Comparison of neural activity for appreciation of Japanese tanka in human brain and artificial intelligence

（ヒト脳と人工知能における短歌の鑑賞に関する神経活動の比較）

Shotaro Shiba Funai　@ Okinawa Institute of Science and Technology

Collaborators (mostly in random order)

This research is carried out as a project of IURIC (Inter-University Research Institute Corporation).

Satoshi Iso (Leader)　@ High Energy Accelerator Research Organization (KEK, 高エネ研)

Junichi Chikazoe, Naokazu Goda, Norihiro Sadato　@ National Institute for Physiological Sciences (NIPS, 生理研)

Daichi Mochihashi, Shinsuke Koyama　@ Institute of Statistical Mathematics (ISM, 統数研)

Masayuki Asahara　@ National Institute for Japanese Language and Linguistics (NINJAL, 国語研)

From outside of IURIC, young researchers also participate in this project:

Teppei Matsui　@ Okayama University & JST PRESTO

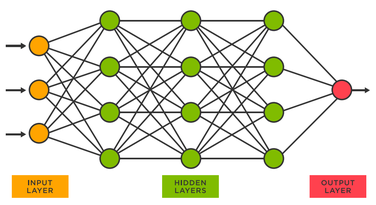
Yutaka Shikano　@ Gunma University, Chapman University & JST PRESTO

Hirono Kawashima　@ Keio University

Human brain and artificial intelligence

- We use machine learning as a method of artificial intelligence.
- Machine learning has an artificial neural network with layers, whose structure is similar to human brain.
- Linguistic machine learning is now applied for text classification, summarization, translation among various languages, …
- It seems to understand not only meaning of words but also context of sentences.

e.g., BERT *[Devlin et al. 2018]*, GPT *[Radford et al. 2018]*



https://www.tibco.com/

Our main question

Does the internal state of linguistic machine learning correspond to human brain activity when machines and humans read verse sentence with indirect implications?

nontrivial context

The verse sentence comprehension by machines is not fully studied yet!
(compared to ordinary sentences)

→ We chose Japanese tanka as an example of poem.

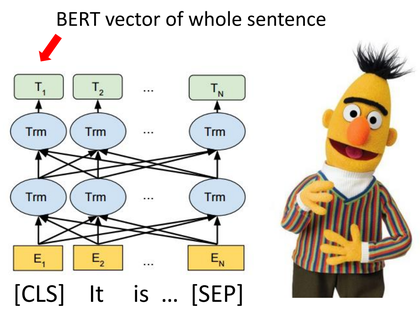
- It is a short poem with only 31 syllables.
- It often has meanings beyond its literal sense to make us emotional.

「研修中」だったあなたが「店員」になり真剣な眼差しがいい
You were “in training” and become a “clerk” then have good intent look.

BERT vector representations

- We use BERT (a popular linguistic machine learning) with the pretrained model (cl-tohoku/bert-Japanese, whole-word-masking, bpe).
- This model is trained with Japanese articles in Wikipedia: almost all the sentences the machine learned are non-poetic.
- BERT has a word embedding layer + 12 encoder layers, and outputs 768-dimensional vector representations.
- The first vector (for [CLS]) is usually regarded as the representation of the whole sentence.
- Some researchers claim that shallow layers grasp syntactic properties while deep layers grasp semantic ones (but no consensus yet).

[Tenney-Das-Pavlick, 2019]

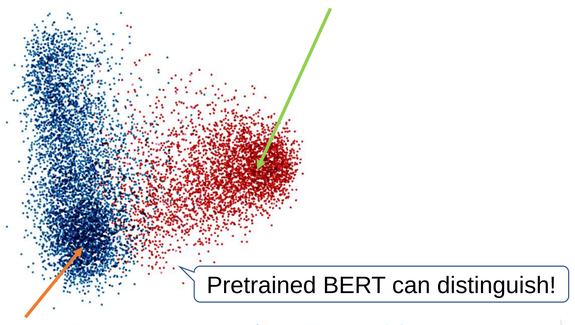


https://medium.com/@cdathuraliya/

Using the BERT vectors, we picked out tankas and nonverse sentences which are relatively similar but can be mostly distinguished.

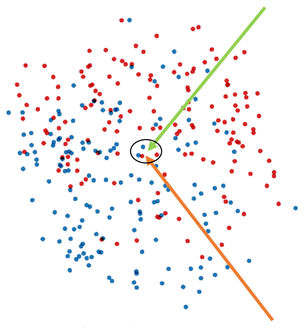
Our database (from NINJAL) → picked out → 150 tankas + 150 nonverse sentences

