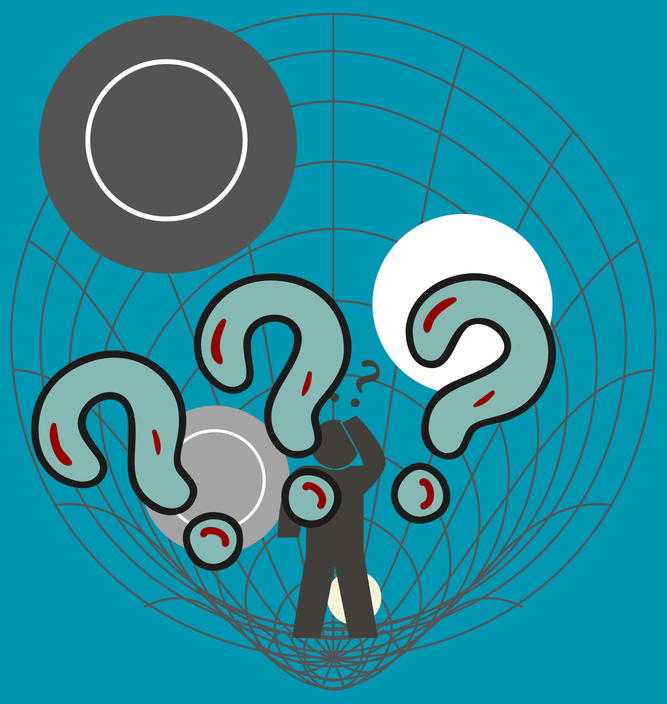
MADALINA ZGREABAN
INSTITUTE FOR LANGUAGE SCIENCES (ILS), UTRECHT UNIVERSITY

WHO WHAT WHEN DID: THE SEMANTICS OF MULTIPLE WH-QUESTIONS IN LARGE LANGUAGE MODELS



INTRODUCTION

- New frameworks account for the semantics of WH-questions (Willis, 2008; Kotek, 2016; Roelofsen & Dotlačil, 2023).
- WH-questions have mention-all or mention-some answers (Schulz & Roeper, 2011; Roelofsen & Dotlačil, 2023).
- Replying with mention-all / mention-some answers, tied to quantification (Schulz & Roeper, 2011) as:
  - **mention-all > universal quantification > exhaustive question** = Who left the party? > Madalina, Bianca and Zgreaban left.
  - **mention-some > existential quantifier > non-exhaustive question** = Where is the bathroom? > On the left.
  - knowledge about semantic quantifiers > ability to elicit exhaustive answers (Foryś-Nogala et al., 2017).

MULTIPLE WH-QUESTIONS (MWHQS)

- Two WH-elements: e.g. Which boy read which book?
- Have mention-all and mention-some answers (Roelofsen & Dotlačil, 2023);
- Languages may (not) allow MWHqs (Schulz & Roeper, 2011): possible (e.g. English, Romanian), not possible (e.g. Italian).
- Subcategorizations:
  - allowing fronted MWHqs (e.g. Romanian, Serbo-Croatian, Polish), or not allowing them (e.g. English; e.g. Cine ce când a făcut? 'Who what when did?')
  - predominantly allowing both mention-all and mention-some answers (e.g. Hindi, Romanian; Boškovic, 1998) or only allowing mention-all answers (e.g. German; Foryś-Nogala et al., 2017)

LLMS

- Q&A important: regarding MWHqs as ungrammatical or providing exhaustive answers to exhaustive questions > improved performance, user satisfaction (Bender et al., 2021) and language diversity.
- MWHqs have not been previously studied in LLMs, and questions are neglected, especially in multilingual systems (Ruder and Sil, 2021),

RQS

What are the semantic abilities of LLMs in WH-questions and, if any, how human-like are they?

SRQS

- Can multilingual LLMs capture the un/grammaticality of MWHqs?
- Do LLMs expect mention-all or mention-some answers depending on the exhaustivity of the question?
- Are LLMs fine-tuned on structures correlated with improved exhaustivity more sensitive to mention-all or mention-some answers?

FIRST EXPERIMENT

- 200 scraped sentences (Romanian and English) allowing mention-all and mention-some answers.
- Generate 200 new synthetic Romanian and English MWHqs with a LLM.
- Both corpora rephrased to have fronted MWHqs and in-situ MWHqs variants for each question > 2 corpora * language.
- Corpora translated to Italian by machine translation (simplified and validated by humans).
  - Chi cosa ha fatto? / Chi ha fatto cosa?
  - Who did what? / Who what did?
  - Cine ce a făcut? / Cine a făcut ce?
- Surprisal and perplexity evaluated after or before the appearance of the second WH-element * each language.
- Chosen models: similar parameters * each language, e.g. mT0 (Muennighoff et al., 2022), mT5 (Xue et al., 2020) GPT-4 (Achiam et al., 2023), and llama 2 (Touvron et al., 2023).

SECOND EXPERIMENT

- Two types of MWHqs sentences: allowing only mention-all (a), or both mention-all or mention-some answers (b), followed by right (c) and wrong answers (d).
  a. Who read which book?
  b. Which of these herbs grows where? (see Roelofsen & Dotlačil, 2023, p. 14)
  c. Who read which book? Madalina read 'Crime and Punishment', and Bianca read 'The Little Prince'.
  d. Who read which book? Madalina read 'Crime and Punishment'.
- Fine-tune models on scraped datasets of sentences containing quantifiers, of different sizes.
- Evaluation: surprisal and perplexity scores for both mention-some and mention-all answers for (non)-exhaustive questions.

PREDICTIONS

EXP1:

- **LLMs capture grammaticality cues:** largest surprisal for MWHqs in Italian, bigger surprisal values for fronted MWHqs in English, no difference in surprisal values for Romanian, in line with Futrell et al. (2019).
- **LLMs do not capture grammaticality cues:** no surprisal score difference across languages or stimuli, in line with Zhou et al. (2023). Influenced by insensitivity to word order, see Sinha et al. (2021).

EXP2:

- **LLMs have semantic knowledge:** bigger averaged surprisal and perplexity scores for mention-some answers given to exhaustive questions. No difference between the answers of non-exhaustive questions allowing both types of answers, in line with Gilbert et al. (2023).
- **LLMs do not have semantic knowledge:** no surprisal score difference across languages or stimuli, in line Saba (2023), and Lam et al. (2023).

LLMs and language ques: fine-tuned models will have bigger surprisal values to mention-some answers provided to exhaustive questions, in line with Frank et al. (2015), Michaelov et al. (2023). No similar ques > no difference, in line with Willems et al. (2016).

CONTRIBUTIONS

- insights into the inter-linguistic diversity of NLP tools by evaluating MWHqs, a previously understudied structure in LLMs;
- two types of new datasets: i.e. a dataset for MWHqs evaluation, and one for the semantics of their answers;
- available computational tools of a current low-resourced language, i.e. Romanian;
- awareness about the current semantic and syntactic abilities of LLMs.

REFERENCES