Natural language processing and its implications in education and medicine

Purdue System Thinkers (PurSysT)

Yan Cong cong4@purdue.edu
School of Languages and Cultures

Leveraging pre-trained large language models for aphasia detection in English and Chinese speakers

Yan Cong, Jiyeon Lee, Arianna N. LaCroix
Purdue University
{cong4, lee1704, anlacroi}@purdue.edu

Introduction

Aphasia: an acquired neurogenic language disorder, most often caused by stroke

Manually assessing language disorders, such as stroke induced aphasia, is labor and cost intensive, especially in low-resource non-English settings.

Introduction

Most aphasia studies with NLP approaches focus on monolingual English speakers (Salem et al., 2023; Purohit et al., 2023; Sanguedolce et al., 2023; Ortiz-Perez et al., 2023).

Fewer on the non-English population (Smaïli et al., 2022; Chatzoudis et al., 2022; Balagopalan et al., 2020).

Aphasia studies in Chinese speakers adopt NLP methods (Balagopalan et al., 2020; Shivkumar et al., 2020; Mahmoud et al., 2020; Qin et al., 2022).

Current study: utilize pre-trained large language models (LLMs) derived *surprisals* to detect aphasia in Chinese speakers, and examine how surprisals relate to the clinical manifestation of aphasia.

Introduction

Surprisals (a token's negative logarithmic probability given previous context) measure the unexpectedness of a sequence in a context. **c.f. page 17**

Surprisal has been discussed in both computational psycholinguistic and clinical literature (Futrell et al., 2018; Rezaii et al., 2023a, 2022; Van Schijndel and Linzen, 2018; Wilcox et al., 2018; Michaelov and Bergen, 2020, 2022a,b; Michaelov et al., 2023; Ryu and Lewis, 2021; Cong et al., 2023; De Varda and Marelli, 2022).

Current study:

(1) Implement LLMs surprisals for aphasia detection in Chinese speakers
(2) Compare LLMs surprisals in Chinese datasets with those in English

Experiment - datasets

All datasets were drawn from the AphasiaBank (MacWhinney et al., 2011 https://talkbank.org/DB/).

Participants: monolingual English or Mandarin Chinese speakers, with a Western Aphasia Battery Revised Aphasia Quotient (WAB-R-AQ, Kertesz, 2007) of 92 or lower in the aphasia group.

Chinese dataset: matched sample (on age, education, sex), 1756 observations for each group (healthy control and aphasia), with tasks picture description and story retelling; subtyping aphasia - randomly sampled balanced sets for Broca's and anomic aphasia (N=86).

English dataset: same methods, N=1586 in aphasia detection and severity measurement, N=86 in aphasia subtyping.

Experiment - aphasia detection

Three tasks in both English and Chinese datasets:

(1) Detecting the presence of aphasia
(2) Detecting aphasia subtypes
(3) Detecting aphasia severity

Logistic regression classifiers classify aphasia and control (task 1) and Broca's and anomic aphasia (task 2).

Elastic net regressions predict WAB-R-AQ scores (task 3).

Experiment - LLMs details

Each LLM read in an utterance in text and output a surprisal score.

Mean surprisal: token-wise surprisals averaged over the utterance.

Hypothesis: higher surprisals, as an indicator of larger amount of grammatical unacceptability, are associated with higher severity of aphasia.

Three pre-trained LLMs:

(1) GPT2 (Radford et al., 2019; Zhao et al., 2019, 2023b)
(2) Llama2-7B (Touvron et al., 2023)
(3) BERT (*bert-base-chinese* for Chinese and *bert-base-uncased* for English) (Devlin et al., 2019, 2018)

Operation: minicons (Misra 2022)

Experiment - feature selection

Predictor variable:

(1) utterance length (MacWhinney et al., 2011; Fromm and MacWhinney, 2023; Fromm et al., 2022, 2020)
(2) utterance level mean surprisal computed by pre-trained LLMs (Rezaii et al., 2023a).

A preliminary experiment focusing on one utility (i.e., LLMs surprisal) in a cross-linguistic clinical setting

Results - LLMs' performance in aphasia presence and subtypes detection

LLMs are more effective in subtyping than detecting the presence of aphasia in **Chinese** speakers.

LLMs showed the inverse pattern for detecting aphasia in **English** speakers.

- Crosslinguistic difference

- Character-level tokenization

Task	Acc	Prec	Rec	F1-score	AUC
Presence	0.61	0.61	0.61	0.61	0.63
Subtype	0.86	0.86	0.86	0.86	0.93

Table 1: Evaluation of logistic regression classifiers using LLMs surprisals in Chinese aphasia detection.

