How Do We Answer Complex Question: Discourse Structure of Long-form Answers

Fangyuan Xu, Junyi Jessy Li, Eunsol Choi

Data & Code

Consider the question: "Why are skyscraper still washed by hand?"

😯 This question is complicated and can't be answered by span-based answer — multiple sentences are needed.
😊 Long-form question answering covers a broad set of questions + comprehensive answer!
☹️ Long-form answer is hard to evaluate: lexical-matching based metrics (ROUGE) misinform, high cognitive load for human to read.
We study: How do human organise the answer? What purpose does each sentence serve to answer the question?

Our work:
✅ A deep understanding of **discourse and pragmatics** of three types of long-form answers (ELI5[1], WebGPT[2], NaturalQuestions[3])
✅ Analysis on **model-generated answer**, showing its different discourse structure compared to human-written answer.
✅ Strong **classification result** for automatically identifying functional roles in long-form answers.
✦ We envision using functional roles for fine-grained, role-level **evaluation** & controllable and concise answer **generation**.

Ontology of Functional Roles for Long-form Answers

Organizational Sentence: inform structure of the answer
Answer Summary: provide main idea of the answer
Answer: directly address the question
Example: provide concrete or hypothetical example
Auxiliary Information: provide background or related information
Miscellaneous: (1) Specify limitation or scope of answer (2) Provide Source of the answer (3) Point to other resource (4) Express sentiment

Dataset and Example

💡 Summary ✏️ Answer 💬 Example 📖 Auxiliary Info 🧩 Misc 🎯 Org

ELI5: Why are skyscraper windows still washed by hand?

I worked on a window-washing robot that cleaned acres of rooftops over a huge commercial greenhouse. Worked great, except when it didn't, and would either break down completely or just get lost and start climbing the wrong parts of the structure. Then repair techs and manual window washers still have to be employed. I think this ends up being a cost/benefit problem where the reliability of our robots and price of implementation isn't quite at the point where it makes this commercially viable for skyscrapers. For what it's worth, I think the Twin Towers actually used a washer robot on the upper floors to limited success.

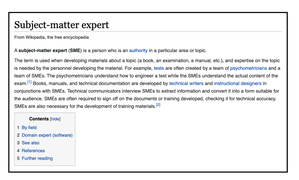
WebGPT: How do our minds estimate the passage of time?

Studies involving the brain's perception of time passage have shown a few things. One suggests the way we perceive time depends on changes in what we see. When the speed of what we see changes, our perception of time will change accordingly [2]. Another study shows the brain reacts to time passing more quickly if we are busy or enjoying something. On the contrary, when bored, the brain will decipher this as time passing more slowly [3]. It is believed we have two different systems for processing time, one for circadian rhythm and another for time passage awareness [1].

1. Your Brain Has 2 Clocks (www.scientificamerican.com)
2. Physics explains why time passes faster as you age (qz.com)
3. A stopwatch on the brain's perception of time (www.theguardian.com)

NQ: what does it mean to be a subject matter expert

A subject - matter expert (SME) or domain expert is a person who is an authority in a particular area or topic. The term domain expert is frequently used in expert systems software development , and there the term always refers to the domain other than the software domain. ...The development of accounting software requires knowledge in two different domains : accounting and software. Some of the development workers may be experts in one domain and not the other. A SME should also have basic knowledge of other technical subjects .



Annotation and statistics

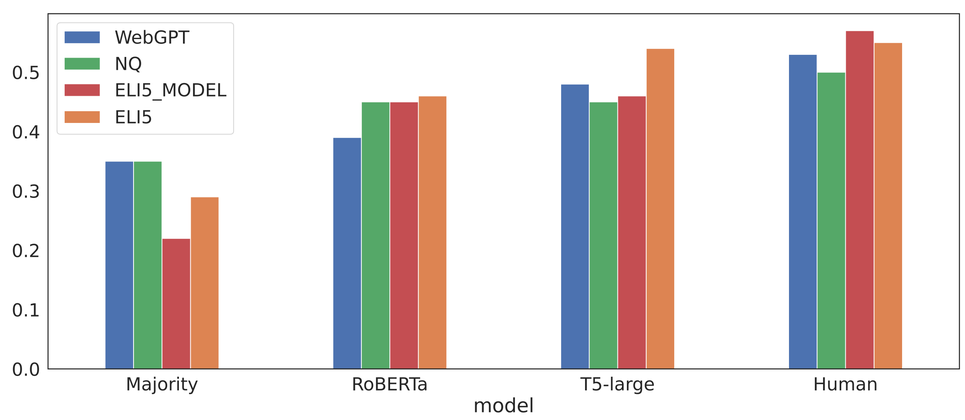
We perform two-stage annotation: (1) validity (2) for valid q/a pairs, we perform sentence-level role annotation. Each pair is three-way annotated by undergraduate linguistic students. Inter-annotator agreement - Fleiss kappa = 0.51 (validity), 0.45 (role).

