ELOQUENT Lab @ CLEF 2024

Task 2: HalluciGen Detection

Task 2, first edition 2024

Motivation

Detecting hallucinated or factually incorrect information may be difficult for humans. LLMs should therefore be rigorously tested for their ability to generate accurate output prior to deployment. One possible approach for detecting hallucinated content, that we will explore in this lab, is the use of LLMs to evaluate LLM output. This task will assess both model awareness of hallucination and the viability of cross-model evaluation.

Goals of the task

This task aims to produce robust LLM-based detectors of hallucination. Given that we wish to perform a cross-model evaluation, the first goal of the task is to develop model *evaluators* that are able to both detect and generate hallucinated content. The second goal of this task is to expose the strengths and weaknesses of the evaluators by testing their detection abilities on challenging cases of hallucination.

Procedure

We define hallucination as generated text that is fluent, that differs semantically from output that would be expected given the specific prompt the model is given.

This task is heavily based on the SHROOM SemEval-2024 Shared Task. Following their work, we define HalluciGen as a hallucination detection *and* generation task that will be performed in two different scenarios: Paraphrase Generation (PG) and Machine Translation (MT).

The data format for both scenarios is the following:

<source sentence><reference><hypothesis+><hypothesis-> where:

- <source sentence> is the original model input
- <reference> is the target translation of the <source sentence> if the subtask is MT, or the target paraphrase of the <source sentence> if the subtask is PG.
- <hypothesis+> is a correct machine-generated translation/paraphrase of the <source sentence>

- <hypothesis-> is a machine-generated, incorrect translation/paraphrase of the <source sentence> with hallucinations

For example, in the following paraphrase example, the <hypothesis+> shares semantic similarity with the source and reference. The <hypothesis->, whilst fluent and sharing some surface-level overlap with the source, diverges in meaning by introducing information that was not present in the source (i.e. “legislation to address”):

> **Source:** Australia is concerned with the issue of carbon dioxide emissions.
>
> **Reference:** The problem of greenhouse gases has attracted Australia’s attention.
>
> **Hypothesis+:** Australia is concerned with greenhouse gases and the problems that they pose.
>
> **Hypothesis-:** Australia has issued legislation to address carbon dioxide emissions.

Our task is split into two years, with year 1 focusing on the builder task and year 2 focusing on the breaker.

Year 1 - builder task

Participants will develop multilingual and monolingual hallucination-aware models (evaluators) that are able to both detect and generate hallucinated content in two scenarios: machine translation and paraphrase generation.

For each scenario (machine translation/paraphrase generation), we ask the participants to develop generative, hallucination-aware models that are able to perform the following:

Detection step: Given a <source sentence> and two hypotheses hyp1, hyp2, the model should decide which one between hyp1 and hyp2 is a hallucination.

The definition of this step changes based on the number of hypotheses provided and whether the <reference> is part of the definition. For the reference-free case, we will have these possible scenarios:

- *Contrastive* : Given the <source sentence> and two hypotheses detect which one is the hallucination
- *Non-contrastive* : Given the <source sentence> and one hypothesis determine if the hypothesis is a hallucination or not

In the reference-based case, the <reference> will be available (in addition to the <source sentence>) in both the contrastive and non-contrastive scenarios.

Update 15th April 2024: We plan to use the contrastive reference-free scenario, for the first year at least.

Generation step: Given a source sentence, generate two LLM hypotheses:

- **<hypothesis+>** that is a correct translation/paraphrase of the source, and
- **<hypothesis->** that is a hallucinated translation/paraphrase of the source.

For details about the scoring of the generation step, please refer to the subsection "Cross-model evaluation of the generation step".

Year 2 - breaker task

The focus of Year 2 will be on the *breaker task*. Participants will develop novel evaluation datasets designed to break the hallucination detection capabilities of one or more baseline evaluator models. The evaluator model(s) could, for example, be the best-performing model(s) from Year 1. This task is inspired by the WMT challenge set subtask and the development of the Adversarial NLI dataset. The focus will be on automated methods for dataset construction, for example leveraging pre-existing datasets and applying techniques such as data augmentation (e.g. using LLMs). This task will explore the use of automated methods for generating high-quality evaluation benchmarks.

