On the unreasonable shallowness of natural language

AKA Adding up word embeddings works far too well

Lyndon White

School of Electical, Electronic and Computer Engineering
The University of Western Australia

We like to think language is very complicated

This is what we do, and complicated models make us feel good and publish well.

and sometimes it is

And so we need the complicated models.

but sometimes it isn’t

And so simple methods obiously work

and often it looks like it is complicated but isn’t

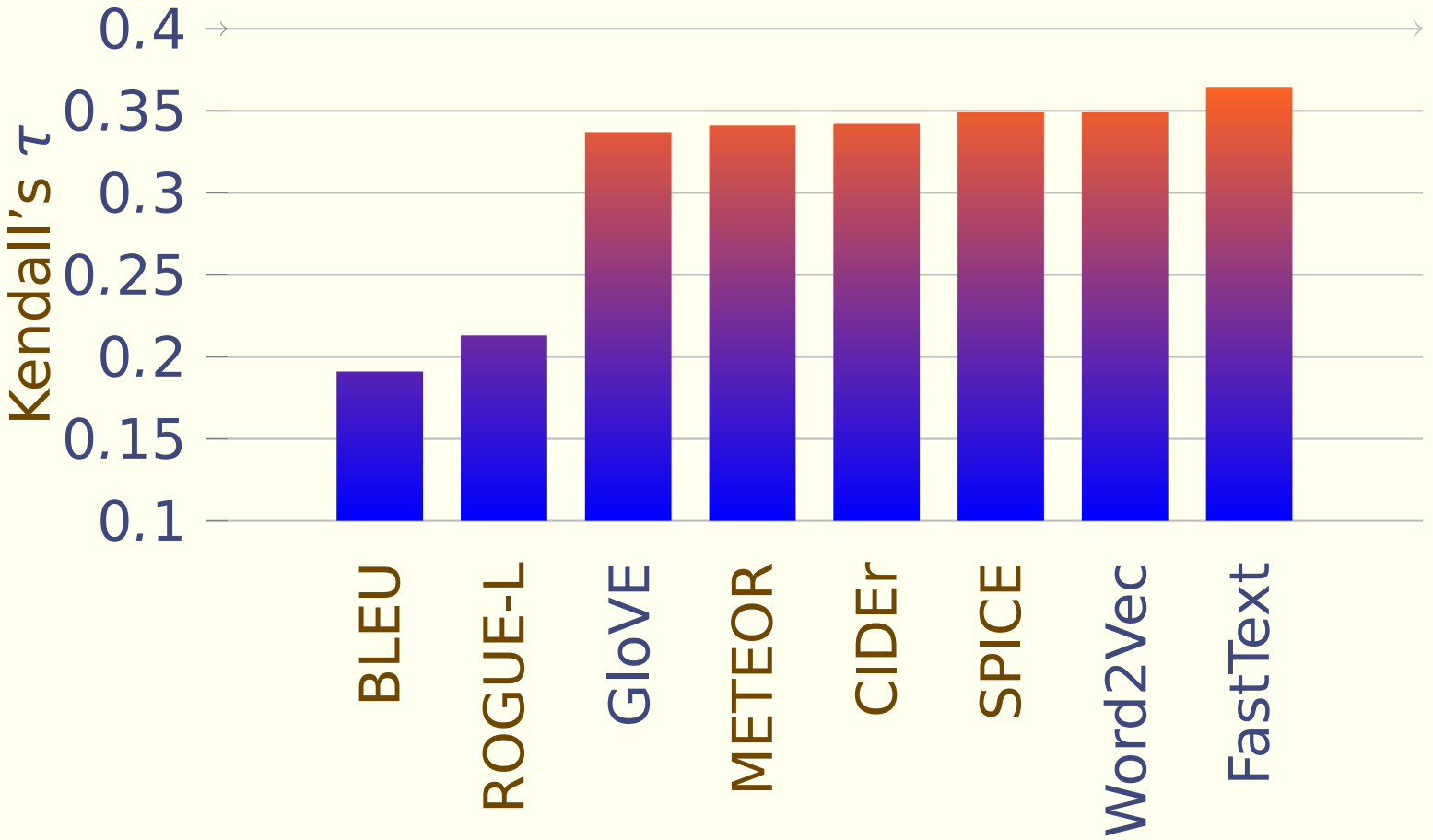
and so using complicated methods is a mistake leading to worse performance.

SkipGram etc. are basically iterative algoirthims for weighted collocation matrix factorization.

When trying to factorize very large matrices numerical linear algebraticians often use iterative methods.

Consider Machine Captioning Evaluation

Correlation with human ranking in the COMPOSITE captioning evaluation dataset. Aditya et al. 2017



∗Forthcoming publication Naeha Sharif, Lyndon White, Mohammed Bennamoun and Syed Afaq Ali Shah.

What is going on? How can a unigram method be beating everything?

Captioning quality can be assessed on **fluency** and on **adequacy**