Halucinace ASR modelů, jejich detekce

Maxim Plička, Michal Pyšík, Michal Bartošák

Brno University of Technology, Faculty of Information Technology

Božetěchova 1/2. 612 66 Brno - Královo Pole

April 9, 2024

- **Definice:**

  „Text generovaný modelem, který je sémanticky nesouvisející s referencí, ale přesto plynulý a smysluplný.“

Reference:	Na poli běží pes.
Chyba přepisu:	Traktor na poli přejel psa.
Fonetická chyba:	Na poli běží les.
Oscilace:	Na poli běží pes pes pes pes.
Halucinace:	Na poli běží pes. Nezapomeň dát odběr.

Proč jsou halucinace problém?

- Zmatení/oklamání uživatele a ztráta jeho důvěry
- Bezpečnostní (např. medicína) a legální (právo) rizika
- Zhoršení produktivity (a ztráta peněz) v korporátním prostředí
- Poškození pověsti firem provozující jazykové modely

Příklad

V lékařském prostředí se ASR model může používat k transkripci lékařských záznamů nebo konzultací. Chyby v transkripci mohou mít závažné následky, například při předpisu léčiv nesouvisejících s léčbou nebo popisu stavu pacienta.

Příčiny vzniku halucinací

- Nedůvěryhodnost trénovacích dat (nesprávnost labelů)
- Šum v testovací nahrávce
  - Především na jejím začátku
- Opakovaná (či velmi podobná) trénovací data
  - viz Whisper Large-V3 a jeho trénování na Youtube datech

Náš nejlepší nález (projekt)

no doubt no doubt returned mr lilburn but if thy right eye offend thee pluck it out and cast it from thee for it is profitable for thee that one of my members should perish and not that **i hope you enjoyed this video if you did please like and subscribe and hit that bell icon so that you can get notified of my next video and i will see you in the next one by bye** body should be cast into hell

Jak detekovat halucinace 1/4

- Těžko formalizovatelný (jak je odlišit od jiných typů chyb?) a doposud málo zkoumaný problém
- Charakteristiky přepisů obsahující halucinace bývají:
  - Dobrá plynulost přepisu
  - Nízká sémantická souvislost se zbytkem přepisu
  - Většinou nižší fonetická podobnost s referencí
- $\Rightarrow$ K řešení problému je možné využít sofistikovanou kombinaci existujících metrik

Jak detekovat halucinace 2/4

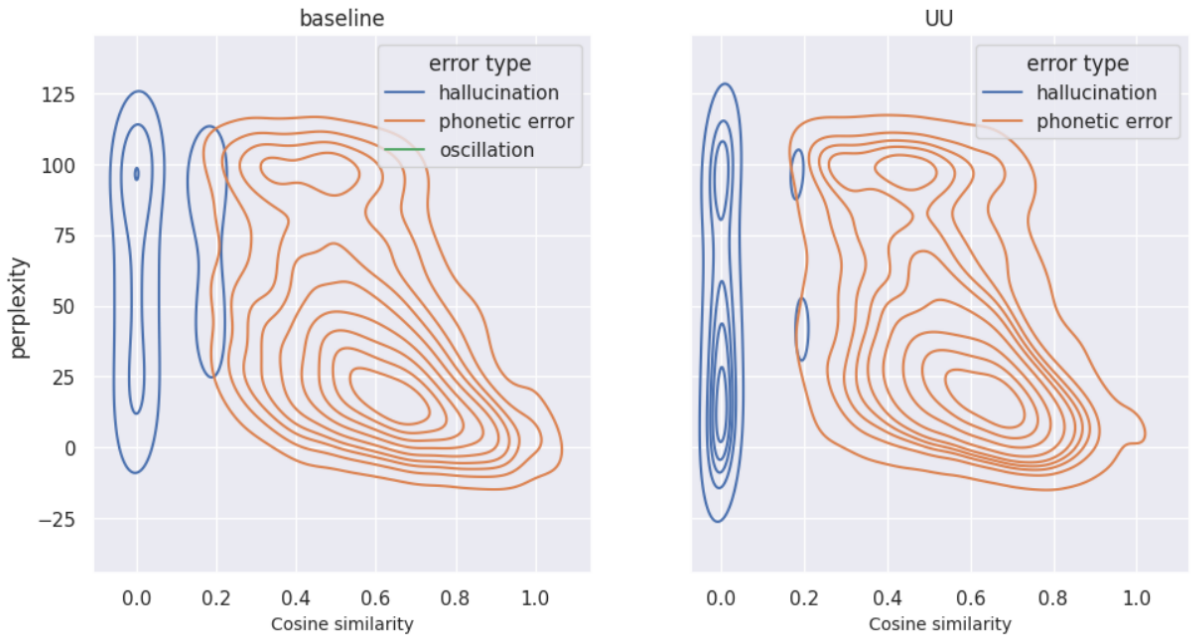
- WER - Hrubá detekce chyb
  - Procentuální podíl chybně rozpoznaných, nahrazených nebo vložených slov ve větě oproti referenční
  - Díky vloženým slovům budou mít halucinace vysokou hodnotu
- Cosine Similarity - Sémantická vzdálenost dvou vět
  - Výpočet vzdálenosti dvou vektorů reprezentující věty
  - Odolnost vůči rozdílné délce vět
  - Halucinace mají nízkou sémantickou souvislost s referencí
- Perplexity - Souvislost věty
  - Vložení věty do jazykového modelu, který na základě kontextu celé věty určuje, jak moc je věta souvislá (čím nižší tím lepší)
  - Problém s horní hranicí (nekonečno)
  - Halucinace budou mít nízkou hodnotu, slouží k vyřazení neplynulých výstupů (word salad)

Jak detekovat halucinace 3/4

1: $y \leftarrow \text{model}(x)$
2: $w \leftarrow \text{WER\_score}(y, t)$
3: **if** $\mathit{WER\_score} > \mathit{tresh}_{WER}$ **then**
4: $\quad c \leftarrow \text{cos\_similarity}(y, t)$
5: $\quad p \leftarrow \text{perplexity}(y)$
6: $\quad$ **if** $c < \mathit{tresh}_{cos}$ and $p < \mathit{tresh}_{perplexity}$ **then**
7: $\quad\quad$ *natural halucination*
8: $\quad$ **end if**
9: **end if**

Jak detekovat halucinace 4/4

- Jak tedy zvolit thresholdy metrik tak, aby detekovaly halucinace?
- Je třeba je statisticky vyhodnotit data a jednotlivé hodnoty nastavit



Obrázek: Cosine similarity a perplexity dvou modelů, ukazující jasný rozdíl mezi fonetickými chybami a halucinacemi.