MULTI-LAYER MATERIALS SCIENCE CORPUS ANNOTATION GUIDELINES: ENTITY TYPES AND RELATIONS

Matteo Finco **Felix Hildebrand** **Annemarie Friedrich**

July 25, 2024

1 Overview

The Multi-Layer Materials Science Corpus consists of scientific publications in the domain of materials science that are annotated with regard to a variety of natural language processing (NLP) tasks. This document specifies the guidelines for the annotation of the **entity mentions** and **relations** focusing on **measurements** in scientific papers within the materials science domain. Materials, their form and morphology, values and citations are considered as entities. In addition, material- and characterization-oriented relations are annotated in the corpus. For an example, see Figure 1.

Aim. The dataset is intended to train machine learning components for named entity recognition and relation extraction, with the aim of providing components for ontology-enhanced search engines.

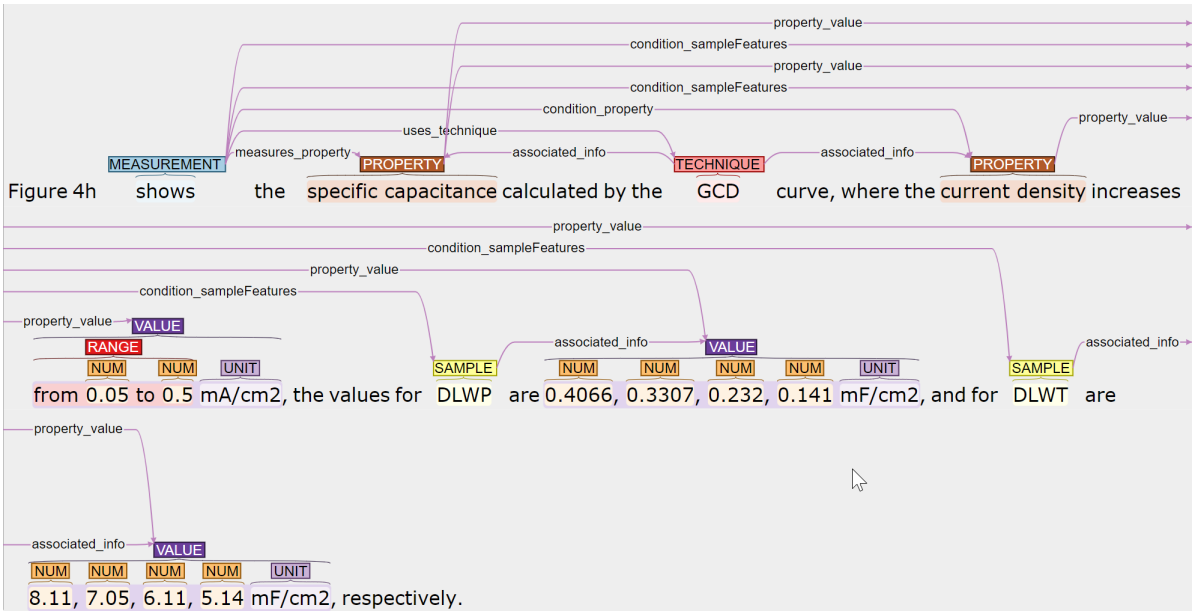


Figure 1: *Annotated sentence reporting a MEASUREMENT (`graphene_02`).*

Contents

2 Entity Types

In this section, we describe the entity type labels that apply to spans of text.

2.1 MATERIAL

The MATERIAL label is used to annotate mentions of materials in a document. These are most often reported using their chemical formula or name as shown in the example in Figure 2. This definition includes also elements and their states (such as: N atoms, copper ions, Mn(VII)) but not the description of electronic states (in "O 2p", only "O" would be marked as material) or functional groups (e.g. hydroxyl, epoxyl, etc.).

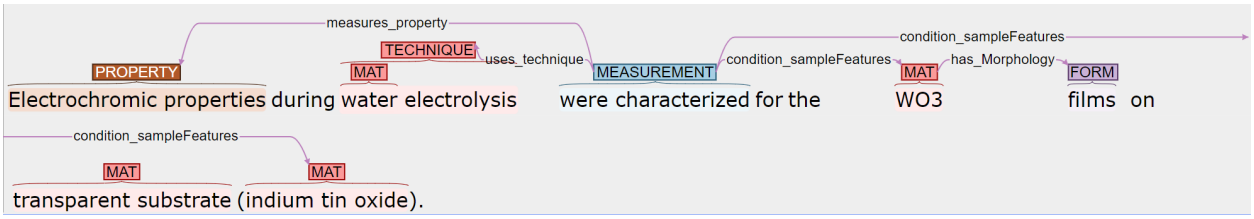


Figure 2: *"WO3" = a MATERIAL entity reported via its chemical formula, "indium tin oxide" = material mention using a chemical name. Example taken from* `steel_08`*.*

2.2 SAMPLE

The SAMPLE of a measurement is the material or the component made of materials that is being studied in the measurement or in the paper. It can either be referred to by a particular material name or composition, or by a more indirect expression, such as through some of its features such as its batch name, the components that are made out of it or the name of the models that are being evaluated. One example of this use can be found in Figure 3.