Task	Acc	Prec	Rec	F1-score	AUC
Presence	0.79	0.79	0.79	0.79	0.86
Subtype	0.54	0.54	0.54	0.54	0.51

Table 2: Evaluation of logistic regression classifiers using LLMs surprisals in English aphasia detection.

Results - LLMs' performance in aphasia severity detection

English tasks: the two decoder LLMs showed negative effects, Llama2 showed the strongest coefficients.

Chinese tasks: utterance length matters, all LLMs showed negative coefficients, Llama2 gave the largest coefficients.

- scaling improves performance in both English and Chinese tasks.

- clinical application: a critical need to pre-train LLMs in the target language

Dataset	MAE	utterance length	GPT2	Llama2	BERT
English	14.97	0.00	-0.55	-3.05	1.56
Chinese	7.61	0.55	-0.03	-0.37	-0.06

Table 3: Elastic net regression models in predicting English and Chinese aphasia severity.

Qualitative error analysis

Participant group	Utterance	LLMs surprisals
(1) Aphasia	mhm. okay. mhm. okay. it's fun. the kid throws the ball. and it. oops. on the window. and the guy's dad wasn't just boom. and here comes the ball. and then he looks up. that's pretty funny.	*Llama2* 3.11; *GPT2* 3.81
(2) Healthy control	a young boy is kicking a ball and crashed through a window. startled the man. and he looked up at the cracked window.	*Llama2* 3.44; *GPT2* 4.51
(3) Aphasia	yeah. yeah. alright. book. mow oh boy. boy. woe. shakes. ball. hey books. yeah. jay balls. balls six. oh boy balls. bugs.	*Llama2* 4.2; *BERT* 19.16
(4) Healthy control	oh gee.	*Llama2* 5.43; *GPT2* 5.26

Table 4: Unexpected output given by the English LLMs in picture description tasks.

Participant group	Utterance	Literal translation	LLMs surprisals
(1) Aphasia	两个人两只动物比谁走跑得快	two people two animals compare who walk run faster	*Llama2* 3.25; *GPT2* 4.4
(2) Healthy control	后来呢兔子和小乌龟比赛跑	then hare and small tortoise compete to run	*Llama2* 4.06; *GPT2* 5.1
(3) Aphasia	就是到医院医院医院然后就是做了这个就是看了这个头	so to the hospital hospital hospital then just do this just look at this head	*Llama2* 3.09; *GPT2* 3.61
(4) Healthy control	不识字呀	do not recognize characters	*Llama2* 5.15; *GPT2* 5.2

Table 5: Unexpected output given by the Chinese LLMs in picture description tasks.

Qualitative error analysis

- Extremely short utterances turn out to give rise to large surprisal scores in both Chinese and English datasets.
- For BERT, the effect of utterance length is not salient.
- Specific words that may lead to outstanding LLMs surprisals: interjection, filler words, low frequency words, and sentence final particles (in Chinese).
- Level of pre-processing matters.

Conclusion

Leverage pre-trained LLMs to detect the presence, subtypes, and severity of aphasia in English and Mandarin Chinese speakers.

Without fine-tuning, taking pre-trained LLMs off-the-shelf can inform us how surprisals distribute in aphasic individuals whose first language is not English.

English LLMs exhibit decent accuracy in detecting the presence of aphasia; the Chinese counterparts demonstrate satisfactory performance in subtyping aphasia.

Pre-trained LLMs have clinical potential (e.g., automatic aphasia diagnosis), especially in the context of multilingual populations.

Acknowledgements

We acknowledge Emily Tumacder's and Cameron Pilla's help with compiling the datasets and optimizing the machine learning pipeline. We thank Emmanuele Chersoni, Sunny Tang, Phillip Wolff, and Sunghye Cho for their inspirations. We appreciate anonymous reviewers' constructive and helpful comments.

Leveraging AI in Language Learning

Leveraging AI in Language Learning

1. The keys to the cabinet are on the table. [GPT2 Surprisal 38.88].
2. The keys to the cabinet is on the table. [GPT2 Surprisal 42.76].
3. Olivia bought a German shepherd. The dog was docile and friendly. However, it bit her hand. [GPTNeo Surprisal 6.38].
4. Olivia bought a German shepherd. The dog was unpredictable and violent. However, it bit her hand. [GPTNeo Surprisal 9.77].

Leveraging AI in Language Learning

1. The keys to the cabinet are on the table. [GPT2 Surprisal 38.88].
2. The keys to the cabinet is on the table. [GPT2 Surprisal 42.76].
3. Olivia bought a German shepherd. The dog was docile and friendly. However, it bit her hand. [GPTNeo Surprisal 6.38].
4. Olivia bought a German shepherd. The dog was unpredictable and violent. However, it bit her hand. [GPTNeo Surprisal 9.77].