mikolov et al., 2013); [c]CaboCha (Kudo and Matsumoto, 2003); [*]Also targeted at technical documents;
Abbreviations: clf. = classification; prox. = proximity; POS = Part-Of-Speech tags; REGEX = REGular EXpressions; FSA = Finite State Automata

Table B2: The methods of the rule-based approaches to measurement extraction per subtask.

		Quantity Extraction (QE)	Meas. Entity Extraction (MEE)	Meas. Property Extr. (MPE)	Qualifier or Context Extraction	Unit of Measurement Extraction	Quantity Modifier Extraction	Grouping or Relation Extraction
Clinical documents	Liu et al. (2021b)	Char.-level BiLSTM-CRF	Char.-level BiLSTM-CRF	–	–	Char.-level BiLSTM-CRF	Dict & rules	Prox. heuristics & random forest
	Bozkurt et al. (2019)	REGEX	CRF (POS, dict)	–	REGEX & CRF (POS, dict)	REGEX	–	Probably prox. heuristics
	Liu et al. (2018)	Learned patterns & REGEX & dict	Learned patterns & dict.-matching	Mixed with MEE	–	Learned patterns & dict.-matching	Learned patterns & Rules	Inherent to patterns & prox. heuristic
	Kang et al. (2017) (EliIE)	CRF	CRF	–	CRF	–	–	SVM on relation pairs
	Kim et al. (2017c) (TUCP, TUCP+Prediction)	MIRA (i.a. pred. of multiple models)	–	MIRA (i.a. pred. of multiple models)	–	–	–	–
	Meystre et al. (2017) (CHIEF ADAHF)	MIRA (i.a. pred of CUIMANDREef)	–	MIRA (i.a. pred of CUIMANDREef)	–	–	–	–
	Yim et al. (2016)	REGEX	CRF	–	CRF	Inherent to QE rules	–	Max. entropy classifier
	Kim et al. (2013) (CHIEF EF)	MIRA (i.a. pred of CUIMANDREef)	–	MIRA (i.a. pred of CUIMANDREef)	–	–	–	–
Diverse web sources	Ho et al. (2022) (QL)	Illinois Quantifier	Open IE[o] & coref. res. & entity linker	Semantic distance between input prop. and Open IE[o] prop.	–	Illinois Quantifier	–	Ranking & filtering of candidates from Open IE[o]
	Ning et al. (2022)	BERT (large, cased)	T5-large	Mixed with MEE	T5-large	–	–	Inherent to sequential approach
	Li et al. (2021) (AnaSearch)	Recognizers-Text[r]	Constituency parse tree & Text Analytics API[z]	Mixed with MEE	Constituency parse tree & rules & Text Analytics API[z]	Recognizers-Text[r]	Recognizers-Text[r]	Inherent to MEE & qualifier extraction
	Ho et al. (2019, 2020) (Qsearch)	Illinois Quantifier & rules	NER tagger	Mixed with Qualifier extraction	BiLSTM	Illinois Quantifier & rules	Mixed with Qualifier extraction	Inferred from rel.-specific tagging
	Gruss et al. (2018)	REGEX	–	Naïve Bayes classifiers	–	Dict & rules	Probably REGEX	Inherent to framing task as quantity span clf.
	Lamm et al. (2018b)	CNN+CRF	CNN+CRF	Mixed with MEE	CNN+CRF	Patterns	CNN+CRF	Rel. embeddings & shallow NN
	Saha et al. (2017) (BONIE)	Illinois Quantifier	Bootstrapping to learn dep. patterns	Bootstrapping to learn dep. patterns & UnitTagger[u] for implicit rel.	–	Illinois Quantifier	–	Inherent to patterns
	Madaan et al. (2016) (NumberTron)	UnitTagger[u]	NER tagger	Graphical model[h] (keywords, dep. tree, POS, num. features)	–	UnitTagger[u]	–	Inherent to MPE
	Intxaurrondo et al. (2015)	CRF & noisy-or[n]	Inherent to semantic frame	Inherent to semantic frame	CRF & noisy-or[n]	Inherent to semantic frame	–	Inherent to modeling task as slot-filling problem
	Roy et al. (2015) (Illinois Quantifier)	Bank of classifiers	–	–	–	Rule-based & coref. res. & SRL	list of phrases	Coref. res. & SRL if unit not in quantity span
	Vlachos and Riedel (2015)	Learned patterns	Learned patterns	Learned patterns (ranking by deviation to value in KB)	–	–	–	Inherent to patterns
	Hoffmann et al. (2010) (LUCHS)	Rel.-spec. CRF extractors (learned lexicons)	Rel.-spec. CRF extractors (learned lexicons)	Rel.-spec. CRF extractors (learned lexicons)	–	–	–	Inferred from rel.-specific tagging
Reg. doc.	Loukas et al. (2022)	BERT trained from scratch	–	Mixed with QE	–	–	–	Inherent to framing QE as entity typing
Product descriptions	Mehta et al. (2021) (LaTeX-Numeric)	BiLSTM-CNN-CRF	–	Inherent to QE (property spec. BIO labels)	–	BiLSTM-CNN-CRF[l]	–	Inherent to QE & MPE
	Opasjumruskit et al. (2019a,b) (ConTrOn)	Rules/patterns (learning by user feedback)	–	Keyword-search (ontology, user feedback)	–	Rules/patterns	N/A	Patterns & prox. heuristic
	Wu et al. (2018); Hsiao et al. (2020)	Rule-based candidate generation	Rule-based candidate generation	Log. regression (doc. structure, tables, images)	–	Rule-based candidate generation	–	Inherent to property relation clf.
	Bakalov et al. (2011) (SCAD)	Probably patterns & dict	N/A	Multi-class log. regression & int. linear program	–	Patterns & dict	–	Inherent to MPE