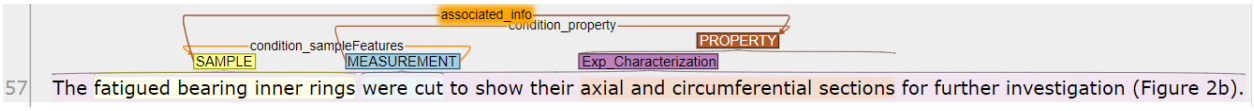


Figure 3: *Example of a typical use case of the entity SAMPLE as it expresses some features of the sample in a MEASUREMENT frame taken from* `steel_08`*.*

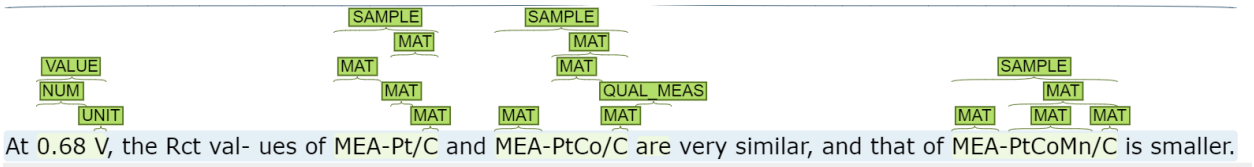


Figure 4: *Add* MATERIAL *if it is applicable in addition. Example taken from* `pemf_02`*.*

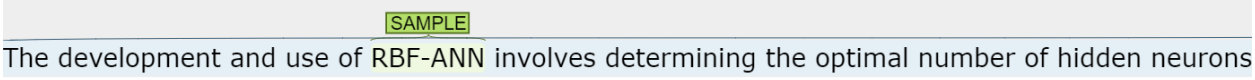


Figure 5: *In simulation papers, the SAMPLE may also refer to the computational model under discussion. Example taken from* `poly_04`*.*

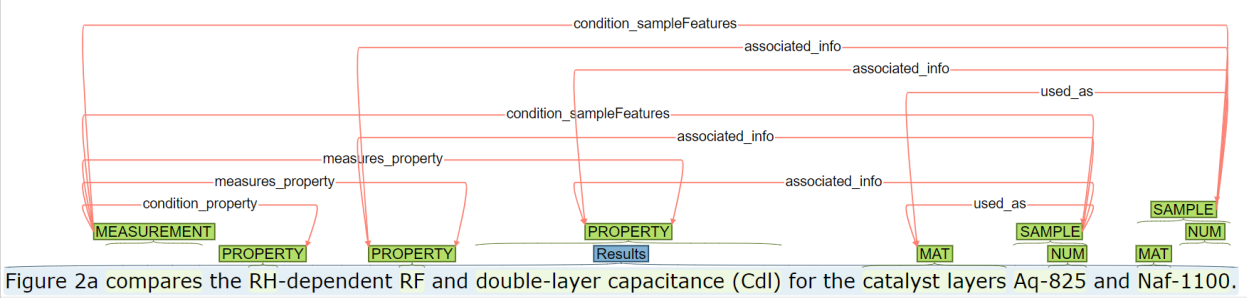


Figure 6: *The SAMPLE may also refer to a batch name as in* `pemfc_14` .

2.3 FORM

This tag is used for mentions of the form or morphology of the material, e.g., *thin film*, *gas*, *liquid*, *cubic*, or *crystals*. An example can be found in Figure 2.

2.4 DEVICE

The label DEVICE is used to identify the devices realized or described in the paper. It does not refer to the instrumentation used for manufacturing or testing or to subcomponents, but the actual target products. For example an electrolyzer is a device, while a cathode is not. As illustrated in Figure 7, this label can be the endpoint of the relation *used_in* to connect the device described to the MATERIAL that is employed in it, whether it is a specific or generic mention of a material.

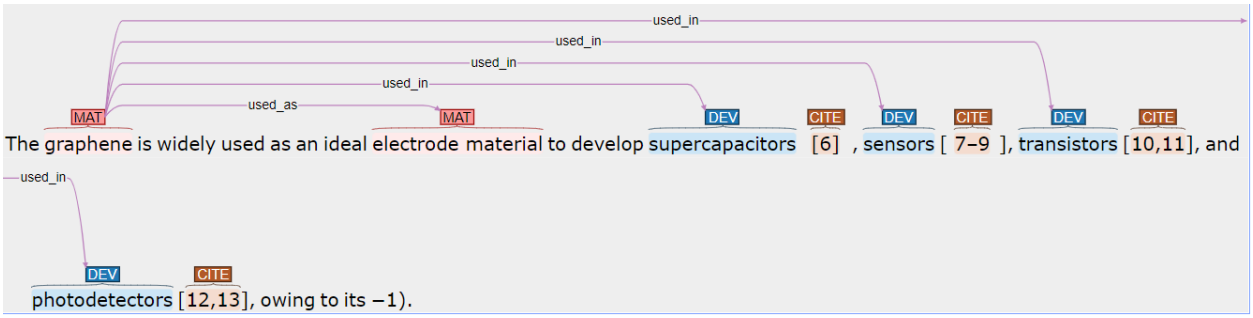


Figure 7: *Examples of devices mentioned in the paper* `graphene_02`.

2.5 NUMERIC (NUM)

The NUMERIC label is used to annotate all the numbers related to measurements. Dates and tables are not included in this definition. A single numerical value should be linked within a single NUMERIC (NUM) label like in Figure 8.

