Automatic Evaluation of Question Under Discussion Discourse Parsers

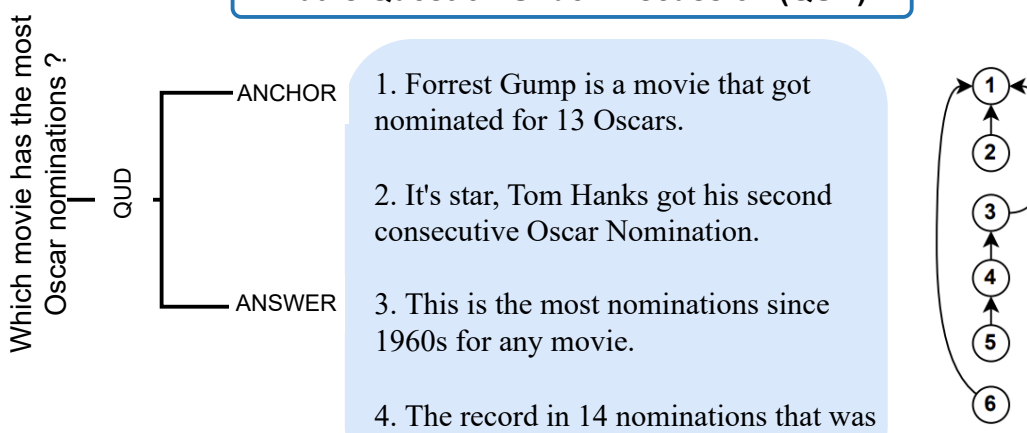
Ashima Suvarna*[1], Xiao Liu*[2], Tanmay Parekh[1], Kai-Wei Chang[1], Nanyun Peng[1]

[1] University of California Los Angeles [2] Peking University

Introduction

A linguistic framework that views every statement in a text as an immediate answer to an implicit or explicit question called QUD.

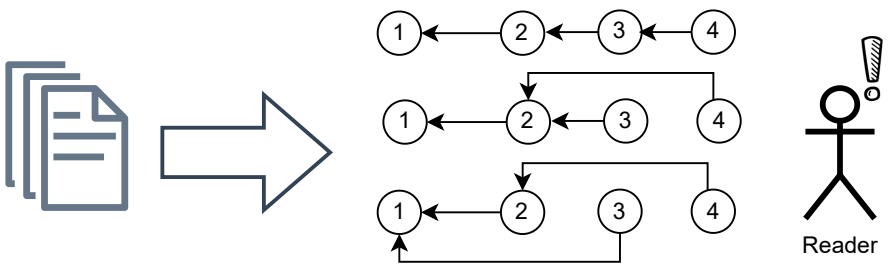
What is Question Under Discussion (QUD) ?



What criterion should a QUD satisfy ?

Answer Compatibility: The focus of answer sentence should adequately answers the QUD explicitly	Sentence 2 is a **direct and focused** answer to the QUD 'Who starred in Foresst Gump ?'
Givenness: QUDs should only contain concepts that are accessible by the reader from prior context or commonsense	'Who is the star of the movie that was shot in LA?' - This QUD **introduced a new concept** that was not in the article
Anchor Relevance: QUD should be relevant or grounded to the part of the context where it was invoked (Anchor)	Sentence 3 **fully grounds** the QUD 'Who has the most nominations at Oscars ?' answered by Sentence 4.

Motivation : Automatic Evaluation of QUD Parsers



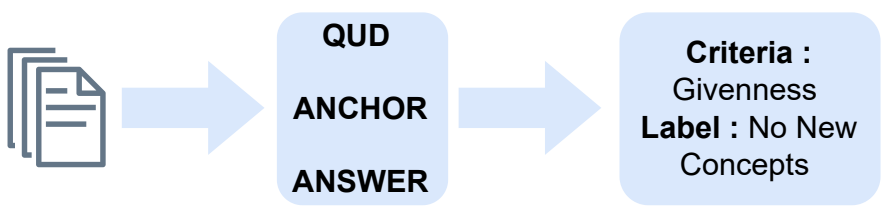
QUDs are subjective as reading depends on the reader

QUDs must satisfy theoretical constraints

Evaluating QUDs is a cognitively heavy task for humans

Experimental Setup

Can LLMs be trained to be reference-free evaluators ?



