Measurement Extraction with Natural Language Processing: A Review

Jan Göpfert\textsuperscript{1,2,*} and Patrick Kuckertz\textsuperscript{1} and Jann M. Weinand\textsuperscript{1} and Leander Kotzur\textsuperscript{1,2}

\textsuperscript{1}Institute of Energy and Climate Research, Techno-economic Systems Analysis (IEK-3), Forschungszentrum Jülich, 52425 Jülich, Germany
\textsuperscript{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).

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...
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 (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 (Maity 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 (Moricelou, 2006; Roy et al., 2015; Lamm et al., 2018b). In MeasEval, phrases that indicate change are regarded as qualifiers. 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).

**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.

---

1https://github.com/harperco/MeasEval/tree/main/annotationGuidelines
### 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.e., 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 sub-word 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 extraction.

\(^2\)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
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. Hereinafter, measured entity, property and qualifier spans had to be identified. Finally, relations between the identified spans had to be extracted. 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).
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’ $\rightarrow 21^\circ C$; 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 $\frac{6}{10}$, respectively” are common (Deus et al., 2017) and measurement extraction might include parsing formulaic expressions (e.g., “\(\nu(29) = -1.85, p = 0.074^*\)”; 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 m$^2$’ 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$\text{ voltaic}$’.

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; Intxaurreondo 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


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


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, Jia Li Ren, Suresh Pal Singh, Fan Xiao, Hui-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.


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 is a product from max.recall and performed 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.

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.

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 is a product from max.recall and performed 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.

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.
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.
<table>
<thead>
<tr>
<th>Scope</th>
<th>Modality</th>
<th>Meas. Property</th>
<th>Qualifier/Unit</th>
<th>Scope</th>
<th>Scope</th>
<th>Scope</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>Full</td>
<td>Entity Meas.</td>
<td>Quantity</td>
<td>Full</td>
<td>Full</td>
<td>Full</td>
</tr>
<tr>
<td>Partial</td>
<td>Partial</td>
<td>Extraction</td>
<td>Extraction</td>
<td>Partial</td>
<td>Partial</td>
<td>Partial</td>
</tr>
<tr>
<td>Not</td>
<td>Not</td>
<td>Extraction</td>
<td>Extraction</td>
<td>Not</td>
<td>Not</td>
<td>Not</td>
</tr>
</tbody>
</table>

**Table B1:** The scopes of systems for measurement extraction with regard to different subtasks.
<table>
<thead>
<tr>
<th>Quantity Extraction</th>
<th>Meas. Entity Extraction (MEE)</th>
<th>Meas. Property Extr. (MPE)</th>
<th>Qualifier or Context Extraction</th>
<th>Unit of Measurement Extraction</th>
<th>Quantity Modifier Extraction</th>
<th>Grouping or Relation Extraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jiang et al. (2020)</td>
<td>REGEX</td>
<td>Rules</td>
<td>Dict.-matching</td>
<td>Keywords</td>
<td>Dict.-matching</td>
<td>Prox. heuristic</td>
</tr>
<tr>
<td>(LATTE)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cai et al. (2019)</td>
<td>Patterns</td>
<td></td>
<td>Mixed with MEE</td>
<td>Rules/patterns</td>
<td>Patents</td>
<td>Prox. heuristic</td>
</tr>
<tr>
<td>(EXTEND)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yehia et al. (2019)</td>
<td>REGEX</td>
<td>Dict.-matching</td>
<td>of noun phrases</td>
<td>Mixed with MEE</td>
<td>Dict.-matching</td>
<td>Prox. heuristic</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liu et al. (2017)</td>
<td>REGEX &amp; dict.-matching</td>
<td></td>
<td>GUI</td>
<td>Temporal expression tagger</td>
<td>Implicit by rules for QE</td>
<td>REGEX</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patterson et al. (2017)</td>
<td>REGEX &amp; dict.-matching</td>
<td></td>
<td>GUI</td>
<td>Temporal expression tagger</td>
<td>Implicit by rules for QE</td>
<td>Patents</td>
</tr>
<tr>
<td>(EchoExtractor)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hao et al. (2016)</td>
<td>REGEX</td>
<td></td>
<td>Rules/patterns</td>
<td>Patents</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(VAlt)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sevrin et al. (2013)</td>
<td>REGEX</td>
<td></td>
<td>Rules/patterns</td>
<td>Patents</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Madaan et al. (2016)</td>
<td>UniTagger†</td>
<td></td>
<td>Rules/patterns</td>
<td>Inherent to QE patterns</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(NaBoMedRule)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Skopinar and</td>
<td>FSA-based grammar</td>
<td></td>
<td>Rules/patterns</td>
<td>Inherent to QE patterns</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Horlevich (2013)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hetanovich and</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Skopinar (2014)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bekar et al. (2009)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nazth et al. (2007)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Morisseau (2006)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teker et al. (2016)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(NASCEPT)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(NASCEPT)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aras et al. (2014)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agatonovic et al. (2008)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zdunek et al. (2016)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(NASCEPT)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kang and Keyuap (2013)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schmidt et al. (2014)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dees et al. (2017)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mayka et al. (2015)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jun et al. (2014)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jones et al. (2014)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NuanShifter</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xiao et al. (2013)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Huwitz et al. (2011)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(ChemicalTagger)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Abbreviations
- REGEX = REGular EXpressions
- FSA = Finite State Automata
- Rules = Rules or patterns
- Dictionary-, gazetteer-, keyword- or ontology-matching
- POS = Part-Of-Speech tags
- MEE = Meas. Entity Extraction
- MPE = Meas. Property Extr.
- QI = Quantity Extraction
- MPE = Meas. Property Extr.
- PWS = Processing step
- N/A = Not applicable
- NA = Not available
- OE = Outlier extraction

