Détection d'Hallucinations dans le Cadre de la Tâche 6 SemEval-SHROOM

Nihed Bendahman, Karen Pinel-Sauvagnat, Gilles Hubert, Mokhtar Boumedyen Billami

Détection d'Hallucinations dans le Cadre de la Tâche 6 SemEval-SHROOM

Nihed Bendahman[1,2], Karen Pinel-Sauvagnat[1], Gilles Hubert[1] and Mokhtar Billami[2]

[1]*Université Paul Sabatier, IRIT, UMR 5505 CNRS, Toulouse, France*
[2]*Berger-Levrault, 64 Rue Jean Rostand, 31670 Labège, France*

Abstract

Cet article présente notre participation à la tâche 6 de SemEval-2024, nommée SHROOM (*a Shared-task on Hallucinations and Related Observable Overgeneration Mistakes*). L'objectif de la tâche est de détecter des hallucinations. Nous avons proposé deux types d'approches pour la tâche: une première basée sur des *embeddings* (ou plongements) de phrases, et une seconde basée sur des LLMs (Large Language Model). Nous observons que les LLMs ne parviennent pas à améliorer les performances obtenues par les modèles de génération de plongements de phrases. Ces derniers surpassent la baseline fournie par les organisateurs, et notre meilleure approche obtient 78% de précision.

Keywords

Détection des hallucinations, Génération de texte, Plongements contextuels de phrases

1. Introduction

Malgré les performances intéressantes atteintes par les derniers modèles de génération de texte, tels que GPT-4 [1] ou Llama 2 [2], un problème crucial concerne les hallucinations dont ils peuvent faire preuve: les segments de texte concernés semblent fluides et naturels mais contiennent des informations incohérentes et inconsistantes par rapport au contexte fourni [3].

Cette année, la campagne d'évaluation SemEval a proposé la tâche SHROOM (*a Shared-task on Hallucinations and Related Observable Overgeneration Mistakes*)[1]. L'objectif de la tâche est de classifier des hypothèses (phrases générées par un modèle) par rapport à une verité terrain dans trois situations différentes (traduction automatique, génération de paraphrases, *definition modeling*[2]). Dans l'exemple du tableau 1, la *source* est l'entrée du modèle génératif, la *cible* est la vérité terrain, l'*hypothèse* est ce qui est généré par le modèle et le *label* est l'annotation humaine associée.

Deux corpus sont fournis : *Model-Aware* et *Model-Agnostic*. Pour le corpus *Model-Aware*, le checkpoint *HuggingFace* [3] du modèle de langue ayant généré l'hypothèse est fourni et peut être utilisé dans le système de classification, contrairement au corpus *Model-Agnostic*. A noter que dans nos approches, cette information n'est pas utilisée. Chacun des corpus est divisé en train/ trial/ dev/test, contenant respectivement 3 0000, 80, 500 et 1 500 échantillons.

CORIA-2024: COnférence en Recherche d'Information et Applications, 03–04 Avril, 2024, La Rochelle, France
nihed.bendahman@irit.fr (N. Bendahman); karen.sauvagnat@irit.fr (K. Pinel-Sauvagnat); gilles.hubert@irit.fr (G. Hubert); mb.billami@berger-levrault.com (M. Billami)

[1] https://helsinki-nlp.github.io/shroom/
[2] Il s'agit de retrouver le concept associé à une définition.
[3] https://huggingface.co/

Table 1
Exemple d'un échantillon de données, contenant la source, la référence et l'hypothèse générée par un modèle de traduction automatique.

Source : J'ai poli le plancher et les meubles.
Cible : I polished up the floor and furniture.
Hypothèse : I've got the floor and the furniture.
Label : Hallucination

2. Description des Approches Proposées

Les corpus d'apprentissage n'étant pas labellisés, nous avons évalué deux grands types d'approches non supervisées sur la tâche :

Approche à base de plongements. Nous avons généré les plongements des verités terrains et des hypothèses générées par le modèle de langue. Ensuite, nous avons calculé la similarité cosinus entre ces deux derniers. Si cette similarité ne dépassait pas un seuil prédéfini, nous avons attribué le label « hallucination » à l'hypothèse. Plusieurs modèles de génération de plongements ont été évalués: (i) Sentence-T5 [4], une variante spécialisée du modèle T5 conçue spécifiquement pour générer des représentations denses de phrases; (ii) BGE-base; (iii) BGE-large [5]; (iv) E5-Base; (v) E5-Large et (vi) E5 SF [6]. Pour chacun des modèles, nous avons fixé de manière empirique sur les collections trial et dev le seuil à partir duquel une hypothèse était classifiée comme hallucination.

Approche à base de LLMs. Nous avons testé deux LLMs : Llama2 [2] et Mistral-7B [7], avec 2 prompts différents inspirés de [8].

3. Résultats

Le tableau 2 décrit nos résultats. Les approches à base de plongements fonctionnent mieux que les LLMs testés. Notre meilleure approche a été classée 22/48 sur le *model-agnostic* et 22/45 sur le *model-aware*. Il est à noter que la première moitié du classement est extrêmement reserrée. Il faut souvent 4 décimales pour départager les participants. Une première perspective d'amélioration de ce travail s'orienterait vers le fine-tuning des modèles de plongements sur les données de la tâche [9].

Acknowledgments

Ce travail a bénéficié de l'accès aux ressources HPC de l'IDRIS dans le cadre de l'allocation 2023-AD011014740 faite par GENCI.

Table 2
Résultats des différents modèles (précision) des deux types d'approches sur les deux sous-corpus : Model-Aware et Model-Agnostic.

Modèle	M-Aw	M-Ag
Baseline (prompt LLM)	0.74	0.70
BGE-Base	0.750	0.75
BGE-Large	0.77	0.77
E5-Base	0.74	0.75
E5-Large	0.75	0.75
sentence t5 xl	**0.78**	**0.78**
SF E5	0.76	0.76

Modèle	M-Aw	M-Ag
Baseline (prompt LLM)	0.74	0.70
Llama-2-13b-chat Prompt 1	0.62	0.56
Llama-2-13b-chat Prompt 2	0.56	0.54
Mistral-7b-instruct Prompt 1	0.63	0.52
Mistral-7b-instruct Prompt 2	0.68	0.62

References

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).

[2] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288 (2023).

[3] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM Computing Surveys 55 (2023) 1–38.

[4] J. Ni, G. H. Abrego, N. Constant, J. Ma, K. B. Hall, D. M. Cer, Y. Yang, Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models, ArXiv abs/2108.08877 (2021). URL: https://api.semanticscholar.org/CorpusID:237260023.

[5] S. Xiao, Z. Liu, P. Zhang, N. Muennighoff, C-pack: Packaged resources to advance general chinese embedding, ArXiv abs/2309.07597 (2023). URL: https://api.semanticscholar.org/CorpusID:261823330.

[6] L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder, F. Wei, Text embeddings by weakly-supervised contrastive pre-training, ArXiv abs/2212.03533 (2022). URL: https://api.semanticscholar.org/CorpusID:254366618.

[7] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de Las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b, ArXiv abs/2310.06825 (2023). URL: https://api.semanticscholar.org/CorpusID:263830494.

[8] P. Manakul, A. Liusie, M. J. F. Gales, Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models, ArXiv abs/2303.08896 (2023). URL: https://api.semanticscholar.org/CorpusID:257557820.

[9] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, F. Wei, Improving text embeddings with large language models, arXiv preprint arXiv:2401.00368 (2023).