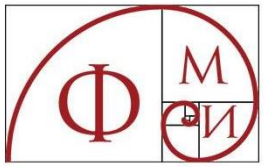
Подходи за обработка на естествен език
SemEval-2025 Task-3 — Mu-SHROOM

Кристиян Симов, ф.н. 4MI3400288
Цветан Цветанов, ф.н. 4MI3400570
Изкуствен интелект

8 февруари 2025 г.



1 Въведение

Големите езикови модели владеят естествените езици и звучат убедително. Но понякога генерират грешни и подвеждащи твърдения, които нямат връзка със заявката на потребителя или реална подкрепа с факти. От друга страна, съществуващите метрики са пригодени по-скоро да описват нивото на владеене на езика от модела, отколкото неговата коректност. Тези феномени често биват наричани "халюцинации".

Халюцинирането е един от ключовите все още неразрешени проблеми на големите езикови модели. Задачата Mu-SHROOM в изданието на SemEval от 2025г. е продължение на SHROOM от миналогодишното издание. През 2024 година задачата пред участниците е била да класифицират дали даден текст е халюцинация (булева класификация). Промяната в сегашното издание е, че се очаква да се предскаже начало и край (диапазони) на халюцинации в изходния текст на конкретни езикови модели.

2 Данни

Използвани са единствено публично достъпните данни от страницата на задачата. Не е правен опит да се генерират синтетични данни. Организаторите са подготвили текстови данни за трениране, тестване и валидация на 14 различни езика - арабски (модерен стандартен), баски, каталонски, китайски (мандарин), чешки, английски, фарси, финландски, френски, немски, хинди, италиански, испански и шведски. Данните са предоставени в jsonl формат. Всеки ред от тези файлове отговаря на json обект:

```
[8]: train_unlabeled_en_df = pd.read_json("data_sets/train_unlabeled/mushroom.en-train_nolabel.v1.jsonl", lines=True)
     train_unlabeled_en_df.head()
```

[8]:

	lang	model_id	model_input	model_output_text	model_output_logits	model_output_tokens
0	EN	togethercomputer/ Pythia-Chat-Base-7B	Do all arthropods have antennae?	Yes, all insects and arachnids (including spi...	[-2.57427001, 5.1865358353, 5.4173498154, 2.32...	[ĠYes, ,, Ġall, Ġinsects, Ġand, Ġar, ach, n, i...
1	EN	togethercomputer/ Pythia-Chat-Base-7B	Do all arthropods have antennae?	Yes, all insects and arachnids have at least ...	[-2.57427001, 5.1865358353, 5.4173498154, 2.32...	[ĠYes, ,, Ġall, Ġinsects, Ġand, Ġar, ach, n, i...
2	EN	togethercomputer/ Pythia-Chat-Base-7B	Do all arthropods have antennae?	Yes, all insects and arachnids (including spi...	[-2.57427001, 5.1865358353, 5.4173498154, 2.32...	[ĠYes, ,, Ġall, Ġinsects, Ġand, Ġar, ach, n, i...
3	EN	togethercomputer/ Pythia-Chat-Base-7B	Do all arthropods have antennae?	Yes, all insects and arachnids (including spi...	[-2.57427001, 5.1865358353, 5.4173498154, 2.32...	[ĠYes, ,, Ġall, Ġinsects, Ġand, Ġar, ach, n, i...
4	EN	togethercomputer/ Pythia-Chat-Base-7B	Do all arthropods have antennae?	Yes, all insects and arachnids (including spi...	[-2.57427001, 5.1865358353, 5.4173498154, 2.32...	[ĠYes, ,, Ġall, Ġinsects, Ġand, Ġar, ach, n, i...

Фигура 1: Английско обучаващо множество

3 Метод

3.1 Обработка на данните

- Токенизиране изхода на модела - с цел генериране речник на отместванията (offset mapping)
- Векторизиране входа на потребителя и изходните токени на модела
- Кеширане на обработените и почистени данни на диска

3.2 Алгоритми

Използвана е невронна мрежа със стохастично градиентно спускане (Adam optimizer) за изчисление на вероятност за потискане на изходни токени базирайки се на съответстващите им логити и на семантична близост. Хиперпараметри на оптимизатора са: наказателен параметър (за да ограничи склонността към потискане), параметър на ентропията (насърчава вероятностите за потискане да са близо до 0 или 1), скорост на обучение и брой епохи. Алгоритъмът на обучение използва крос-ентропия като функция на загубата между вероятностното разпределение на вгражданията на входа и съответното разпределение на вгражданията на маскираните изходи. Токените с високи вероятности на потискане (над определена стойност на праговия хиперпараметър) се сливат в диапазони. Опционално, високите вероятности се разпространяват.

4 Експерименти

- Опити с различни стойности на хиперпараметрите и сравнение на резултатите с цел постигане на най-висока точност за всички езици
- Vectara Hallucination model приложен към двойки от входни заявки и изходни токени за изчисление вероятност за наличие на халюцинации
- Чистене на аномалии чрез регулярен израз - технически токени в изходните токени и съответните им логити
- Евристики за разпръскване на определено количество вероятностна маса над определен праг в съседни вероятности (с цел "заразяване" на съседни токени).
- Обучаване на модела с различни подмножества от обучаващите множества езици
- Генериране диапазони на халюцинация с и без речници на отстъпа

5 Резултати

Публично Github хранилище с кода от проекта Официално класиране

6 Заключение и бъдеща работа

Както се вижда от предната секция, модел o1_toy постига най-добра точност за английски (≈ 17% IoU №39), немски (≈ 13% IoU №28), арабски (≈ 11% IoU, №29). Тези по-добри резултати в сравнение с останалите езици се дължат на наличието отворени токенизатори за съответните езикови модели, с които могат да се генерират речници за отстъпа. За сравнение без речници: английски (≈ 13% IoU), немски (≈ 10% IoU) арабски (≈ 7% IoU). Корелацията на Спиърман (Cor) между генерираните от модела вероятности, назначавани на всеки предсказан диапазон съдържащ халюцинации, и вероятностите от златните корпуси като цяло е ниска и дори отрицателна.

Идеи за бъдещо подобрение:

- Генериране на синтетични данни и анотиране на тестови множества
- Генериране на речници за отстъпа (offset mappings) за всички езици
- Vectara за вграждания и вероятност за халюцинации
- RAG за проверка на фактологията
- Смяна BERT модела за изречения all-mpnet-base-v2 за по-сполучливи вграждания
- По-фино оптимизиране на хиперпараметрите
- Съкращаване времето за обучение на невронната мрежа
- Изчисляване с точни метрики представянето на нашия модел - f1, precision, recall.

7 Индивидуален принос

Нашият екип вярва в гъвкавите принципи на разработка като extreme programming(XP) и по конкретно pair programming. Поради ограниченото време, с което разполагаше екипът преди крайния срок за участие в SemEval, много голяма част от проучването, разработката, експериментите и писането на документация беше споделена във виртуална среда. Това улесни комуникацията и синхронизацията, освен това доведе до бързо тестване на прототипи и ранно откриване на грешки.