Evaluating language understanding in Danish LLMs based on semantic dictionaries

BOLETTE SANDFORD PEDERSEN, NATHALIE C. HAU SØRENSEN, SUSSI OLSEN & SANNI NIMB

In this paper, we describe how we have generated a number of evaluation datasets – a so-called benchmark – in order to evaluate certain reasoning and understanding capacities of Danish language models. We hypothesize that the semantic knowledge already given in existing Danish dictionaries can be conceived as a 'ground truth' for the semantics of the Danish vocabulary. Our method therefore regards turning the semantic dictionaries into a number of evaluation datasets that can be used for testing how well the models understand Danish. More specifically, we examine how well the models i) understand synonymy and semantic association between concepts, ii) make inferences in relation to conceptual knowledge and inheritance structures from super-concept to sub-concepts, iii) make correct inferences in relation to acts and events, iv) disambiguate correctly when words have multiple meanings, and v) treat 'sentiment', i.e. positive and negative connotation, in running text. We test our datasets on ChatGPT 3.5 turbo and ChatGPT 4.0 and find that the models are challenged in an adequate manner even if ChatGPT 4.0 does perform very well on several of the datasets.

KEYWORDS: Danish language models; evaluation; language understanding; semantic lexicons; benchmark; ChatGPT

DOI 10.7146/nys.v1i65.145394