美しき声より孤児となりゆきし少女なるべしいま秋立つ
あふれつつ四国の海の鳴る夜を汝が追憶は断たねばならぬ



後継者も息子たちの中から選ばれる可能性が高い
健康診断やワクチン接種は保険から給付されない

好きだったあなたのくれたハンカチの匂い月面着陸をした
すでに豊作を報じる者とかかわりなく一本一本稲植える農婦



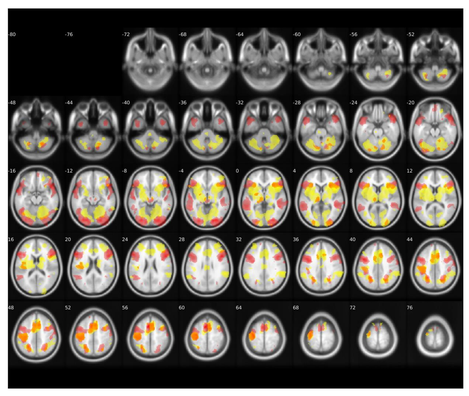
ロまわりの感覚が実に繊細になるのがひしひしとわかる
椿の花そして島の中心にある活火山の三原山

fMRI experiment

- Participants: 21 healthy young adults
- Functional Imaging: 3.0T scanner (at NIPS)
  GE-EPI, TR = 0.75s, TE = 31ms, flip angle = 55°,
  72 slices, multiband factor = 8, voxel size = 2.0 x 2.0 x 2.0 mm
- Preprocess: Realigned, slice timing corrected, normalized, using SPM12

We divide a tanka (or a sentence) into 3 lines and show them with interval of 3 seconds.
After that, we ask the participants if they feel it is poetic or not.

Result of fMRI measurement



Red: reaction to tanka stimulation

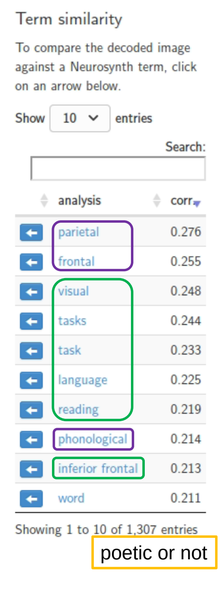
Language area and visual cortex (as expected)

Yellow: difference of reaction between poetic and non-poetic

Neurosynth suggests terms related to language processing and cognition, as expected.

Notable brain activations were found in precuneus, ventromedial PFC, left temporoparietal junction.

This may show that interaction of emotion and cognition takes place in these regions.



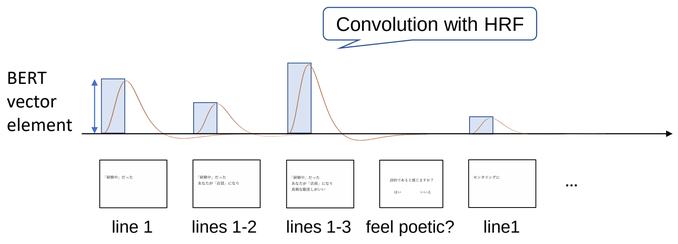
Correspondence to BERT layers

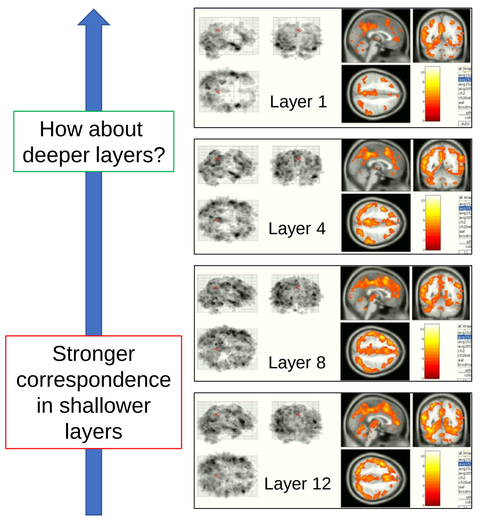
BERT vectors (768 dims)

- For line 1, lines 1-2, all 3 lines of a tanka/sentence
- In layer 1 (deepest), 2, 3, …, 12 (shallowest)

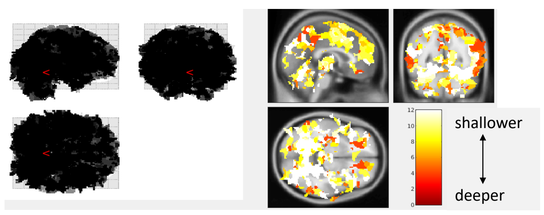
PCA (reduction to 100 dims) and use elements in each direction

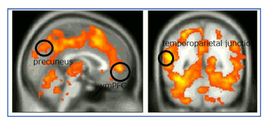
Regressors for regression analysis of fMRI data





We expect that deeper layers of BERT correspond to brain area correlated with the judgement (poetic or not) but found only weak correspondence.

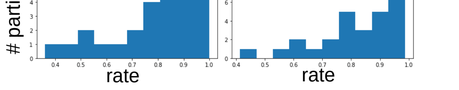




Brain area related to “poetic or not”:
Precuneus, ventromedial PFC,
left temporoparietal junction, …

A possible way to find stronger correspondence:
Finetuning BERT by each participant’s judgements

- The judgements depend on the person (the histogram has large dispersion), but we neglect it when using the pretrained model.
- Then we expect to obtain a different result after finetuning, hopefully showing clear correspondence.



Summary

- ✓ We compared the neural activity in human brain (fMRI) and artificial intelligence (BERT) when the participants and the machines read Japanese tankas.
- ✓ We found that shallower layers of the pretrained BERT are strongly correlated with brain reactions in various area.
- ✓ We specified the brain area correlated with the judgements whether poetic or not, but didn’t find its clear correspondence to deeper layers of BERT (which presumably grasp semantic properties).
- ✓ We are now finetuning BERT with each participant’s judgements and will check if deep layers of the finetuned model show the correspondence to the brain area as we expect.