Language models can outperform empirical predictability in predicting eye movement data

Poster · August 2019
DOI: 10.13140/RG.2.2.27252.50562

CITATIONS 0
READS 98

4 authors:

Markus J. Hofmann
Bergische Universität Wuppertal
59 PUBLICATIONS 1,616 CITATIONS

Steffen Remus
University of Hamburg
15 PUBLICATIONS 167 CITATIONS

Chris Biemann
University of Hamburg
215 PUBLICATIONS 2,306 CITATIONS

Ralph Radach
Bergische Universität Wuppertal
110 PUBLICATIONS 2,741 CITATIONS

Some of the authors of this publication are also working on these related projects:

Project
Modelling eye movements in reading View project

Project
Text and Trend Mining, Language Technology, Computational Linguistics View project

All content following this page was uploaded by Markus J. Hofmann on 15 July 2019.
The user has requested enhancement of the downloaded file.
Language models can outperform empirical predictability in predicting eye movement data

Markus J. Hofmann1*, Steffen Remus2, Chris Biemann2 & Ralph Radach1

1 General and Biological Psychology, University of Wuppertal; 2 Language Technology, University of Hamburg; * Email: mhofmann@uni-wuppertal.de

Introduction: Language Models provide a deep explanation of semantic long-term memory, because they reflect the 3 basic stages of the memory process...

Table 1: Sample sizes of EM and CCP data

<table>
<thead>
<tr>
<th>Training Set</th>
<th>Number of Data Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>WIKI</td>
<td>54M</td>
</tr>
<tr>
<td>SUB</td>
<td>7.3M</td>
</tr>
<tr>
<td>TVT</td>
<td>7K</td>
</tr>
<tr>
<td>PSC</td>
<td>7K</td>
</tr>
<tr>
<td>SFD</td>
<td>70-1200m</td>
</tr>
</tbody>
</table>

Table 2: Significantly better predictions than previous model in blue, asterisks indicate significance levels (SRs). In PSC, all GAMS provided better predictions than the previous GAMS. Therefore, we only report (relative) large increases in explained variance (in blue). If SR provided significant increases in the same comparison (*P < 0.05).

Table 3: Significant predictors in M30 (*P < 0.05; **P < 0.01; ***P < 0.001).

References

- 3 WIKI-based LMs together provide better predictions than CCP (M27; Table 2)
- No general training corpus advantage, SUB better in 3, Wiki in 1 out of 6 comparisons (M28)
- Taking together both training corpora improves the predictions: general knowledge + spoken language (M29)
- > Replicating and extending Hofmann et al.’s (2017) linear approaches to SFD, N400 and CCP) data for the PSC
- When LMs are trained by larger English corpora, a small CCP sample doesn’t improve EM prediction (SRC). In a large CCP sample and smaller German training corpora, CCP improves EM predictions (PSC; M30)
- Single LMs vs. CCP: LMs win in SRC (6 vs. 5), but loose in PSC (2 vs. 4, smaller LM training corpora, larger CCP-sample). CCP is better for GD data, but LMs are often better for predicting SFD data (Table 4)
- M30: In a large EM data set, all predictors except some item random terms are significant predictors: Nearly all “item-level” variance explained? (Table 3)
- Topic: Lower F-values than N-gram & RNN in SRC, but highest in PSC for TVT. Late long-range semantics? Effects are distributed across training corpora, EM measures and position (WIKI last→GD, SUB→SFD,TVT).
- RNN-based predictability of next word is best in WIKI training: general(ized) knowledge.

N-gram: conditional probability of present and last word is best in WIKI training: general(ized) knowledge.

Topic: Probability of words to occur together in documents is approximated by N = 200 topic identities. p(w|z) = ∑ topic-document-matrices Word-topic & topic-document-matrices

Table 4: Deviance (d) and F (F0.05). Blue = LM is better; red = CCP is better.

Table 5: Significant predictors in M30 (*P < 0.05; **P < 0.01; ***P < 0.001).

Stepwise GAMs

Baseline predictors: Random terms: Based (M01), and

Different types of “predictable”

Clear topic matches to previous sentence context account for long-range semantics?

N-gram:

Conditional probability that a word N occurs, given the two previous words, e.g.

WIKI: Wikipedia dump, 5M documents/articles, 87M sentences, 1.8B words

SUB: Subtitles from opensubtitles.org, 110K documents/articles, 128M sentences/utterances, 716M words

VERB: spoken language.

> 16B-frequency correlates better with lexical decision/naming times (e.g. Brysbaert et al., 2011)

German: WIKI: 114K articles / 7.7M sentences / 180M words

SUB: 7K movies / 7.3M utterances / 54M words

Visualizations

- No general training corpus advantage, SUB better in 3, Wiki in 1 out of 6 comparisons (M28)
- Taking together both training corpora improves the predictions: general knowledge + spoken language (M29)

- > Replicating and extending Hofmann et al.’s (2017) linear approaches to SFD, N400 and CCP) data for the PSC
- When LMs are trained by larger English corpora, a small CCP sample doesn’t improve EM prediction (SRC). In a large CCP sample and smaller German training corpora, CCP improves EM predictions (PSC; M30)
- Single LMs vs. CCP: LMs win in SRC (6 vs. 5), but loose in PSC (2 vs. 4, smaller LM training corpora, larger CCP-sample). CCP is better for GD data, but LMs are often better for predicting SFD data (Table 4)
- M30: In a large EM data set, all predictors except some item random terms are significant predictors: Nearly all “item-level” variance explained?

- Topic: Lower F-values than N-gram & RNN in SRC, but highest in PSC for TVT. Late long-range semantics? Effects are distributed across training corpora, EM measures and position (WIKI last→GD, SUB→SFD,TVT).
- RNN-based predictability of next word is best in WIKI training: general(ized) knowledge.

N-gram: conditional probability of present and last word is best in WIKI training: general(ized) knowledge.