### Notes
1. This table is derived from Table 7 in the original paper.
2. The methods are categorized based on their extraction techniques.
3. The leftmost column lists the authors and their methods.
4. The subsequent columns indicate the types of methods used, including REGEX, FSA, and other techniques.
5. The rightmost column identifies the grouping or relation extraction methods.
<table>
<thead>
<tr>
<th>Clinical Domain</th>
<th>Quantity Extraction (QE)</th>
<th>Meas. Entity Extraction (MEE)</th>
<th>Meas. Property Extraction (MPE)</th>
<th>Qualifer or Context Extraction</th>
<th>Unit of Measurement Extraction</th>
<th>Quantity Modifer Extraction</th>
<th>Grouping or Relation Extraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liu et al. (2021b)</td>
<td>Character-level BiLSTM &amp; CNN</td>
<td>–</td>
<td>–</td>
<td>Character-level BiLSTM &amp; CNN</td>
<td>DT &amp; rules</td>
<td>POS, dict</td>
<td>Random forest &amp; SVM, rules &amp; random forest</td>
</tr>
<tr>
<td>Bozkurt et al. (2019)</td>
<td>CNN, POS, dict</td>
<td>–</td>
<td>–</td>
<td>CNN, POS, dict</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Kang et al. (2017)</td>
<td>(Elie)</td>
<td>–</td>
<td>–</td>
<td>(Elie)</td>
<td>–</td>
<td>–</td>
<td>SVM on relation pairs</td>
</tr>
<tr>
<td>Kim et al. (2017a,c)</td>
<td>MIRA (i.a. pred. of multiple models)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Meyeste et al. (2017)</td>
<td>MIRA (i.a. pred. of multiple models)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Yin et al. (2016)</td>
<td>Bootstrapping to learn dep. patterns</td>
<td>–</td>
<td>–</td>
<td>Bootstrapping to learn dep. patterns &amp; UnitTagger for implicit rel.</td>
<td>–</td>
<td>–</td>
<td>Inferred to MPE</td>
</tr>
<tr>
<td>Kim et al. (2013)</td>
<td>(CHEF-EF)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

