The Primacy of Data in Deep Learning NLP for Conversational AI

Mark Johnson
Oracle Digital Assistant
Oracle Corporation
Mark.mj.Johnson@Oracle.com

ABSTRACT

Computational Linguistics and Natural Language Processing have changed considerably in the past few decades. Early research focused on representing and using linguistic knowledge in computational processes such as parsers, while these days the field focuses on practically-useful tasks such as information retrieval and chatbots. Currently our Deep Learning models have little to do with linguistic theory.

For example, the Oracle Digital Assistant is built on top of generic "Foundation" Deep Learning models. An intermediate Focusing step adapts these models to specific enterprise domains. Transfer Learning is used to refocus these models onto specific customer-oriented tasks such as Intent Classification, Named Entity Recognition, as well as more advanced models such as text-to-SQL sequence-to-sequence models. These technologies have revolutionised the application of NLP to practical problems with commercial relevance, enabling us to build better systems faster and cheaper than ever before.

Linguistic insights aren't gone from the field, however; they play a critical role in data manufacturing and evaluation. This talk explain how we use hundreds of different evaluations to understand the strengths and weaknesses of our models in the Oracle Digital Assistant, and how we automatically use this in hyper-parameter tuning. It also describes areas where additional research is still required before we can claim that NLP has become an engineering field.

CCS Concepts/ACM Classifiers

• Computing methodologies ~Artificial intelligence ~ Natural language processing; • Computing methodologies ~Artificial intelligence ~Natural language processing ~ Language resources

Author Keywords

Deep Learning NLP; Foundation Models; NLP Evaluation; Linguistics in NLP; Data and Evaluation in NLP

CIKM '21, November 1–5, 2021, Virtual Event, Australia.

ACM ISBN 978-1-4503-8446-9/21/11.
https://doi.org/10.1145/3459637.3482496

BIOGRAPHY

Mark Johnson is Chief AI Scientist, Oracle Digital Assistant and a Professor of Language Sciences, Dept of Computing at Macquarie University in Sydney, Australia.

He has been active in Computational Linguistics and Natural Language Processing for decades. He was President of the Association for Computational Linguistics (ACL) and ACL's SIGDAT, a Founding Fellow of the ACL, and an Editor-in-Chief for the Transactions of the ACL. His research spans topics such as syntactic parsing, semantic analysis and knowledge representation. He has over 200 publications across a wide range of NLP and AI topics. Most recently his research focuses on practical applications of Deep Learning to Conversational AI.

REFERENCES

Oracle Digital Assistant: https://www.oracle.com/au/chatbots/

Oracle Labs Machine Learning Research Group: https://labs.oracle.com/pls/apex/f?p=94065:12:100886501524386:7