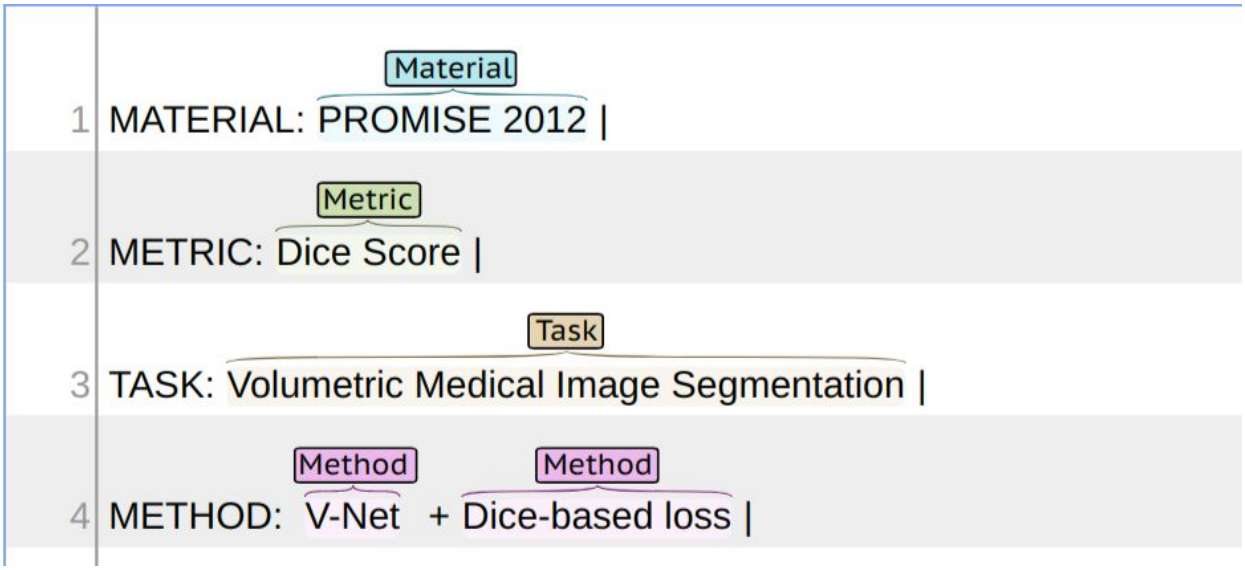
Annotation Interface :

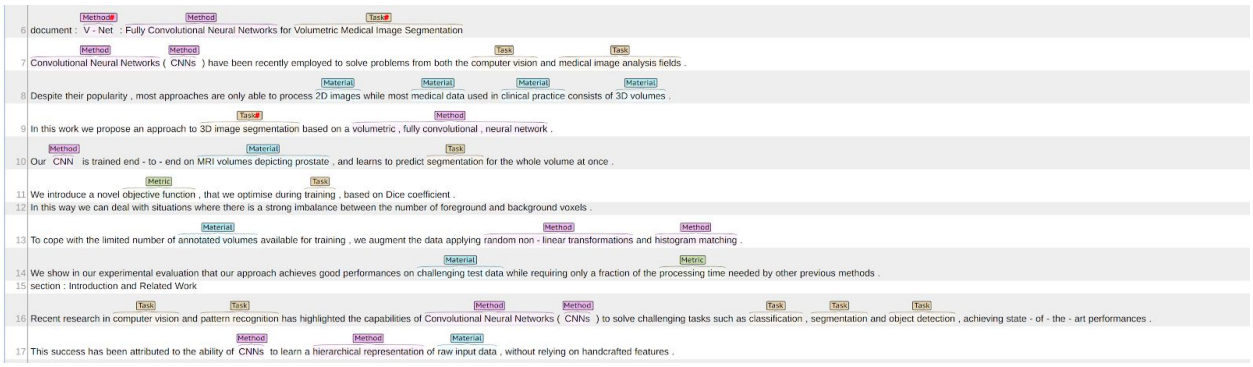
Header: True PwC Entity names that needs to be marked. In top 4 lines we show the true entity names that PwC provide us, categorised by Entity types (Material, Metric, Task, Method)