### Disease Web Resources

| Bier et al. (2022) (QL) | Illinois Quantifier | Open Bi- & entity-linked & query-based & Text Analytics API | Semantic distance between input prop. & Open Bi- & prop. Mixed with MIE | Illinois Quantifier | – | Ranking & filtering of candidates from Open Bi- & Inherent to sequence alignment approach |
| Bier et al. (2022) | CNN & CRF & CNN & CRF | CNN & CRF & CRF & CRF | CNN & CRF & CNN & CRF | CNN & CRF | CNN & CRF | Rel. embeddings & shallow NN |
| Li et al. (2021) | Illinois Quantifier | NE & CRF & Text Analytics API | Constituency parse tree & rules & Text Analytics API | Illinois Quantifier | – | Inferred from rel-term specific tagging & Inherent to ranking task, on quantity span off |
| Ho et al. (2019, 2020) (Qecch) | Illinois Quantifier | NE & CRF | Mixed with Qualification extraction & Text Analytics API | Illinois Quantifier | – | Inferred from rel-term specific tagging & Inherent to ranking task, on quantity span off |
| Ho et al. (2019, 2020) (Qecch) | Illinois Quantifier | NE & CRF | Mixed with Qualification extraction & Text Analytics API | Illinois Quantifier | – | Inferred from rel-term specific tagging & Inherent to ranking task, on quantity span off |
| Loukas et al. (2018b) | CNN & CRF | CNN & CRF | CNN & CRF | CNN & CRF | CNN & CRF | Rel. embeddings & shallow NN |
| Saha et al. (2017) (BONE) | Illinois Quantifier | Bootstrap to learn dep. patterns & UnitTagger for implicit rel. | Bootstrap to learn dep. patterns & UnitTagger for implicit rel. | – | – | Inherent to patterns |
| Li et al. (2021) | Illinois Quantifier | NE & CRF | Graphical model & POS, lexicons | Illinois Quantifier | – | Inherent to MPE |
| Li et al. (2021) | Illinois Quantifier | NE & CRF | Graphical model & POS, lexicons | Illinois Quantifier | – | Inherent to MPE |
| Hoffmann et al. (2010) (LUCHS) | Rel-spec. Ex. extractors (learned lexicon) | Rel-spec. Ex. extractors (learned lexicon) | Rel-spec. Ex. extractors (learned lexicon) | Rel-spec. Ex. extractors (learned lexicon) | Rel-spec. Ex. extractors (learned lexicon) | Inferred from rel-term specific tagging |
| John et al. (2020) | Illinois Quantifier | NE & CRF | Graphical model & POS, lexicons | Illinois Quantifier | – | Inherent to MPE |
| Mehta et al. (2021) (LaTeX-Numeric) | BiLSTM & CNN & CRF & CRF | – | – | – | – | Inherent to QE & MPE |
| Opasannunukul et al. (2019a,b) (ConTrOn) | Rule-paths (learning by user feedback) | – | – | – | – | Patterns |
| Wu et al. (2018) (Hsiao et al. 2020) | Rule-based candidate generation | Log. regression (doc. structure, tables, images) | – | – | – | Inherent to property relation class |
| Bakalov et al. (2011) (SCAD) | Probably patterns & N/A | Multi-class log. regression & int. linear program | – | – | – | Inherent to MPE |

