Mixed Orthographic/Phonemic Language Modeling: Beyond Orthographically Restricted Transformers (BORT)

Robert C. Gale
galer@ohsu.edu

Alexandra C. Salem
salem@ohsu.edu

Gerasimos Fergadiotis
gf3@pdx.edu

Steven Bedrick
bedricks@ohsu.edu

NLP for SLP

In the field of speech language pathology, professionals work to diagnose and treat language disorders. Transcripts from people with language disorders often include speech errors transcribed with phonemes. In order to automate assessment we need models that can represent phonemes.

BORT

is a pre-trained LLM for

clinical language assessment

designed for use on mixed

phonemes & orthography

Acknowledgements

This work was funded by NIH/NIDCD award #R01DC015999 awarded to Steven Bedrick and Gerasimos Fergadiotis

Pre-Training

DATA

CMUDict pronunciations

Wikipedia orthographic context

INITIAL WEIGHTS

BART (Lewis et al., 2021)

WORD TRANSFORMS

Pronounce replace w/ IPA

Spell letter separation

Noise ins/del/repl tokens

SELF-SUPERVISED TASKS

Pron.	*Spell*	*Noise*	Example of Transformed Text
10%	—	—	he retaliates by ·s·pɹædɪŋ false rumours
—	10%	—	he ·R·E·T·A·L·I·A·T·E·S by spreading false rumours
10%	10%	—	he ·R·E·T·A·L·I·A·T·E·S by ·s·pɹædɪŋ false rumours
10%	—	5%	he retaliates by ɔɪs·pædɪŋ false rumours
—	10%	5%	he ·E·T·A·K·I·A·T·E·S by spreading false rumours
10%	10%	5%	he ·E·T·A·K·I·A·T·E·S by ɔɪs·pædɪŋ false rumours

PRE-TRAINED BORT MODELS

- **BORT-PR**
- **BORT-SP**
- **BORT-PR-SP**
- **BORT-PR-NOISE**
- **BORT-SP-NOISE**
- **BORT-PR-SP-NOISE**

Fine-Tuning

DEFINITIONS

aphasia: an acquired neurogenic language disorder, typically the result of a stroke

paraphasia: unintended production indicative of word retrieval deficit

ADD'L DATA

CMUDict pronunciations

CLINICAL DATA

AphasiaBank language assessments

OUR USE CASE

To understand the underlying cognitive/linguistic deficits of a patient, first we need to infer the target word of a paraphasia (what the person intended to say).

"EASY" TASK: *PRONUNCIATIONS*

she made a **<pʌm·p·kɪn >** (**pumpkin**) the beautiful ·hɔɹs (horse) drawn ·kæɹɪdʒ (carriage)

"HARD" TASK: *PARAPHASIAS*

she made a **<pʌm·pɪn >** (**pumpkin**) the beautiful ·hɔ (horse) drawn ·kæsəl (carriage)

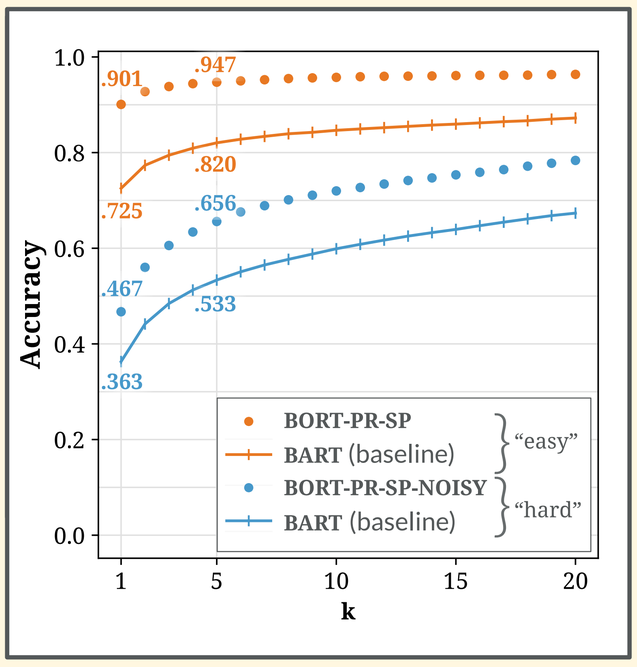
"EASY" RESULTS

Pre-training Configuration	CER Top 1	Accuracy Top 1	Accuracy Top 5
BORT-PR	.083	.869	.931
BORT-SP	.106	.843	.906
BORT-PR-SP	.057	**.901**	.947
BORT-PR-NOISE	.089	*.863*	.925
BORT-SP-NOISE	.096	.848	.911
BORT-PR-SP-NOISE	.060	.895	.947
BART-BASE	.228	.725	.820

CER = character error rate

Boldface values were significantly different from all other models, with the exception of those italicized, according to McNemar's test.

TOP-K ACCURACY



"HARD" RESULTS

Pre-training Configuration	CER Top 1	Accuracy Top 1	Accuracy Top 5
BORT-PR	.462	.451	.634
BORT-SP	.526	.401	.579
BORT-PR-SP	.447	*.456*	.641
BORT-PR-NOISE	.452	*.458*	.640
BORT-SP-NOISE	.469	.446	.625
BORT-PR-SP-NOISE	.420	**.467**	.656
BART-BASE	.606	.363	.533

MODELS & RESOURCES:

https://github.com/rcgale/bort