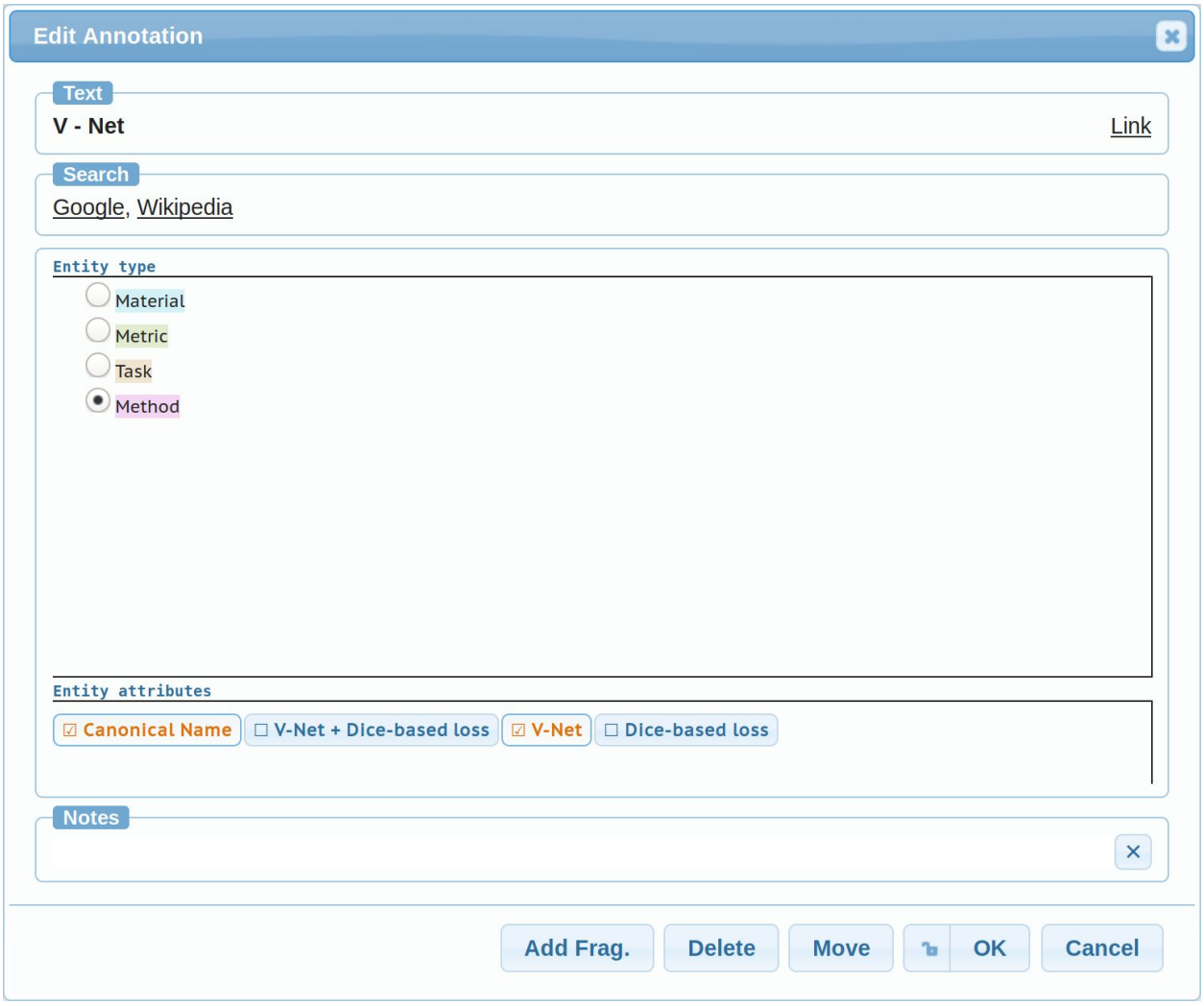
Method names can sometimes be composed of multiple parts where individual parts may be present as separate spans in document. We have used a heuristic method to split the model names into their subparts (If the splitting is wrong please delete current split annotations and resplit them).

If splitting or resplitting needs to be done on a model name, please annotate the correct split and move on to the next document. We will do a second pass of annotations for these documents.



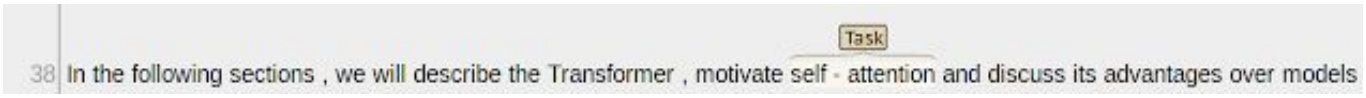
Content : This is the document to be annotated. The document contains spans of text marked with entity types. Some of the spans are also marked with red **#** symbol which denotes that the span matches one of the pwc entities (or one of the splits of pwc entities) - we call these spans as **linked** .

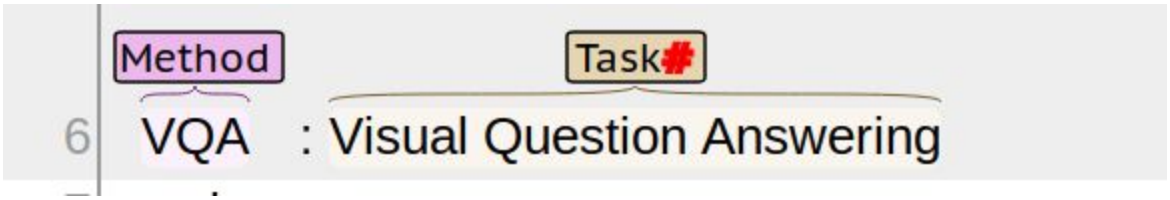




Entity Typing :

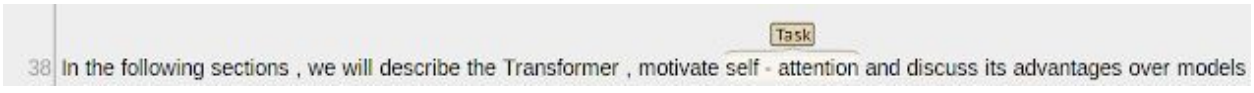
1) Correct entity types (for both matched / unmatched spans).