Continued on the next page...

= Rules or patterns; = Dictionary-, gazetteer-, keyword- or ontology-matching; = Constituency or dependency parse tree analysis; = External model; = Concepts are related based on proximity heuristics; = Not distinguished between subtasks; = Subtask is indirectly fulfilled; = CRF-based model; = BiLSTM-based model; = Transformer-based model; N/A = Aspect not evident to the authors;

[u]UnitTagger (Sarawagi and Chakrabarti, 2014); [n]noisy-or (Surdeanu et al., 2012); [r]https://github.com/microsoft/Recognizers-Text;

[h]Based on MultiR (Hoffmann et al., 2011); [z]https://azure.microsoft.com/en-us/services/cognitive-services/text-analytics;

[o]Open IE (Saha et al., 2017; Saha and Mausam, 2018); [*]Also targeted at technical documents;

[l]LaTeX-Numeric uses a unit list for matching in a distant supervision setting, however, only during training;

Abbreviations: MIRA = Margin-Infused Relaxed Algorithm; KB = Knowledge Base; rel. = relation; clf. = classification; prox. = proximity; char. = character; coref. res. = coreference resolution; SRL = Semantic Role Labeling; NN = Neural Network; CNN = Convolutional Neural Network; POS = Part-Of-Speech tags; SVM = Support Vector Machine; REGEX = REGular EXpressions

