Using Language Models for Relevance Labelling

Paul Thomas
Microsoft

ABSTRACT

Relevance labels – annotations that say whether a result is relevant to a given search – are key to evaluating the quality of a search engine. Standard practice to date has been to ask in-house or crowd workers to label results, but recently-developed language models are able to produce labels at greatly reduced cost. At Bing we have been using GPT-4, with human oversight, for relevance labelling at web scale. We find that models produce better labels than third-party or even in-house workers, for a fraction of the cost, and these labels let us train notably better rankers. In this talk I'll report on our experiences with GPT-4, including experiments with in-house data and with TREC-Robust. We see accuracy as good as human labellers, and similar capability to pick "interesting" cases, as well as variation due to details of prompt wording. High accuracy makes it hard to improve, and I'll also discuss our work on high-quality "gold" labels and on metrics for the labels themselves.

BIOGRAPHY

Paul Thomas is a senior applied scientist at Microsoft, where he works on measurement for Bing. His research is in information retrieval: particularly in how people use web search systems and how we should evaluate these systems, as well as interfaces for search including search with different types of results, search on mobile devices, and search as conversation.