Data

The dataset used in this task will be compiled from existing datasets around Paraphrase Generation (for the monolingual scenario) and Machine Translation (for the cross-lingual scenario). Examples of datasets that we are planning to base our datasets on: SHROOM for Paraphrase Generation, and ACES for Machine Translation. Update on 18th April: We request that participants do not use either of these datasets for training/tuning their models. Instead, please use the trial data provided for the HalluciGen task.

Data examples that lack the <hypothesis+> or the <hypothesis-> will be complemented by synthetically generated entries, provided by the organizers. Each <hypothesis+> will be assigned to a hallucination type. We are reusing the set of hallucination types defined in ACES and extend it to include negation and tense. The final set of hallucination categories is the following: *addition*, *named entity*, *number*, *conversion*, *date*, *tense*, *negation*, *gender*, *pronoun*, *antonym*, *natural*.

For each task scenario (paraphrase/machine translation), we are going to provide a trial and a test split per step (hallucination generation and detection). The trial and test splits will become available through HuggingFace by 22nd April 2024.

Participant input

- Data: We are open to suggestions of other suitable datasets to include.
- Baseline models: We are actively seeking hosted models with APIs to use as baselines.
- Language cover: We have proposed a set of base languages but we are open to additions to this set.

Submission format (updated on 22nd April)

Important information:

- Submit all files as **comma-separated CSV.**
- Submit all files using **this submission form**, where you will also be asked to show your prompt and elaborate on the LLM you used. You can reuse the same submission form to submit results for different steps, different scenarios, and different languages**.**
- Submit **a separate CSV file for each language or language-pair**. For example, if the task data is in English and Swedish, there will be one CSV file per language.
- The technical specifications below are the same for both Hallucination and Machine Translation scenarios.

For the detection step

Schema:

COLUMN	TYPE	INFORMATION
id	`int`	This is the id of the source sentence. Evaluation will be done by id, so make sure to include it!
label	`str` / categorical	This is the detected label. Only these two values are allowed: - `hyp1` - `hyp2`

explanation	`str`	This is an **OPTIONAL** column. If your model produces some kind of explanation as to why it detected a hallucination in one hypothesis and not the other, you are free to include it here. You **do not** have to have them for every example, and they **will not** be evaluated

Example file:

```
id,label,explanation
1,hyp1,"The first hypothesis switches the dates around"
2,hyp2,"Hypothesis 2 contains a hallucination because it introduces new information not present in the source sentence. While Hypothesis 1 accurately paraphrases the absence of birds in the sky, Hypothesis 2 replaces ""birds"" with ""cats,"" which diverges from the original meaning and introduces false information about cats flying in the sky, constituting a hallucination."
3,hyp1,
4,hyp1,
```

For the hallucination generation step

Schema:

COLUMN	TYPE	INFORMATION
id	`int`	This is the id of the source sentence. We need the id to match each set to the source sentence, so make sure to include it!
hyp+	`str`	This is a hypothesis supported by the source sentence
hyp-	`str`	This is a hypothesis **not** supported by the source sentence (in other words, it contains a hallucination)

Example file:

```
id,hyp+,hyp-
1,"The upcoming Berlin summit, with its exclusive concentration on Agenda 2000, holds considerable significance for the forthcoming trajectory of the European Union.","The Berlin summit, which will exclusively prioritize Agenda 2000, is anticipated to overlook critical matters concerning economic reforms within the European Union."
2,Agenda 2030 does not include a chapter on renewable energy.,Agenda 2000 does not include a chapter on renewable energy.
```

For the cross-model evaluation of the generation step

The schema and example are the exact same as the detection step.

Multilinguality

We plan to provide data and baseline models that support a subset of languages included in the ACES and SHROOM datasets. ~~The list of base languages will be refined at a later stage.~~

Update 15th April 2024: We are going to release data that cover the following languages:

- Paraphrase: English, Swedish
- Machine Translation: English-German, English-French, French-English, German-English

Task scoring

Hallucination detection step: The model output is limited to a couple of tokens [1]

Example input to the participant model (**paraphrase** scenario):