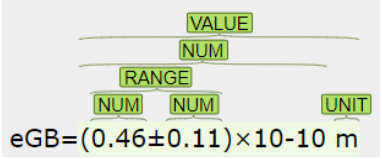


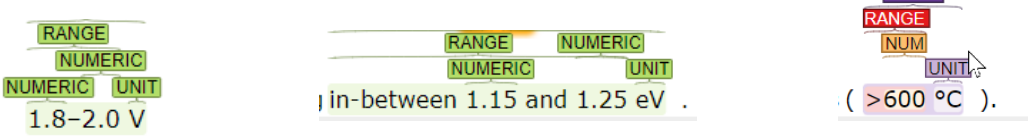
Figure 8: *Complete measure including quantity, error and order of magnitude. (Source:* `oastm_01`*)*

2.6 UNIT

The UNIT label is used to annotate the units associated to a numerical VALUE. It is worth noting that also percentage symbols have been included under this label. For examples, see Figure 9.

2.7 RANGE

The RANGE label is used whenever a range is specified, for example by two NUMERIC boundary values. Therefore, the annotated span includes two NUMERIC labels and (possibly) the words describing the range such as in the example in Figure 9.

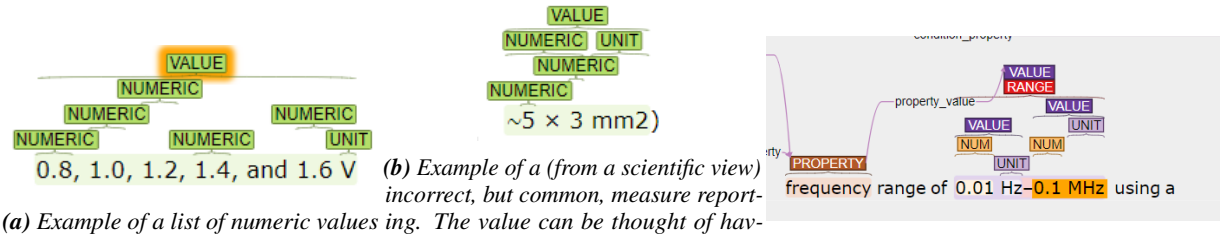


(a) *Example of annotation of a range.* ***(b)*** *Example annotation of a range where words express the relation between numerical values.* ***(c)*** *Example annotation of a range expressed using a "greater than" symbol.*

Figure 9: *Range types commonly found in materials science literature.*

2.8 VALUE

The VALUE label encompasses NUMERIC labels (or RANGE labels in case it has been used) and respective UNITs. In the following, we describe some decisions we made with regard to special cases. First, one VALUE is associated to exactly one UNIT and vice versa as shown in Figure 10. Alternatively, there is the possibility that a range needs multiple units to be expressed since its upper and lower boundary differ of many orders of magnitude as in Figure 10 c). In this case, both sub-values as well as the total expression are marked as VALUEs.



(a) *Example of a list of numeric values sharing the same unit, making them te same VALUE.* ***(b)*** *Example of a (from a scientific view) incorrect, but common, measure reporting. The value can be thought of having been expressed as 5mm * 3mm = 15mm2.* ***(c)*** *Example of range bracketing two distinct VALUE.*

Figure 10: *Value mentions commonly found in the literature.*

2.9 INSTRUMENT

This label is used to mark the names of the instruments used to perform a measurement.

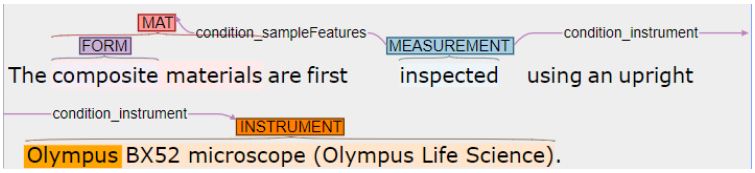


Figure 11: *Examples of an annotated sentence taken from* `electrolysis_02` *reporting the INSTRUMENT used to perform the measurement.*

2.10 PROPERTY

The PROPERTY label is used to annotate properties that are addressed (i.e., measured) in a measurement. Figure 15 contains an example.

The *measures_property* is used to connect a PROPERTY annotation to the corresponding MEASUREMENT annotation. In order to resolve ambiguity towards multiple measured properties connected to the same measurememt, a relation *property_value* is added to the tagset aiming to match the measured PROPERTY to the corresponding VALUE through respective relations. A PROPERTY can also refer to a condition of a measurement (see *condition_property*).

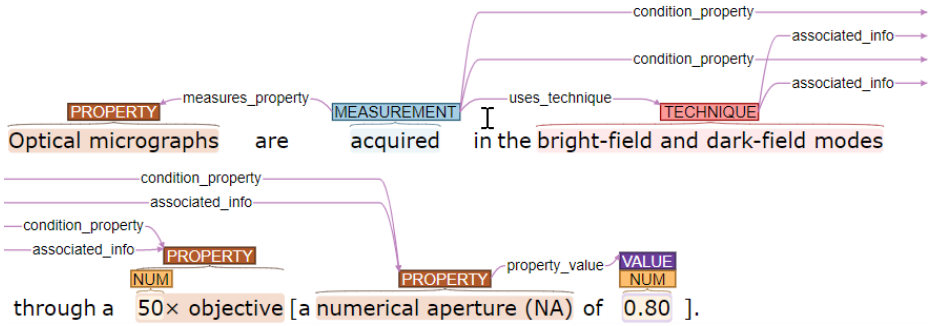


Figure 12: *Examples of an annotated sentence taken from* `electrolysis_02` *reporting the PROPERTY that is measured in the experiment, as well as PROPERTY mentions corresponding to conditions of the measurement.*

2.11 TECHNIQUE

The TECHNIQUE label is used to annotate the experimental TECHNIQUE(s) employed in the characterization steps. For an example, see Figure 12 or 13.

