Why The Pond Is Not *Outside* The Frog?

Grounding In Contextual Representations By Neural Language Models

Mehdi Ghanimifard

Department of Philosophy, Linguistics and Theory of Science

Thesis submitted for the Degree of Doctor of Philosophy in Computational Linguistics, to be publicly defended, by due permission of the dean of the Faculty of Arts at the University of Gothenburg, on May 27, 2020, at 15:15, in C350, Lilla Hörsalen, Humanisten, Renströmsgatan 6, Gothenburg. Faculty opponent: Parisa Kordjamshidi, Department of Computer Science and Engineering at Michigan State University.

UNIVERSITY OF GOTHENBURG

Title Why the pond is not *outside* the frog?
Grounding in contextual representations by neural language models
Author Mehdi Ghanimifard
Language English
Keywords Computational linguistics, Language modelling, Spatial language, Deep neural networks, Neural language model, Computer vision, Vision and language Meaning representation, grounded language modelling
ISBN 978-91-7833-916-7 (PRINT) 978-91-7833-917-4 (PDF)
http://hdl.handle.net/2077/64095

Abstract

In this thesis, to build a multi-modal system for language generation and understanding, we study grounded neural language models. Literature in psychology informs us that spatial cognition involves different aspects of knowledge that include visual perception and human interaction with the world. This makes spatial descriptions a compelling case for the study of how spatial language is grounded in different kinds of knowledge. In seven studies, we investigate *what* and *how* neural language models (NLM) encode spatial knowledge.

In the first study, we explore the traces of functional-geometric distinction of spatial relations in uni-modal NLM. This distinction is essential since the knowledge about object-specific relations is not grounded in the visible situation. Following that, in the second study, we inspect representations of spatial relations in a uni-modal NLM to understand how they capture the concept of space from the corpus. The predictability of grounding spatial relations from contextual embeddings is vital for the evaluation of grounding in multi-modal language models. On the argument for the geometric meaning, in the third study, we inspect the spectrum of bounding box annotations on image descriptions. We show that less geometrically biased spatial relations are more likely to deviate from the norm of their bounding box features. In the fourth study, we try to evaluate the degree of grounding in language and vision with adaptive attention. In the fifth study, we use adaptive attention to understand if and how additional bounding box geometric information could improve the generation of relational image descriptions. In the sixth study, we ask if the language model has an ability of systematic generalisation to learn the grounding on the unseen composition of representations. Then in the seventh study, we show the potentials in using uni-modal knowledge for detecting metaphors in adjective-nouns compositions.

The primary argument of the thesis is built on the fact that spatial expressions in natural language are not always grounded in direct interpretations of the locations. We argue that distributional knowledge from corpora of language use and their association with visual features constitute grounding with neural language models. Therefore, in a joint model of vision and language, the neural language model provides spatial knowledge that is contextualising the visual representations about locations.