		Quantity Extraction (QE)	Meas. Entity Extraction (MEE)	Meas. Property Extr. (MPE)	Qualifier or Context Extraction	Unit of Measurement Extraction	Quantity Modifier Extraction	Grouping or Relation Extraction
Scientific publications	Avram et al. (2021)[M] (UPB)	RoBERTa+CRF	RoBERTa (QA)	RoBERTa (QA)	RoBERTa (QA)	Char.-level BiLSTM	Char.-level BiLSTM	Inferred from rel.-specific tagging
	Cao et al. (2021)[M] (CONNER / jarvis@tencent)	RoBERTa enc. & PointerNet[†] & CRF	RoBERTa enc. & rel.-spec. tagger	RoBERTa enc. & rel.-spec. tagger	RoBERTa enc. & rel.-spec. tagger	Rules	RoBERTa enc. & plain classifier	Inferred from rel.-specific tagging
	Davletov et al. (2021)[M] (LIORI)	LUKE NER[†]	RoBERTa (QA)	RoBERTa (QA)	RoBERTa (QA)	RoBERTa (QA)	RoBERTa (QA)	Inferred from rel.-specific tagging
	Gangwar et al. (2021)[M] (Counts@IITK)	SciBERT+CRF	SciBERT+CRF	2nd SciBERT+CRF	3rd SciBERT+CRF	Char.-level BiLSTM	BERT	Inferred from rel.-specific tagging
	Harper et al. (2021)[M] (MeasEval Baseline 1)	spaCy NER[s]	spaCy NER[s]	spaCy NER[s]	spaCy NER[s]	Dict.-matching	–	Prox. heuristics
	Karia et al. (2021)[M] (KGP)	BioBERT	2nd BioBERT	2nd BioBERT	2nd BioBERT	Dict.-matching	Keywords & rules	Inferred from rel.-specific tagging
	Kohler and Jr (2021)[M] (GPT-3)	GPT-3 (few-shot)	GPT-3 (few-shot)	GPT-3 (few-shot)	–	GPT-3 (few-shot)	–	Inherent to prompt
	Kuniyoshi et al. (2021)	ELMo for materials synthesis[e] & rules	–	Inherent to framing QE as entity typing	ELMo for materials synthesis[e]	Rules	Inherent to normalization rules	–
	Lathiff et al. (2021)[M] (CLaC-np)	DGCNN on dep. parse trees	DGCNN on dep. parse trees	DGCNN on dep. parse trees	DGCNN on dep. parse trees	Dict.-matching	SciBERT (from Therien et al., 2021)	Inferred from rel.-specific tagging
	Liu et al. (2021a)[M] (Stanford MLab)	BERT large	–	–	–	BERT large	Multi-label clf.	–
	Mavračić et al. (2021); Court and Cole (2018) (ChemDataExtractor 2.0)	Snowball pattern learning	Probably chem. NER tagger (CRF, dicts & REGEX)	Snowball pattern learning	Snowball pattern learning	Snowball pattern learning	N/A	Inherent to phrase parsing grammars
	Panapitiya et al. (2021)	SciBERT+CRF	SciBERT+CRF	Inherent to tagging ME & Q for one spec. property	–	SciBERT+CRF	–	–
	Petersen et al. (2021)[G] (Geo-Quantities)	Grobid-quantities	Grobid-quantities	–	CRF	Grobid-quantities	Grobid-quantities	Prox. heuristics
	Therien et al. (2021)[M] (CLaC-BP)	SciBERT	SciBERT	SciBERT	SciBERT	Rules	2nd SciBERT	Prox. heuristics & rules
	Friedrich et al. (2020)	BiLSTM or SciBERT	Inherent to semantic frame?	BiLSTM or SciBERT	BiLSTM or SciBERT	–	–	Inherent to modeling task as slot-filling problem
	Foppiano et al. (2019a)[G]	Grobid-quantities	CRF & chem. NER tagger	Quantity clf. with dict.-matching	–	Grobid-quantities	Grobid-quantities	Prox. heuristics
	Foppiano et al. (2019b)[G,*] (Grobid-quantities)	CRF & unit dict	Dep. tree analysis (later CRF)[a]	Mixed with MEE	–	Char.-level CRF & unit dict	CRF	Inherent to MEE
	Kononova et al. (2019)	REGEX & dict.-matching	–	Implicit by rules for QE	NN (dep. tree, Word2Vec[w])	Probably implicit by rules for QE	Probably implicit by rules for QE	Dep. tree analysis
	Hundman and Mattmann (2017)[G] (Marve)	Grobid-quantities	Patterns (POS, dep. tree)	Mixed with MEE	Mixed with MEE	Grobid-quantities	Grobid-quantities & Mixed with MEE	Inherent to patterns
	Kim et al. (2017b,a)	NN, dict & DB matching & existing models	NN, dict & DB matching & existing models	NN, dict & DB matching & existing models	NN, dict & DB matching & existing models	NN, dict & DB matching & existing models	–	Dep. tree analysis, prox. heuristics
	Swain and Cole (2016) (ChemDataExtractor)	Rule-based phrase parsing (NER, POS)	Chem. NER tagger (CRF, REGEX & dict.-matching)	Rule-based phrase parsing (NER, POS)	Rule-based phrase parsing (NER, POS)	Rule-based phrase parsing (NER, POS)	–	Inherent to phrase parsing grammars
	Dieb et al. (2014, 2015) (NaDevEx)	CRF (POS, REGEX, unit dict)	CRF (POS, REGEX, chem. NER tagger)	CRF (POS, REGEX, dict, SVM clf.)	–	–	–	–
	Dieb et al. (2012)	YamCha[y] (POS, REGEX)	YamCha[y] (POS, REGEX, chem. NER tagger)	YamCha[y] (POS, REGEX)	–	–	–	–

= Rules or patterns; = Dictionary-, gazetteer-, keyword- or ontology-matching; = Constituency or dependency parse tree analysis; = External model; = Concepts are related based on proximity heuristics; = Not distinguished between subtasks; = Subtask is indirectly fulfilled; = CRF-based model; = BiLSTM-based model; = Transformer-based model; N/A = Aspect not evident to the authors;

[†]Ensemble; [M]Part of MeasEval; [G]Related to Grobid-quantities; [s]Spacy (Montani et al., 2022);

[e]ELMo for materials synthesis (Kim et al., 2020); [y]http://chasen.org/~taku/software/yamcha/; [*]Also targeted at technical documents;

[a]Experimental feature described at https://grobid-quantities.readthedocs.io/en/latest/guidelines.html;

Abbreviations: DB = Database; QA = Question Answering; enc. = encoder; rel. = relation; clf. = classification; prox. = proximity; char. = character; NN = Neural Network; DGCNN = Deep Graph Convolution Neural Network; POS = Part-Of-Speech tags; SVM = Support Vector Machine; REGEX = REGular EXpressions

Table B3: The methods of the learning-based approaches to measurement extraction per subtask. Note that this table starts on the previous page.