Anti-surprise : Une métrique complémentaire pour évaluer l'apprentissage lexical des (grands) modèles de langue

Nazanin Shafiabadi[1] Guillaume Wisniewski[2]
(1) ISIR, 4 place Jussieu, 75 005 Paris, France
(2) LLF, 8 Rue Albert Einstein, 75 013 Paris, France
nazanin.shafiabadi@isir.upmc.fr, guillaume.wisniewski@u-paris.fr

RÉSUMÉ

Un grand nombre de travaux s'appuient sur l'analyse des courbes de surprise pour évaluer la manière dont les modèles de langue capture le sens des mots au cours de leur apprentissage. Toutefois, cette approche ne considère que la capacité d'un modèle à prédire un mot dans des contextes appropriés, sans prendre en compte sa capacité à ne pas produire ce mot dans des contextes inappropriés. Pour combler cette lacune, nous introduisons une nouvelle mesure complémentaire, que nous appelons l'*anti-surprise*, qui évalue la capacité d'un modèle à ne pas utiliser un mot dans des contextes où il serait surprenant voire erroné. Nous montrons que l'analyse conjointe des courbes de surprise et d'anti-surprise permet de mieux caractériser l'acquisition du lexique par les modèles de langue [1].

ABSTRACT

Anti-surprisal : a Complementary Metric for Evaluating Lexical Learning in (Large) Language Models

Many studies have explored when and how LLMs learn to use specific words, primarily by examining their learning curves. While these curves capture a model's capacity to use words correctly in context, they often neglect the equally important skill of avoiding incorrect usage. In this paper, we introduce a new metric, *anti-surprisal*, which measures a model's capacity to refrain from using words in inappropriate or unexpected contexts. By examining both correct usage and error avoidance, we offer a more comprehensive perspective on the learning dynamics of LLMs.

MOTS-CLÉS : acquisition lexicale, surprise, anti-surprise.

KEYWORDS: lexical acquisition, surprisal, anti-surprisal.

1. Lien vers la publication acceptée pour COLING 2025 : https://aclanthology.org/2025.coling-main.443/