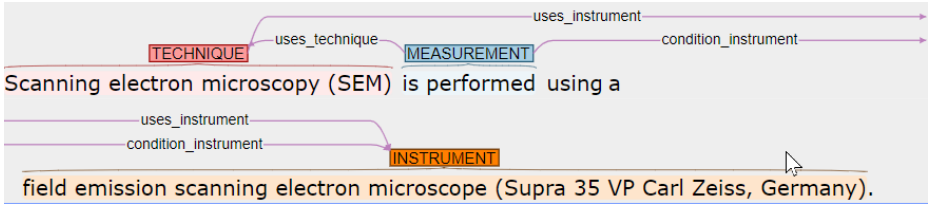
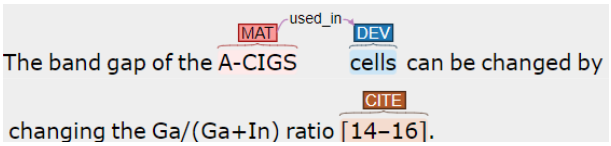


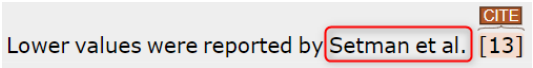
Figure 13: *Examples of an annotated sentence taken from* `electrolysis_02` *reporting the TECHNIQUE that is used in the measurement.*

2.12 CITATION

The CITATION label is attributed to in-text references which might be in numeric form. In our dataset, there are no author-year citations, though they are sometimes added to the numeric citations (see Figure 14(b)). We only annotate the latter in these cases as well as shown in Figure 14.



(a) *Example of numeric in-text citation of papers. (Source:* `electrolysis_01`*)*



(b) *Example of citation via authors' name, in the red box. (Source:* `oastm_01`*)*

Figure 14: *Citation types commonly found in the materials science literature.*

2.13 MEASUREMENT

Finally, we describe the label MEASUREMENT, which is not intended to mark entity types but instead predicates (mostly verbs) that indicate that some measurement has taken place. In linguistic terms, we treat MEASUREMENTs as a frame consisting of a frame-evoking element and several arguments. The MEASUREMENT label marks the frame-evoking element. This annotation is then the root of a graph describing the relevant arguments in the context of a characterization of a material. The role of it is primarily to act as a starting point for the relations related to measurements. For an example, see Figure 15. Therefore, annotations of this label can be linked to annotations of the types PROPERTY, VALUE, INSTRUMENT and TECHNIQUE respectively through the relations *measures_property*, *result_of*, *condition_instrument* and *uses_technique*. All of these relations originate at the MEASUREMENT annotation.

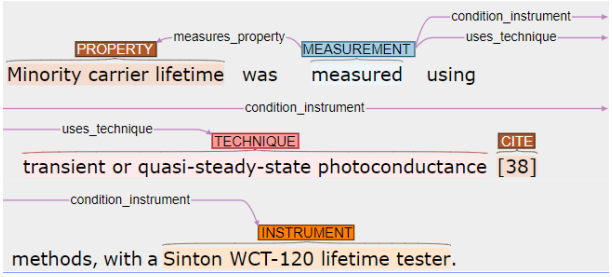


Figure 15: *Example of a sentence with a MEASUREMENT frame annotation. (Source:* `oastm_11`*) uses_instrument link missing? check: relevant if multiple techniques + instruments in one sentence?*

2.14 QUALITATIVE MEASUREMENT

This additional label is used in place of the MEASUREMENT entity type if the sentence presents only qualitative description of the measurement or of the results, but no detailed numeric results, such as shown in Figure 16. Its use and function though is generally identical to the MEASUREMENT, however, we only mark these mentions for future reference and do not mark/link their arguments (even if some of them occur in the sentence).

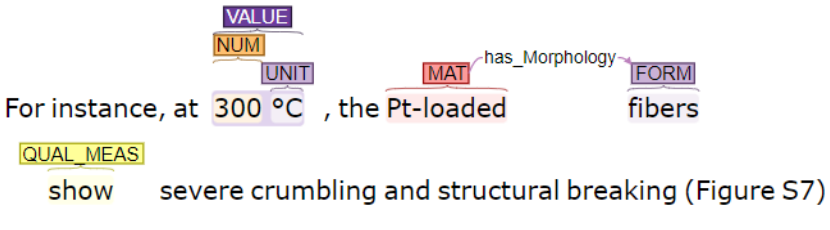


Figure 16: *Example of sentence with a qualitative result, not focused in this annotation round. (Source:* `electrolysis_02`*)*

3 Implicit Entity Mentions

The feature ***implicit*** is used whenever there is an implicit mention of any entity that must be mentioned as an additional argument of a measurement. In Figure 17, the technique of the measurement is referred to implicitly; in Figure 18, the measured property is mentioned only implicitly. For information extraction purposes, we need to know what the author refers to (domain experts gather this information from the context). Hence, we add a placeholder annotation to the same span of the corresponding MEASUREMENT annotation, add the relevant information to this annotation, which receives the feature ***implicit***.

The corresponding value for TECHNIQUE (*leed* in Figure 17) or PROPERTY (*presence* in Figure 18) is then selected from an extensible set of labels.