Data	Validity	Role
ELI5	1,035 (6,575)	411 (2674)
WebGPT	100 (562)	98 (551)
NQ	263 (1,404)	131 (698)
Total	**1,398 (8,541)**	**542 (3,372)**

Data statistics: # answer paragraphs (# sentences)

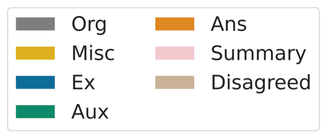
Automatic Role Classification

Task: Given a question q and its long form answers consisting of sentences s_1, s_2, … s_n, assign each sentence s_n one of the six roles.
Data: Train: ELI5; Evaluation: ELI5/WebGPT, NQ, ELI5_MODEL
Model: (1) Classification model: We use [CLS] token from RoBERTa and encodes each sentence separately to predict the role (2) Seq2Seq model: We use T5 model to encode the entire answer paragraph and output the roles sequentially.
Results: (1) In-domain test performance is comparable to human performance. (2) Performance degraded on OOD dataset, including model-generated answers.



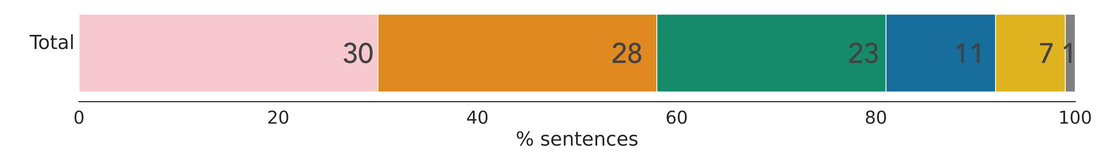
Accuracy on in-domain and out-of-domain data

Analysis

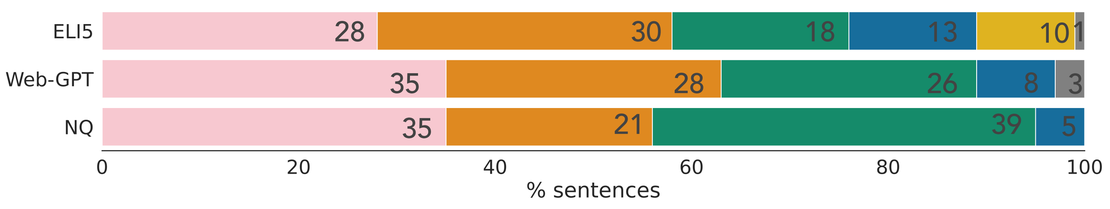


(a) Human-written answer

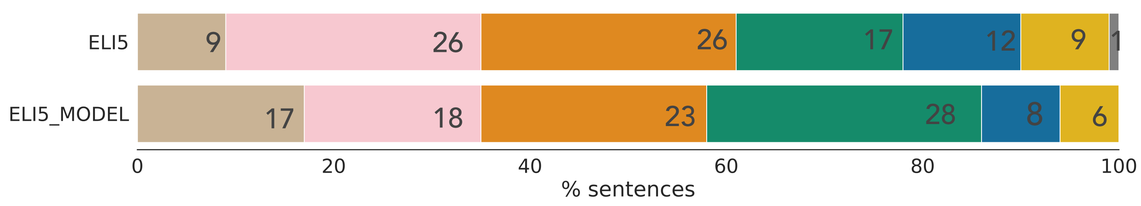
(1) ~50% sentences serve roles other than answering the question.



(2) Role distribution varies for different types of long-form answers.



(b) Model-generated answer: we annotated 194 model generated answers from a state-of-the-art LFQA system [4]. 114 of them passed validity check and are annotated with sentence-level roles:



(1) Human annotators disagree with each other more when annotating model-generated answers (kappa=0.31) -> discourse structure is less clear.
(2) Answer role distribution is different for model generated answer: more auxiliary information and less summary, example.

Reference

[1] Angela Fan et al. 2019. ELI5: Long form question answering.
[2] Reiichiro Nakano et al. 2021. Webgpt: Browser-assisted question answering with human feedback.
[3] Tom Kwiatkowski et al. 2019. Natural questions: A benchmark for question answering research.
[4] Kalpesh Krishna et al. 2021. Hurdles to progress in long-form question answering.