Modeling performance differences on cognitive tests using LSTMs and skip-thought vectors trained on reported media consumption

Maury Courtland[1], Aida Davani[2], Melissa Reyes[3], Leigh Yeh[2], Jun Leung[3], Brendan Kennedy[2], Morteza Dehghani[23], and Jason Zevin[13]

University of Southern California 1: Dept. of Linguistics 2: Dept. of Computer Science 3: Dept. of Psychology

Introduction

Cognitive tests seek to:

- Measure a latent construct of interest
- Minimize idiosyncrasies to maximize validity

One size does *not* fit all in tests using language:

- Tests are uniform: standardized to static corpora
- **BUT** subjects have diverse language backgrounds

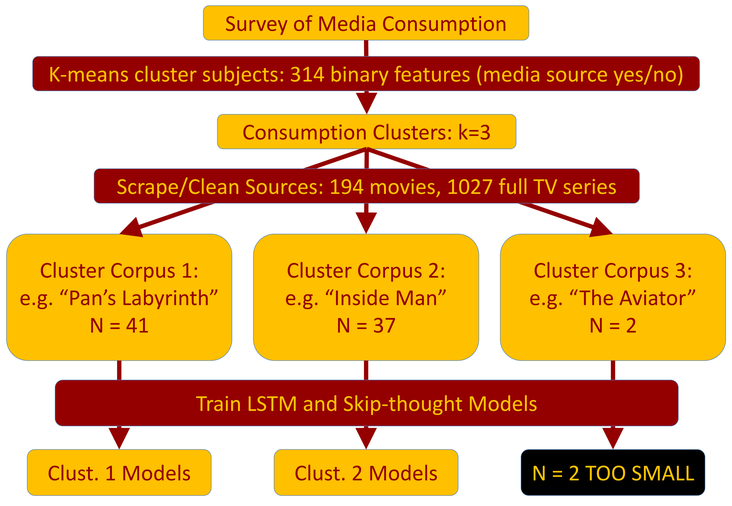
Diversity of language exposure *matters* for testing:

- We previously used the SPiN test to measure hearing: the task is to report the **last word** of a sentence in noise
- Ability to predict **last word** from context offers advantage: subjects whose language fits the test are better at this
- We surveyed what media people consume and found it has a significant effect on their performance, $F(1,76) = 7.30$, $p < .01$
- Effect was orthogonal to demographics (SES, schooling, etc.): **your media choices uniquely predict your test performance**

Hypothesis: Our findings reflect differences in language background reflected in the media people report consuming

Aim: To train models on the language of media sources to predict test performance, supporting the role of media diet

Methods

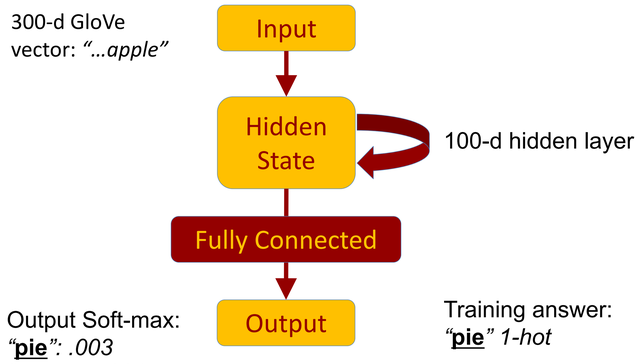


Models

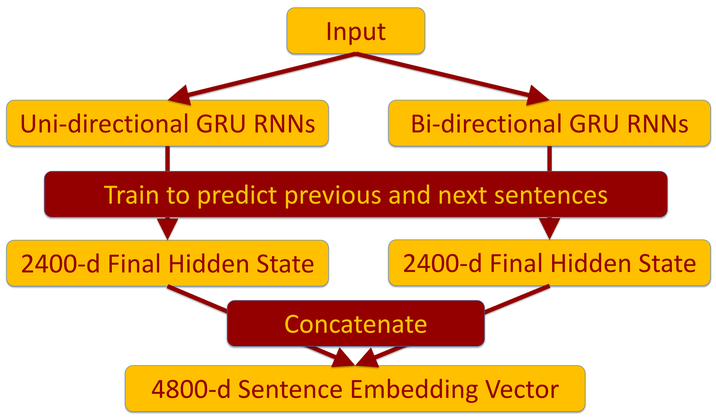
Neural Cloze Modeling: LSTMs

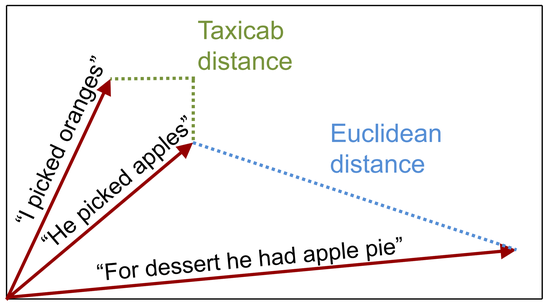
- Predicts sentence-final words given the rest of the sentence: Crucial to SPiN test (last word is test answer)

Ex. "For dessert he had apple **pie**."



Semantic Modeling: Skip-thought Vectors





Skip-thoughts encode semantic similarity as distance

Results

LSTM

- Soft-max activation of each test item's last word is rank correlated with cluster performances on that item

Skip-thoughts

- Each item's mean distance from all corpus items is rank correlated with cluster performance on that item

Rank correlation coefficients (ρ)

	Cluster 1	Cluster 2
Vanilla LSTM	.39	.46
Attn. LSTM	.31	.29
Taxicab	.49	.52
Std. Euclid.	.41	.44

- All models are significantly correlated with test performance
- Attention model correlations are all weaker
- High cloze model activation examples:
  - Clust. 1: "The dealer shuffled the **cards**."
  - Clust. 2: "For dessert he had apple **pie**."

Conclusion & Future Work

Takeaway:

- Our results indicate that the language you consume in media can significantly predict your performance on the SPiN test
- We do not currently know if this is causal (i.e. you learn from media sources) or just correlated (i.e. media is a proxy)

Next Steps:

- Increase number of sources in corpora
- Decrease model correlation to increase specificity
- Retrain language models on larger corpora
- Generate test items that are equally hard rather than uniform to ensure test fairness across diverse language backgrounds

References:

Poster Copy: