Guidelines for claims classification

1 Task description

In this project, your task is to **label sentences** according to the **type of claims they contain**. Some sentences **do not contain any**, so you must assign them the label *no claim* (NC), excepted for sentences about **related works** which fall into the more precise category *related works* (RW). Otherwise, whenever at least one claim is identified, it must be categorized according to these four categories (which will be defined just below):

- *positive* claims (POS)
- *negative* claims (NEG)
- *factual* claims (FACT)
- *prospective* claims (PROSP)

You are allowed to use **multiple labels**, if and only if the sentence actually contains **multiple claims** (e.g two clauses separated by "and") from different categories. If you identify **one claim** but hesitate on its category, please choose only one that seems the most appropriate (see **Ambiguous cases** at the very end of this guide for examples).

As a **general methodology**, we recommend that you:

1. Decompose the sentence to annotate into simpler clauses (if relevant)
2. For each clause, decide if it is a claim or not. If no claims are found in the sentence, use the label (NC) or (RW)
3. Else, for each claim, identify its category and assign the corresponding label to the sentence
4. If you hesitate, please refer to these guidelines (definition of a claim, categories, and ambiguous cases). Also note that **you have access to the previous and next sentence** (in the paper's full text) **of the one you are annotating** as well as the **name of the paper section** from which it was extracted in the metadata pannel (bottom right) of the annotation page, which can allow you to better contextualize it.

1.1 What is / isn't a claim

A **claim** is a statement (at phrase-level) found in a sentence. Independently of its category, it must meet following requirements:

1. It is a **statement** emitted by the paper's authors **on the basis of their work / findings / reflexion** and NOT on previous knowledge or related works (in this last case, with an explicit reference to related works, it must be labelled (RW))

   claim: *we [prove/ suggest/ find] that Y*
   (NC): *[It is assumed in general that] Y*
   (RW): *In work X, authors [prove/ suggest/ find] that Y*

2. It captures an **essential result, contribution, or conclusion** that the authors put forward to the community for general acceptance. It can also be a **critical reflexion** about the work itself (quality, impact) or an **anticipation of future directions** to be explored.

   claim: *we successfully trained a LLM to perform multilingual translation using few-shot learning*
   claim: *in the future, we plan to further improve our method by exploring other model architectures*

   But it is **NOT** an **explanation** (introduction to the field, definitions, operation of a system, etc.) or a **description** (of a model, data, methodology, paper structure, etc.)

   (NC): *extractive summarization techniques extract key sentences from a text to produce a summary*
   (NC):*we trained for 50 epochs and used a learning rate of 0.001*
   (NC): *we used dataset X from source Y*
   (NC): *Table 1 shows our main results*

3. It implies (more or less of) the **subjectivity** of the authors, so it can be presented as only possible or plausible (*this could indicate that Y, we hypothesize that Y*)

4. It must fit in **one** of the categories (POS), (NEG), (FACT) or (PROSP). If not, it is necessarilty (NC) or (RW).

Please note that complex sentences may be broken down in phrases, some of them being claims, and some being either (RW) or (NC). In this case, multiple labels can be assigned among (POS), (NEG), (FACT), (PROSP) and (RW). The (NC) label is strictly reserved to sentences containing 0 claim.

1.2 Claim categories

Positive claims (POS)

These are claims announcing **main results, findings** and **analyses / interpretations / conclusions** that derive directly from them, or **original working hypotheses** proposed by the authors as the basis of their work. They are "positive" in the sense that they contribute to the establishment of **new knowledge** for the scientific community. They anwer to the question: *What do the authors [maybe] show / establish ?*

> *It shows that our system achieves the performance of 84.8 % / 66.7 % / 74.7 in precision / recall / fmeasure on relation detection.*
>
> *Our findings highlight the importance of equipping dialogue systems with the ability to assess their own uncertainty and exploit in interaction.*

Negative claims (NEG)

These are claims by which the authors **acknowledge some [potential] limitations** of their work or findings (often in order to nuance some *positive* claims, hence the "negative" label). They answer to the question: *What [are / could be] some limitations of the authors' [work / findings] ?*

> *The results do not necessarily apply to other encoder-decoder models or autoregressive models such as GPT series*