Note: For modeling purposes in our paper, we transform these cases involving *measures_property* or *condition_property* into *measures_property_value* and *condition_property_value* links starting at the MEASUREMENT annotation and ending at the corresponding VALUE annotation.

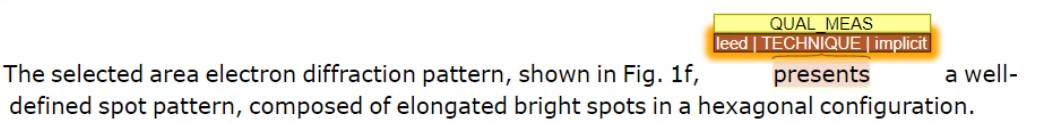


Figure 17: *Implicit mention of an experimental TECHNIQUE. In this case, a domain expert can tell that this technique was used because LEED is commonly used to investigate electron diffraction patterns and the Figure displays the corresponding result. (Source:* `graphene_02`*)*

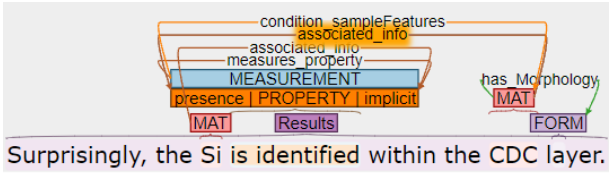


Figure 18: *Implicit mention of a PROPERTY and relative annotation of the needed features. (Source:* `oastm_02`*)*

4 Relations

This section describes the set of relations adopted in the annotations providing descriptions and examples for clarifying their practical use. We both explain relations that are used on their own and relations that are part of the measurement frame.

4.1 Measurement-related relations

This section describes the set of relations that are part of the measurement frame (see Section 2.13 for an explanation of the frame).

4.1.1 measures_property

The *measures_property* relation is used to indicate the properties that are measured. This relation links the MEASUREMENT annotation and the PROPERTY annotation. An example for this can be found in Figure 15, in Figure 12 or in Figure 19.

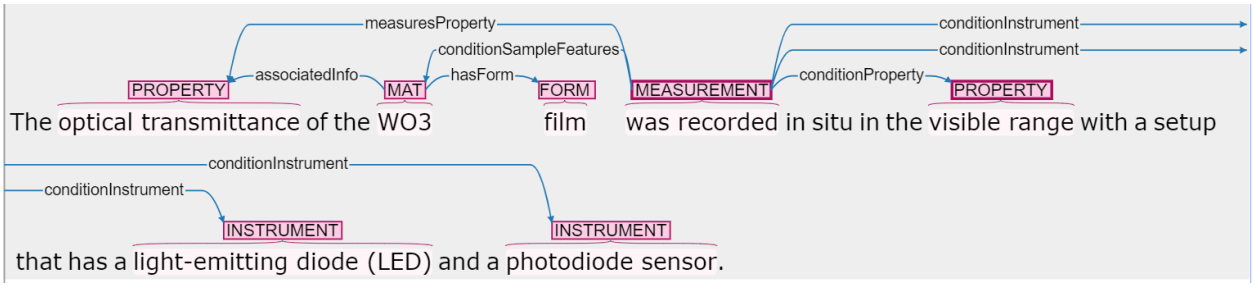


Figure 19: *Example of a sentence with a measures_property relation. (Source:* `electrolysis_01`*)*

4.1.2 same_meas

The *same_meas* relation is used to indicate if two sentences express the same measurement (usually mentioning a different set of slots). When more than two sentences describe the same experiment, the connections always start from the first one and then end on the second, third, etc.. This relation can only be established between MEASUREMENT entity types.

4.1.3 property_value

Sometimes, a sentence reports the measured values of several properties, which are hence linked to a single MEASUREMENT annotation. As there will be several PROPERTY and VALUE annotations linked to this MEASUREMENT, we resolve ambiguities by adding *property_value* relations linking the PROPERTY and VALUE mentions that belong together. The *property_value* relation starts from measured PROPERTY and ends at the VALUE entity.

