Perfect Language Model

Celestine Preetham Lawrence, PhD
Caring Machines

Abstract

I seek support for the development of a Perfect Language Model (PLM), as a sustainable alternative to Large Language Models (LLM).

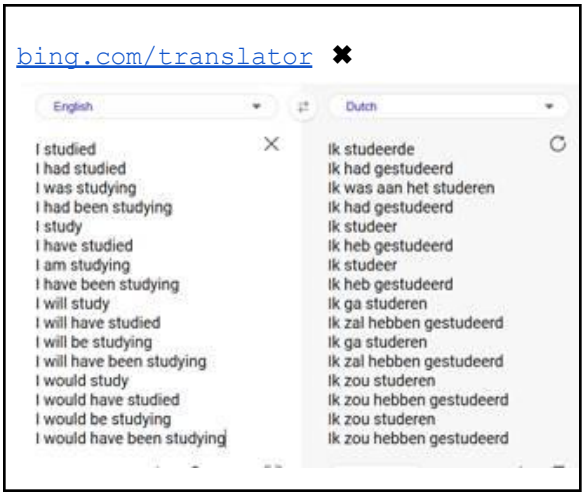
Introduction

LLMs are based on polysemantic languages, where words have multiple meanings depending on the context. This ambiguity is unsuitable for scientific discourse, which requires a monosemantic (perfect) language, such as the *characteristica universalis* envisioned by Leibniz in the 17th century. I identified faults in previous attempts to create perfect languages, like Loglan (1955) and Lojban (1987),and have developed a novel algorithm. This stems from having over a decade of research expertise in energy-efficient, unconventional, and neuromorphic computing.

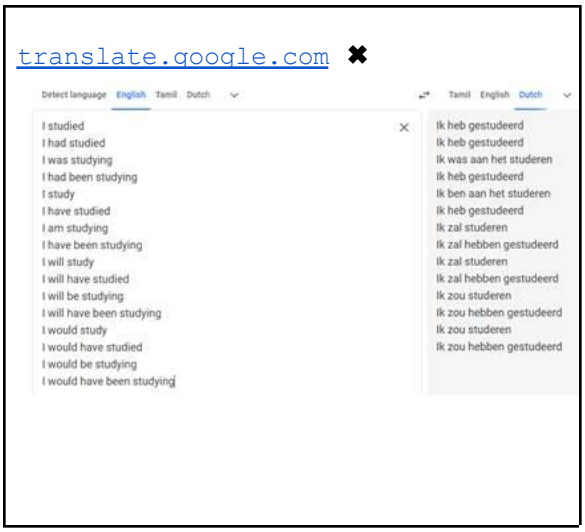
- Relevance to Wikimedia: Fact-checking the internet is not only a laborious process, but also leads to culture wars and differences in political opinion. The development of a Perfect Language Model will help build software tools that can define objective truths and refine subjective wisdom.
- Specific research agenda: I will focus on two concrete and simple tasks, for this proposal.

(1) Subject-verb translations. As shown in the following figures, none of the existing AI translators are capable of accurately expressing the nuances of the differences between subject-verb sentences like “I will study” and “I will be studying” in Dutch. An algorithm developed by me at Caring Machines [1] helps solve this issue in multiple languages, and I seek funding to complete my work. I have found that relying on grammarists [2] to translate rarely used expressions in Dutch is not readily accepted by the public. Hence, we need to survey native speakers and not solely rely on obscure grammar rules to build a PLM.

bing.com/translator ✖



deepl.com ✖



Perfect language model ✔

Ik studeerde
Ik had gestudeerd
Ik was aan het studeren
Ik was wezen studeren
Ik studeer
Ik heb gestudeerd
Ik ben aan het studeren
Ik ben wezen studeren
Ik zal studeren
Ik zal hebben gestudeerd
Ik zal aan het studeren zijn
Ik zal wezen studeren zijn / Ik zal aan het studeren zijn geweest
Ik zou studeren
Ik zou hebben gestudeerd
Ik zou aan het studeren zijn
Ik zou wezen studeren zijn / Ik zou aan het studeren zijn geweest

(2) Multi-lingual syntax trees. I have used GPT4.5 to generate 1000s of sentences in 7 different languages and split the sentences into coloured phrases, where each color is supposed to denote the words playing a particular syntactic role. It does NOT work (you may zoom and verify the following figures). Thus, we need human annotation to build an unambiguous and multilingual syntax tree.

Date: July 1, 2025 to June 30, 2026.

Related work

As shown in the figures above, much of the contemporary work towards using AI for Language Modelling is by deploying deep learning methods. Logical forms of word nets through manual annotation are available in English (https://wordnet.princeton.edu/) but are ambiguous. Multilingual word nets are even more underdeveloped (https://omwn.org/).

Methods

I will employ several research assistants to collect data by surveying language enthusiasts and academics. The analysis and roadmap to build a PLM are under development and will remain a trade secret until maturity. I

occasionally blog about intermediate insights [3].

Expected output

- A website with millions of multilingual subject-verb translations (an extension of https://chimera.university/plm)
- A website with thousands of multilingual and unambiguous syntax trees
- Presentations in Wikimania and other conferences, universities.
- A whitepaper on how a PLM will help in fact-checking the internet.

With additional funding of $500, I can convert the above task-specific PLM to a public API for https://api.wikimedia.org/

Risks

The project relies on my unique expertise and know-how. It lacks a second author.

Community impact plan

I have envisioned the proposal to employ 10 research assistants to increase the collaborative reach of the project. We will tap into our social networks and attend language cafes in different cities in the Netherlands and Europe, with branded attire, to generate buzz about the important goal of building a PLM, and how WikiMedia supports it. I hope PLMs will inspire volunteers who are worried about the overreach of big-tech controlled LLMs. I am open for further suggestions, as the project is being delivered.

Evaluation

Scientifically, a positive review from two academics about the outputs of the project is necessary to establish credibility.

Personally, if I have 1000 users willing to donate some amount after using the PLM under development, I will consider the project a success.

Budget

A total budget of 7980 EUR (~ $9000) is requested for 10 part-time research assistants, travel and dissemination.

[PLM] Research Fund Budget

References

[1]https://www.linkedin.com/posts/caring-machines_perfect-language-model-activity-7314417459996827648-d8CO?utm_source=share&utm_medium=member_desktop&rcm=ACoAACewgxMBIhmw-zNgBCCf4fDYeDVG9_AYCyI
[2]https://www.dutchgrammar.com/en/?n=Verbs.Ot03
[3]https://www.linkedin.com/pulse/essential-verbs-perfect-language-model-caring-machines-fzaqc