"This is a BERT. Now there are several of them."

Presented by

Coleman Haley

Prior work has claimed BERT has morphological knowledge...

...How do we know this knowledge isn't memorized?

We use **non-words** in singular and plural inflections in a **number agreement** task to probe BERT's accuracy at identifying whether a word was singular or plural based on the form of the word.

We consider **five languages** and **eight BERT models** (including mBERT). The languages vary in the complexity of their plural construction.

PLURAL FORMATION

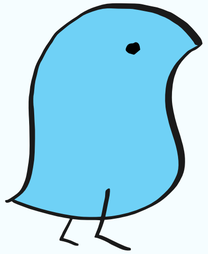
English: (usually) add -s.

French: (usually) add -s,
determiner marks for plurality

Spanish: add -s or -es,
determiner marks for plurality

Dutch: add -en or -s based on stress,
determiner informative

German: highly lexical, several patterns,
determiner infomative



BERT models can pass Wug™ tests
in English, Spanish, Dutch, French, and German
in a zero-shot context.

Language	$\text{Accuracy}_{\text{Wug}}$
English	**87%** (-13%)
French	**95%** (-1%)
Spanish	**85%** (-9%)
Dutch	**81%** (-16%)
German	**75%** (-25%)

Agreement accuracy on simple sentences
(e.g. "The author laughs." -- Det N V)

Can BERT use a priming sentence to do better?

No. Usually, it makes things worse.

Anything else worth noting?

Yes! The BERTje tokenizer (Dutch model) seems to output [UNK] for some words, despite appearing to have been trained with default settings from SentencePiece, which is supposed to be open-vocab.

Do models have a bias towards singular/plural?

It depends. BERT-Base showed a bias to plurality in English, but German models had a singularity bias.

Is there a typical pattern of error?

No. Some models misclassified most words sometimes, some models misclassified a few words all the time.

Do all models for a language perform consistently?

To a degree, but not entirely. For instance, FlauBERT for French gets only 92% accuracy on agreement even in the real-word case.

Does this have any real-world implications?

Potentially. We took the German real-word agreement data and removed the capitalization from the nouns, and found an average agreement accuracy degredation of 13%.

Condition	Stimulus	Candidates
No prime, non-words	The bik [MASK].	laugh/**laughs**
Prime, real words	This is a pilot. the pilots [MASK].	**laugh**/laughs
Prime, non-words	This is a bik. the biks [MASK].	**laugh**/laughs

Sample agreement stimuli in representative conditions in English. Correct completion is in **bold.**

[This is a BERT. Now there are several of them. Can they generalize to novel words?](#)

Coleman Haley. BlackboxNLP 2020.