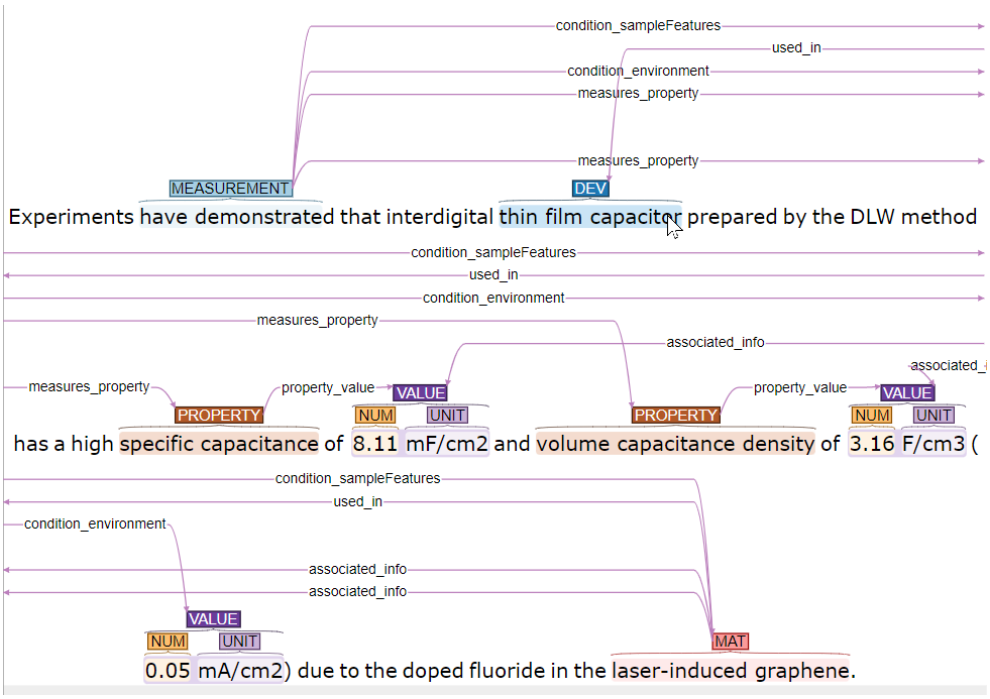


Figure 20: *Example in which the same sentence reports values of different properties in* `graphene_02`*.*

4.1.4 uses_technique

This relation extends between MEASUREMENT and TECHNIQUE, indicating the technique used in a measurement. See Figure 15 for an example. The *uses_technique* relation is used to link the characterization TECHNIQUE(s) to the correct MEASUREMENT frame.

4.1.5 uses_instrument

This relation is used whenever one INSTRUMENT is mentioned to be used to apply a certain TECHNIQUE. Specifically, the relation must start from the entity type TECHNIQUE and end in the entity type INSTRUMENT. See Figure 21 and Figure 22 for examples.

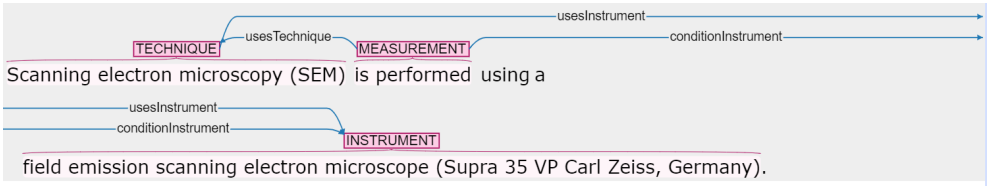


Figure 21: *Example reporting an instrument used in a particular TECHNIQUE (*`electrolysis_02`*).*

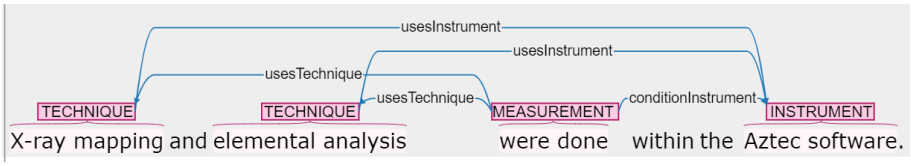


Figure 22: *Example reporting an instrument used (*`electrolysis_01`*): an instrument can also be a particular software in simulation papers.*

4.1.6 condition_environment

The *condition_environment* relation is used to indicate the environmental conditions that distinguish the experimental setup. This relation starts at MEASUREMENT annotations and ends at MATERIAL, VALUE. An example is given in Figure 23.

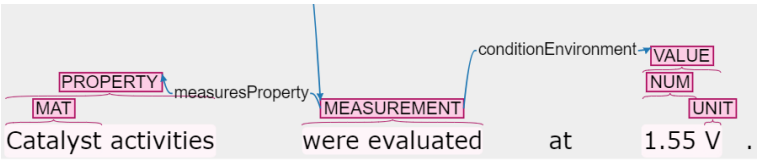


Figure 23: *Example reporting a condition environment value (`electrolysis_03`).*

4.1.7 condition_sampleFeatures

The *condition_sampleFeatures* relation is used to link the experimental setting to the material used as a sample in the measurement. This relation is used between MEASUREMENT and MATERIAL or SAMPLE. For an example, see Figure 20.

4.1.8 condition_instrument

The *condition_instrument* relation is used to link the used equipment to the MEASUREMENT annotation. This relation extends between MEASUREMENT and INSTRUMENT. For an example, see Figure 11.

4.1.9 condition_property

This relation, *condition_property* is used whenever a physical or chemical property of the experimental setup is described. This relation starts at a MEASUREMENT entity types and ends at a PROPERTY annotation. For an example, see Figure 12.

4.1.10 taken_from

The relation *taken_from* has the aim to connect the MEASUREMENT to the bibliographic reference from which the setup has been inspired or taken through the entity CITATION. An example is shown in Figure 24.

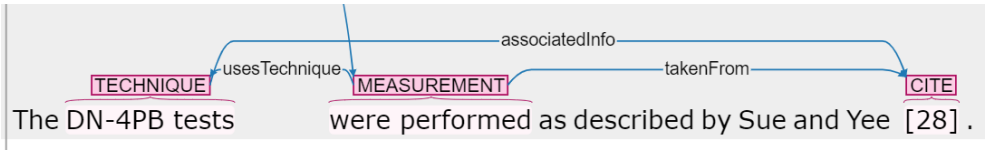


Figure 24: *In this example, the setup of the measurement has been taken over from related work (`oastm_04`).*

4.2 Relations (independent of measurements)

4.2.1 has_form

The goal of this relation is to link the MATERIAL annotation and FORM annotations in order to associate the respective material and its attributes. The relations of this type are directed from MATERIAL mentions towards the FORM (as seen in Figure 2).