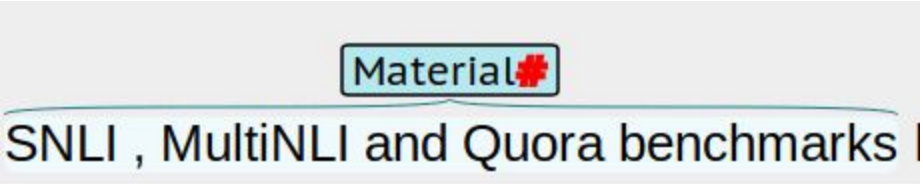
2) Delete Spans that are incorrectly marked as entity (both matched / unmatched spans).

3) Add spans and types (if a span match a PwC Entity name). Eg. Transformer below -



4) Important to focus on span that can be matched to one of PwC Entity names.

5) If a span contains multiple entities, split them into independent spans.



Type Specific Guidelines :

1) Datasets/Material span should be labeled as “Material” if they are being used and evaluated in the paper and/or refer to a named dataset.

PwC Entity Name Matching :

1) Each marked span can be matched to atmost one PwC entity name. We call these spans **linked** (marked with red **#**).. Please select the best match if multiple feasible are available.

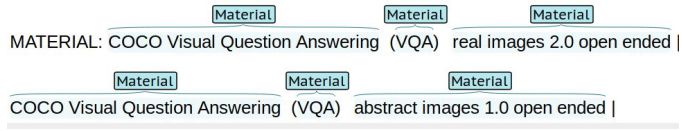
2) **We know that some entities do not occur at all in the document so please make best effort to search for a span for each PwC entity name but it is NOT a requirement.**

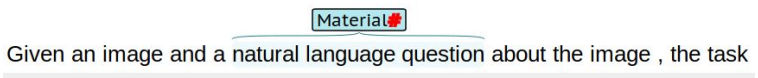
3) Note: If a single entity can match multiple PwC Entity names, that implies that name may need to split into parts.

4) Canonical name: An extra attribute is present in annotation box to mark if the span correspond to PwC Entity name as canonical (since PwC Entity name may not occur exactly as provided, we want to find the best mention in the document of that PwC Entity) .

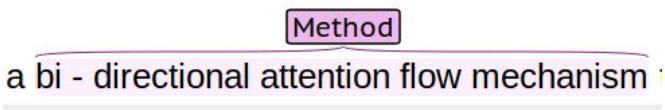
5) You can mark multiple spans in the document as canonical - we will consider the last span marked as canonical as true one.

6) Delete matched links if wrong for spans already annotated.
Example : Wrong Material Match

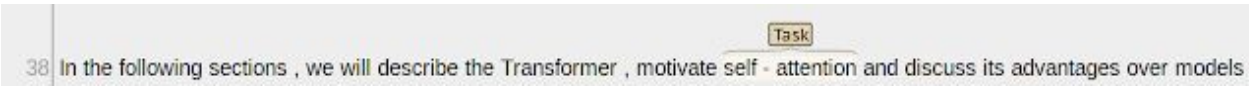




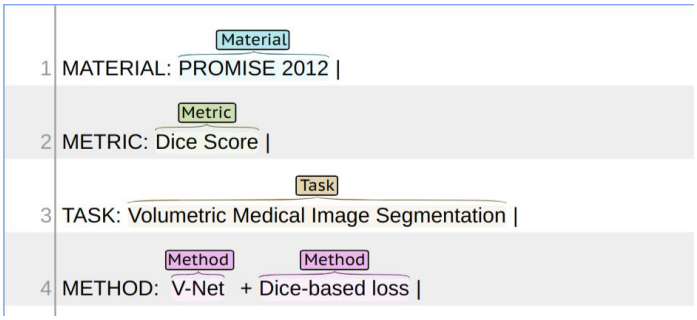
7) Add links to spans that match a PwC Entity name but not linked (important to take care of Abbreviations).

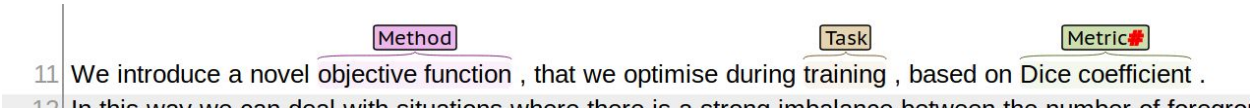


8) Add Spans that are not already annotated if they match a PwC Entity name.



9) In case where same PwC Entity name appears in multiple types , please be careful to select the correct type for a span that matched that name. For example, below “Dice score” is both a method and a metric. The preexisting annotation marked “Dice Coefficient” as metric, but according to its context it is actually behaving as method. Therefore its type should be changes to Method and it should be linked to “Dice-based loss”





Metrics to keep track of:

1) Please keep track of time for each document annotation to be completed. We can keep using the previous spreadsheet and put in start and end timestamps for each document.