Continued on the next page...
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Aumann et al. (2021)⁷</td>
<td>RoBERTa+</td>
<td>RoBERTa+</td>
<td>RoBERTa+</td>
<td>RoBERTa+</td>
<td>Char-level</td>
<td>Char-level</td>
<td>Inferred from rel.-specific tagging</td>
</tr>
<tr>
<td>Case et al. (2021)⁶</td>
<td>RoBERTa+ &amp; PointerNet⁷</td>
<td>RoBERTa+ &amp; rel.-spec. tagger</td>
<td>RoBERTa+ &amp; rel.-spec. tagger</td>
<td>RoBERTa+</td>
<td>BiLSTM</td>
<td>BiLSTM</td>
<td>Inferred from rel.-specific tagging</td>
</tr>
<tr>
<td>Darletis et al. (2021)⁶</td>
<td>RoBERTa+</td>
<td>RoBERTa+</td>
<td>RoBERTa+</td>
<td>RoBERTa+</td>
<td>BiLSTM</td>
<td>BiLSTM &amp; plain classifier</td>
<td>Inferred from rel.-specific tagging</td>
</tr>
<tr>
<td>Gangwar et al. (2021)⁶</td>
<td>SciBERT-3D</td>
<td>SciBERT-3D</td>
<td>SciBERT-3D</td>
<td>SciBERT-3D</td>
<td>Char-level</td>
<td>BiLSTM</td>
<td>Inferred from rel.-specific tagging</td>
</tr>
<tr>
<td>Karia et al. (2021)⁶</td>
<td>BioBERT</td>
<td>BioBERT</td>
<td>BioBERT</td>
<td>BioBERT</td>
<td>2nd BioBERT</td>
<td>2nd BioBERT</td>
<td>Inferred from rel.-specific tagging</td>
</tr>
<tr>
<td>Kohler and Jt (2021)⁶</td>
<td>GPT-3</td>
<td>GPT-3</td>
<td>GPT-3</td>
<td>GPT-3</td>
<td>2nd GPT-3</td>
<td>2nd GPT-3</td>
<td>Inferred from rel.-specific tagging</td>
</tr>
<tr>
<td>Kamiyoshi et al. (2021)</td>
<td>ELMo</td>
<td>ELMo</td>
<td>ELMo</td>
<td>ELMo</td>
<td>BiLSTM</td>
<td>BiLSTM</td>
<td>Inferred from rel.-specific tagging</td>
</tr>
<tr>
<td>Liu et al. (2021)⁶</td>
<td>DGCNN on dep. parse trees</td>
<td>DGCNN on dep. parse trees</td>
<td>DGCNN on dep. parse trees</td>
<td>DGCNN on dep. parse trees</td>
<td>BiLSTM</td>
<td>BiLSTM</td>
<td>Inferred from rel.-specific tagging</td>
</tr>
<tr>
<td>Mavridou et al. (2021)</td>
<td>Snowball pattern learning</td>
<td>Snowball pattern learning</td>
<td>Snowball pattern learning</td>
<td>Snowball pattern learning</td>
<td>N/A</td>
<td>Inferred to phrase parsing grammars</td>
<td></td>
</tr>
<tr>
<td>Peterson et al. (2021)⁶</td>
<td>Grobid-quantities</td>
<td>Grobid-quantities</td>
<td>Grobid-quantities</td>
<td>Grobid-quantities</td>
<td>Grobid-quantities</td>
<td>Grobid-quantities</td>
<td>Inferred to phrase parsing grammars</td>
</tr>
<tr>
<td>Friedel et al. (2020)</td>
<td>BERT-base</td>
<td>BERT-base</td>
<td>BERT-base</td>
<td>BERT-base</td>
<td>BERT-base</td>
<td>BERT-base</td>
<td>Inferred to phrase parsing grammars</td>
</tr>
<tr>
<td>Kononova et al. (2019)</td>
<td>REGEX</td>
<td>REGEX</td>
<td>REGEX</td>
<td>REGEX</td>
<td>2nd REGEX</td>
<td>2nd REGEX</td>
<td>Inferred to phrase parsing grammars</td>
</tr>
<tr>
<td>Hundled and Mattmann (2017)⁷</td>
<td>NN, DR &amp;</td>
<td>NN, DR &amp;</td>
<td>NN, DR &amp;</td>
<td>NN, DR &amp;</td>
<td>2nd DR &amp;</td>
<td>2nd DR &amp;</td>
<td>Inferred to phrase parsing grammars</td>
</tr>
<tr>
<td>Kim et al. (2017b)⁷</td>
<td>DB Matching &amp; existing models</td>
<td>DB Matching &amp; existing models</td>
<td>DB Matching &amp; existing models</td>
<td>DB Matching &amp; existing models</td>
<td>2nd DB Matching &amp; existing models</td>
<td>2nd DB Matching &amp; existing models</td>
<td>Inferred to phrase parsing grammars</td>
</tr>
<tr>
<td>Swan and Cole (2016)</td>
<td>DGCNN</td>
<td>DGCNN</td>
<td>DGCNN</td>
<td>DGCNN</td>
<td>BiLSTM</td>
<td>BiLSTM</td>
<td>Inferred to phrase parsing grammars</td>
</tr>
<tr>
<td>Dieb et al. (2014, 2015)⁷</td>
<td>DCNN</td>
<td>DCNN</td>
<td>DCNN</td>
<td>DCNN</td>
<td>2nd DCNN</td>
<td>2nd DCNN</td>
<td>Inferred to phrase parsing grammars</td>
</tr>
<tr>
<td>Dieb et al. (2012)⁷</td>
<td>DCNN</td>
<td>DCNN</td>
<td>DCNN</td>
<td>DCNN</td>
<td>2nd DCNN</td>
<td>2nd DCNN</td>
<td>Inferred to phrase parsing grammars</td>
</tr>
</tbody>
</table>

*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; Grobid-quantities = Grobid-quantities.*

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