4.2.2 refers_to_material

The goal of this relation is to link the MATERIAL label and FORM label in order to associate the respective material to its implicit mentions. *refers_to_material* is needed since often the name of a material is omitted and it is referred only through its form, e.g., when the material has been mentioned in an earlier sentence. The *has_Form* label is used when FORM and MATERIAL are mentioned in the same sentence; *refers_to_material* is used when the form occurs in a different (later) sentence (in linguistic terms, these are occurrences of bridging coreference). The scope of this is to prepare initial data for implicit mentions recognition. In addition, this relation has been used to link a chemical formula to the extended name of a chemical, establishing a MATERIAL-MATERIAL relation.

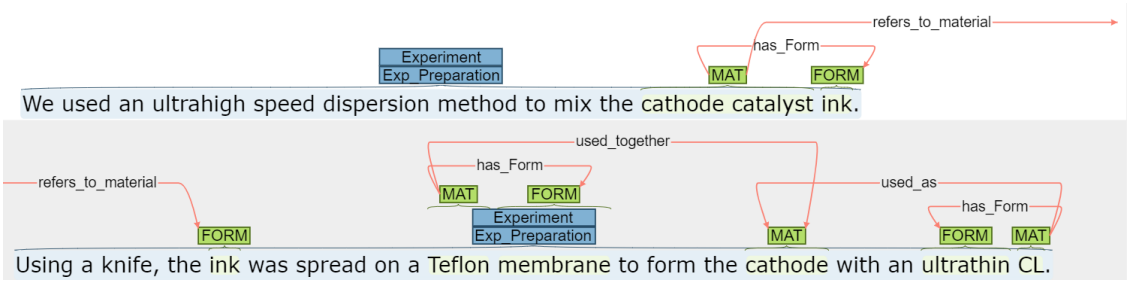


Figure 25: *Example of the use of the relation refers_to_material in the paper* `pemfc_02`*.*

4.2.3 doped_by

This relation is used to indicate which elements are dopants of the main material. For this reason this relation originates from the main material and is addressed to all the relative dopants. Both ends of the relation are therefore entities annotated with the MATERIAL label.

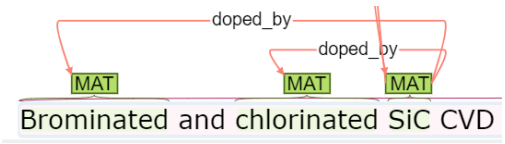


Figure 26: *Example of the use of the relation doped_by in the paper* `semi_06`*.*

4.2.4 used_in

This relation is used to relate the MATERIAL to the DEVICE in which it is used, as reported in the example Figure 27.

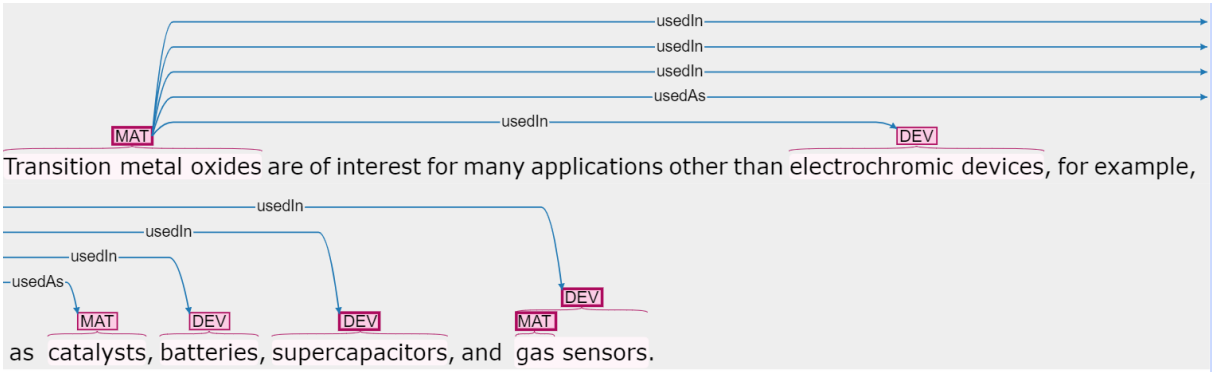


Figure 27: *Example of the use of the relation used_in (*`electrolysis_01`*).*

4.2.5 used_as

The relation *used_as* is meant to start from a specific MATERIAL annotation and target the respecitve generic MATERIAL annotation.

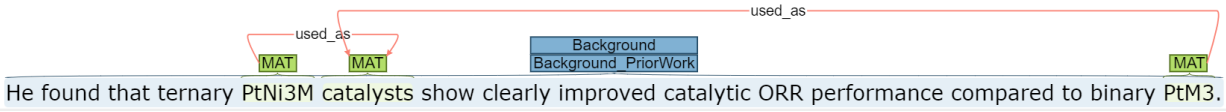


Figure 28: *Example of the use of the relation used_as in the paper* `pemfc_02`.

4.2.6 used_together

The relation *used_together* is used to link two MATERIAL annotations when they are used together in the same experiment. Every material that is a starting point for this relation identifies a distinguished set of materials. Therefore, the direction of the relation is used to identify univocally the different sets of materials. In lay terms, this relation links materials that are mixed in a solution, in gases, or are adjacent in gas/solid combinations, but they do not merge to become a new material (i.e., a composite). New "merged" materials (composites) are marked as nested MATs.