DATASET

- Each instance consists of a QUD-Anchor-Answer tuple with labels for each criterion
- **2500 QUD-Anchor-Answer** Tuples split into train/test/dev - 1300/600/600

LLM-as-EVALUATORS

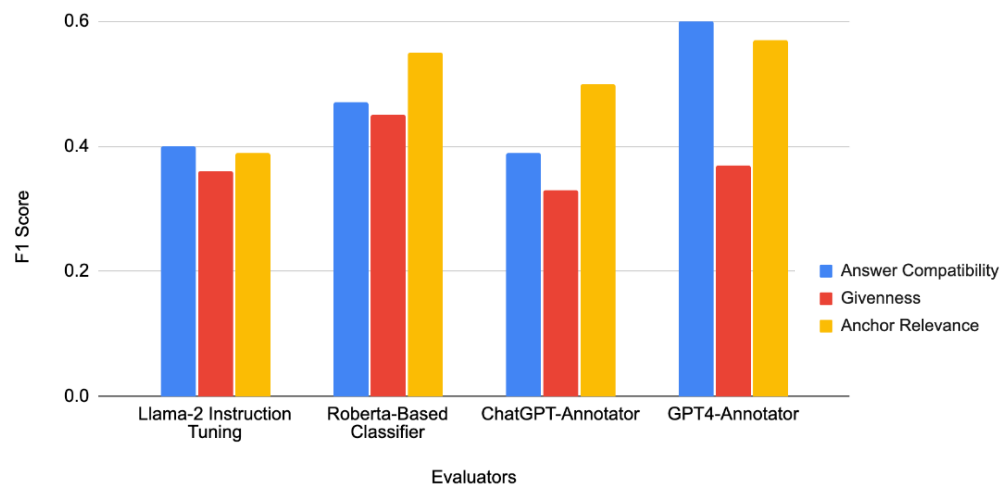
- **ChatGPT** & **GPT4** are **prompted** in a few-shot manner to evaluate
- **LLaMa2-7b-chat instruction tuned** as an evaluator for criterion
- **Roberta-large** model **finetuned** as a classifier to predict the label for each criterion given a input pair

Experimental Results

QUD Parsers

- **Human** : QUD Annotations done by linguistic students
- **ChatGPT** & **GPT4** are prompted in a 2 step approach : Anchor Prediction & Question Generation
- **Alpaca-7B** is prompted in a 2 step approach : Anchor Prediction & Question Generation
- **Ko et al 2023** refers to a pipeline approach where MLM models are used for Anchor Prediction & Question Generation

We align LLMs to the human evaluations and treat these as reference-free evaluators. We report F1 scores on a withheld test split from the human evaluation data.



We compute the system-level rankings of LLMs on the test split by comparing the percentage of 'best' class defined during human evaluations.

Answer Compatibility				Anchor Relevance			
Llama-2-IT	Human	Roberta	GPT4	Llama-2-IT	Human	Roberta	GPT4
Human	GPT4	Chatgpt	Alpaca	Ko et al	Human	Ko et al	Ko et al.
Chatgpt	Chatgpt	GPT4	Ko et al.	GPT-4	Ko et al.	Human	ChatGPT
GPT-4	Human	Human	Human	Human	Chatgpt	Chatgpt	Human
Ko et al	Ko et al	Ko et al	Chatgpt	Chatgpt	GPT-4	GPT4	Alpaca
Alpaca	Alpaca	Alpaca		Alpaca	Alpaca	Alpaca	

Givenness		
Llama-2-IT	Human	GPT4
Human	Human	ChatGPT
Chatgpt	Ko et al.	Ko et al.
GPT-4	Chatgpt	Human
Ko et al	GPT-4	Alpaca
Alpaca	Alpaca	

X	Answer Compatibility	Givenness	Anchor Relevance
LLaMa-2-IT	70%	80%	70%
RoBerta	90%		90%
GPT4	0%	50%	67%

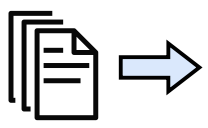
Pairwise Agreement between Human Evaluators and X where X is an LLM-based Evaluator. We also report the overall system-level rankings generated by each evaluator per criterion.

QUD Parser Development

PIPELINE

Identify Anchors from Article → Generate QUDs for Anchor-Answer

JOINT PARSING



→ Generate QUD-Anchor-Answer Tuple

GENERATE-SELECT

Generate QUD for each pair of sentences → Best QUD selection

We use MLM models like BERT and Longformer or LLaMa2-7B-Chat for QUD Parser Development

QUD Parsers	Answer Compatibility	Anchor Relevance
Pipeline (Ko et al.)	0.81	0.94
Pipeline (Llama2-7b)	0.82	0.87
Sentence-level Joint (Llama2-7b)	0.83	0.83
Article-level Joint (Llama2-7b)	0.72	0.85
Generate-Select (Comp.)	0.92	0.68
Generate-Select (Relv.)	0.74	0.99

Performance of Developed QUD Parsers evaluated by the Roberta-based Classifier. Performance is indicated by the percentage of generations classified as the 'best' label by the evaluator (Higher the Better)