> *Which one of hyp1 and hyp2 is not supported by src?*
>
> ***src:*** *The fact is that a key omission from the proposals on agricultural policy in Agenda 2000 is a chapter on renewable energy.*
>
> ***hyp1:*** *Agenda 2030 does not include a chapter on renewable energy.*
>
> ***hyp2:*** *Agenda 2000 does not include a chapter on renewable energy.*
>
> ***Accepted answers:*** *hyp1 or hyp2*

Expected output:

> ***hyp1***

Example input to the participant model (**translation** scenario):

> *Which hypothesis is not supported by src: Hyp1 or hyp2?*

[1] If your model generates an explanation, you are welcome to include it as an extra optional column in your submission file.

src: *Er wurde in aufeinanderfolgenden Coups durch Olusegun Obasanjo (1975) und Murtala Mohammed (1976) ersetzt.*

Hyp1: *He was replaced by Olusegun Obasanjo (in 1975) and Murtala Mohammed (in 1976) in successive coups.*

Hyp2: *He was replaced in successive coups by Olusegun Obasanjo (1975) and Murtalea Mohammed (1976).*

Accepted answers: *hyp1 or hyp2*

Expected output:

hyp2

Scoring procedure (same for both scenarios):

- We ask the participants to submit a `.csv` file with the answers. (In addition, in the online submission form, we will ask participants to submit the prompts they used.)
- Then we will calculate the F1-score/accuracy based on the gold labels that denote which hypothesis (hyp1/hyp2) contains the hallucination. In the case of the paraphrase scenario the examples have been human annotated. In the case of the translation scenario the labels were generated using a combination of human and automated annotation.

Hallucination generation step: All participant models are provided with a set of source sentences and they are asked to generate both a good hypothesis (<**hypothesis+**>) and a hallucinated one (<**hypothesis-**>) for each source sentence.

Example for the **paraphrase** task:

Given the src below, generate a paraphrase hypothesis hyp+ that is supported by src and a second paraphrase hyp- that is not supported by src.

src: *The fact is that a key omission from the proposals on agricultural policy in Agenda 2000 is a chapter on renewable energy.*

Expected output:

Paraphrase hypothesis supported by src (hyp+): One notable absence in the agricultural policy proposals of Agenda 2000 is a section addressing renewable energy.

Paraphrase hypothesis not supported by src (hyp-): Agenda 2000 lacks comprehensive measures for addressing climate change impacts within agricultural policy, which could significantly hinder the transition to renewable energy sources.

Example for the **translation** task:

Given the src below, generate a translation hypothesis hyp+ that is supported by src and a second translation hyp- that is not supported by src.

src: Es ist der Sitz des Bezirks Zerendi in der Region Akmola.

target language: *English*

Expected output:

Translation hypothesis supported by src (hyp+):
It is the seat of the district of Zerendi in Akmola region.

Translation hypothesis not supported by src (hyp-):
It will be the seat of the Zerendi District in Akmola Region.

Cross-model evaluation of the generation step : In this step each of the generated hypothesis pairs (**hyp+, hyp-**) is given as an input to another participant model for evaluation. The cross-model evaluation process follows the detection step, with the only difference being that (**hyp+, hyp-**) are generated by participant models. After formatting the new cross-model dataset consisting of the **src**, **hyp+** and **hyp-**, we will release it to the participants for evaluation. Note that the organizers will collect, anonymize and shuffle all responses, before releasing the data to the participants. Clarification on 22nd April 2024: We will measure agreement to assess:

1. Consistency of the model predictions: are the participant evaluator models consistent with each other in their predictions of which hypothesis contains the hallucination(s)?

2. Accuracy of the model predictions: are the participant evaluator models' predictions accurate with respect to the labels (hypothesis-/hypothesis+) produced by the generator model?

Submission dates

1. **May 1st:** Participants submit their results for the **hallucination generation step**
2. **May 3rd:** Organisers collect the submitted outputs of the hallucination generation step and redistribute them to the participants so that they can get started with the **cross-model evaluation phase** of the hallucination generation step
3. **May 7th:** Participants submit their results for the **hallucination detection step**
4. **May 10th:** Participants submit their results for the **cross-model evaluation of the hallucination generation step**
5. **May 22nd:** Organisers send results to participants.

Plans for coming years

This task is split into two years. In the third year, we aim to build on the progress of the previous years and run the *builder* and *breaker* subtasks in parallel.