Factual claims (FACT)

These are claims by which the authors announce the **nature of their main contributions** in terms of what they have **actually realised or produced** (a model, a survey, a corpus, a method, etc.). These claims are mostly **found in the abstract, introduction, or concluding parts** of a paper. Sometimes, they are also **repeated in other sections**. But they should NOT be description of sub-experiments that weren't mentionned as main contributions in the beginning of the paper (e.g *We performed ablation studies* → minor contribution, part of a bigger one). They answer to the question: *What kind of contribution did the authors make in this work ?*

> *We present the first challenge set and evaluation protocol for the analysis of gender bias in machine translation (MT)*
>
> *We propose the novel task of automatic source sentence detection and create SourceSum [. . .]*

Prospective claims (PROSP)

These are claims that **anticipate** possible **consequences / impact** of the presented work or **suggestions** of **future continuations / directions**. They

answer to the question: *What [will / could] this work [become / provoke / evolve into]?*

> *The proposed method may be an important module for future applications related to time.*

> *We believe the isarcasm dataset , with its novel method of sampling sarcasm as intend by its author , shall revolutionise research in sarcasm detection in the future.*

1.3 Non-claim categories

***Related works* (RW)**

All statements based on the results of **other works**, with an **explicit mention** of these works, either quoted or simply vaguely evoked.

> *According to [X et al. 2020], the best model for task X is ...*

> *Recent works have shown that X*

***Not a claim* (NC)**

All the other sentences.

1.4 Ambiguous cases

Multiple labels

Multiple labels are to be used only when **a sentence actually contains more than one claim**. In the example below, two claims are identified: a *factual* claim (FACT) and a *positive* claim (POS). So, we should assign **both labels** (FACT) and (POS).

> *We created a new model for task X and achieved an accuracy of Z on dataset Y.*

Multiple labels shouldn't be used in case of an **hesitation on the category to choose** for **one single claim**: then, please read carefully the guidelines and following ambiguous cases to decide and choose the most appropriate labe.

Wrong sentence segmentation

In case of a poor sentence segmentation of the paper, you may encounter "sentences" that are actually incomplete because they were split, e.g

(1) *We created a corpus based on the work of (X et al.*

(2) *2020) and introduce a novel method for MT*

If you find that the overall sentence contains some claims, **please annotate all the sub-sentences accordingly** in the same manner. If you encounter sub-sentences but do not find their other parts in the surrounding documents to be annotated, and **you can't make sense of them**, please assign them the (NC) label.

Hypotheses / Arguments / Opinions

Hypotheses (or arguments/opinions) are to be distinguished according to the way they are used by the authors. If they are used to **justify the grounds** of the work (initial working hypotheses, main thesis statement), or to **analyze results** (consecutive to these results), they are considered as (POS) claims.

> *We hypothesize that task X can benefit from Y. [In this work, we will do Y and show ...]*

> *[We find result X]. We argue that this is mainly due to the dataset size.*

If they are used to imagine/assert what the presented work could be/provoke (with more **hindsight** than the previous examples), they are considered as (PROSP). If they are used to imagine/assert some limitations of the presented work, they are considered as (NEG). Please note that hypotheses (or arguments/opinions) shouldn't (in principle) fall within the (FACT) category.

1.5 Summary

category	question	characteristics
Positive claims (POS)	*What do the authors show / establish ?*	- main results, findings, conclusions - analysis / interpretation of results - key working hypothesis
Negative claims (NEG)	*What are some limitations of this work ?*	- limitations explicited by the authors
Factual claims (FACT)	*What kind of contributions did the authors make in this work ?*	- main contributions in terms of what has been realised / produced (a model, a survey, a corpus, etc.) - should be in abstract / introduction / concluding part OR repeating a FACT claim that was already found in abstract / introduction. If none of these criteria apply, probably not a *main* contribution
Prospective claims (PROSP)	*What will this work become / provoke in the future ?*	- future directions - consequences / impact (on the community / the public)

Table 1: Claim categories

category	characteristics
Related works (RW)	- statement based on other works - explicit mention of these other works (precise reference or vague mention *e.g recent works*)
No claim (NC)	every sentence that doesn't fit in claim categories or in (RW)

Table 2: Non-claim categories