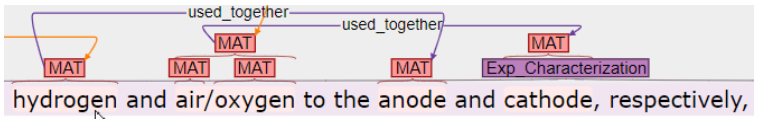


Figure 29: *Example of the use of the relation used_together (*`pemfc_07`*).*

5 Additional Examples

In this section, additional complex and useful examples are explained.

Example 1: Annotating values reported with an uncertainty and written in scientific notation.

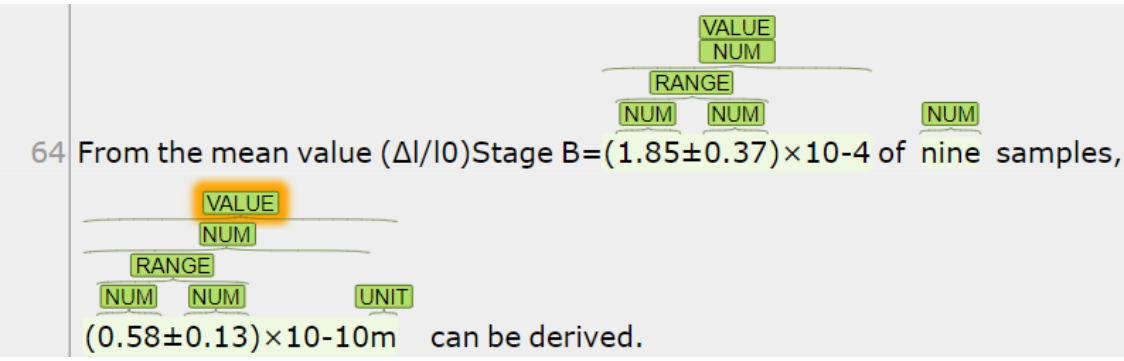


Figure 30: *Sentence taken from* `oastm_01` *containing a measure and its uncertainty written in scientific notation*

In this case (Figure 30), in order to grasp the structure of the data, a particular structure must be defined for consistency throughout the corpus. The measured quantities are labelled as NUMERIC. This definition includes the quantity and the uncertainty. Since the two of them define a range within which the real value is expected to lie, they are bracketed with RANGE. Then, another NUMERIC label is used to include also the power of 10 and this label is then linked it to the quantity.

Example 2: Ranges in scientific notation Figure 31.

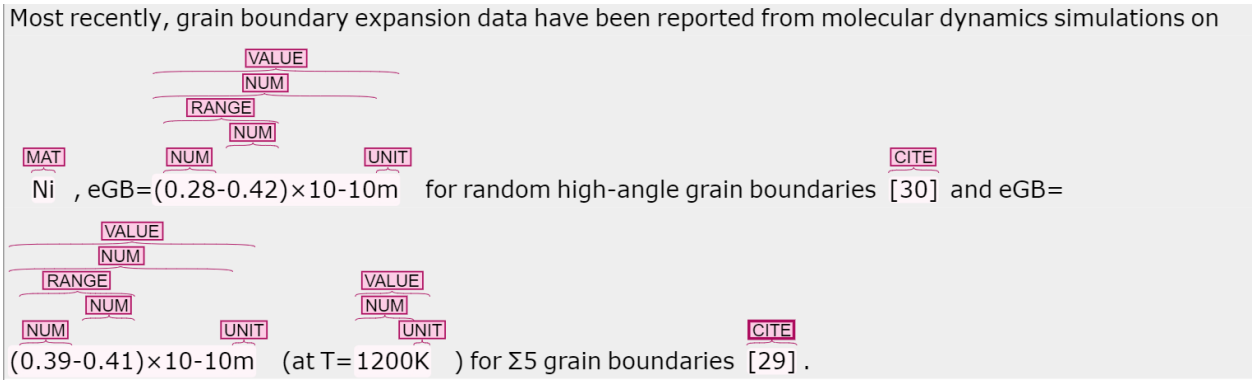


Figure 31: *Sentence taken from* `oastm_01` *containing a range written in scientific notation.*

In this example (Figure 31) the range is explicit but expressed in scientific notation.

Example 3: Indication of the order of magnitude

In this example (Figure 32), the unit is used to indicate the order of magnitude of a value, therefore no numeric entities are included. On the right side it is possible to observe how a value reported in words is annotated the very same way as one expressed using numbers and unit symbols.

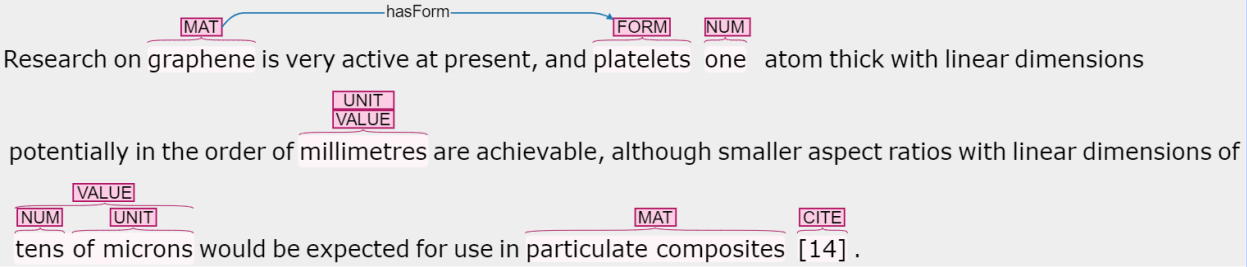


Figure 32: *Span taken from oastm_07 containing values expressed using the wording rather than the symbolic expression.*