Towards Filling the Gap in Conversational Search: From Passage Retrieval to Conversational Response Generation

Weronika Łajewska and Krisztian Balog
University of Stavanger, Norway

CIKM'23, Birmingham

This study

- **Problem setting:** Conversational response generation
  - It extends beyond passage retrieval + summarization
- **Goal:** snippet-level annotations of relevant passages, to enable
  1. the training of response generation models that are able to ground answers in actual statements
  2. the automatic evaluation of the generated responses in terms of completeness
- **Main contributions:**
  1. Crowdsourcing task design and protocol to collect high-quality annotations
  2. A dataset of 1.8k query-passage pairs annotated from the TREC 2020 and 2022 Conversational Assistance track

CAsT-snippets sample

Query: I remember Glasgow hosting COP26 last year, but unfortunately I was out of the loop. What was the conference about?

Passage: HOME - UN Climate Change Conference (COP26) at the SEC – Glasgow 2021 Uniting the world to tackle climate change. The UK will host the 26th UN Climate Change Conference of the Parties (COP26) in Glasgow on 1 – 12 November 2021. The COP26 summit will bring parties together to accelerate action towards the goals of the Paris Agreement and the UN Framework Convention on Climate Change. The UK is committed to working with all countries and joining forces with civil society, companies and people on the frontline of climate change to inspire climate action ahead of COP26. COP26 @COP26 · May 25, 2021 1397069926800654339 We need to accelerate the #RaceToZero Join wef, MPPindustry, topnigel & gmunozabogabir for a series of events demonstrating the need for systemic change to accelerate the global transition to net zero. Starting May 27th Learn more #ClimateBreakthroughs | #COP26 Twitter 1397069926800654339 COP26 COP26 · May 24, 2021 1396737733649846273 #TechForOurPlanet is a new challenge programme for #CleanTech startups to pilot and showcase their solutions at #COP26! Innovators can apply to six challenges focusing around core climate issues and government priorities.

CAsT-snippets sample

Query: I remember Glasgow hosting COP26 last year, but unfortunately I was out of the loop. What was the conference about?

Passage: HOME - UN Climate Change Conference (COP26) at the SEC – Glasgow 2021 Uniting the world to tackle climate change. The UK will host the 26th UN Climate Change Conference of the Parties (COP26) in Glasgow on 1 – 12 November 2021. The COP26 summit will bring

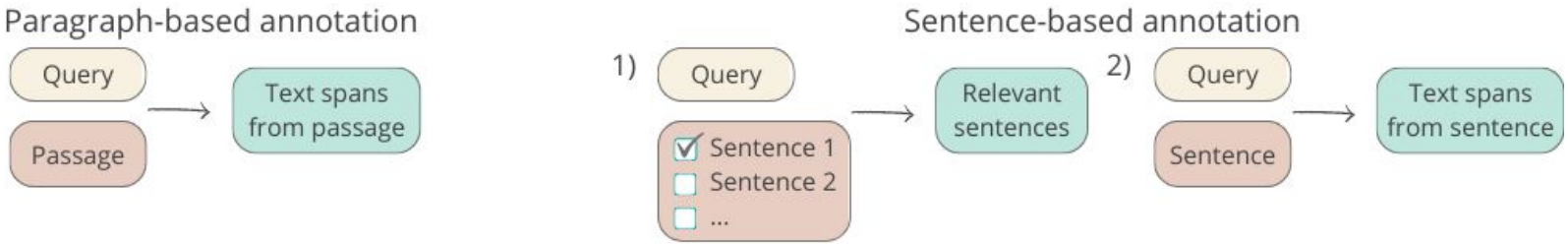
The seemingly straightforward task of highlighting relevant snippets turns out to be not that simple.

We need to accelerate the #RaceToZero Join wef, MPPindustry, topnigel & gmunozabogabir for a series of events demonstrating the need for systemic change to accelerate the global transition to net zero. Starting May 27th Learn more #ClimateBreakthroughs | #COP26 Twitter 1397069926800654339 COP26 COP26 · May 24, 2021 1396737733649846273 #TechForOurPlanet is a new challenge programme for #CleanTech startups to pilot and showcase their solutions at #COP26! Innovators can apply to six challenges focusing around core climate issues and government priorities.

Preliminary study

A comparison of different task designs, platforms, and worker pools

- **Task designs**: paragraph-based vs. sentence-based annotation

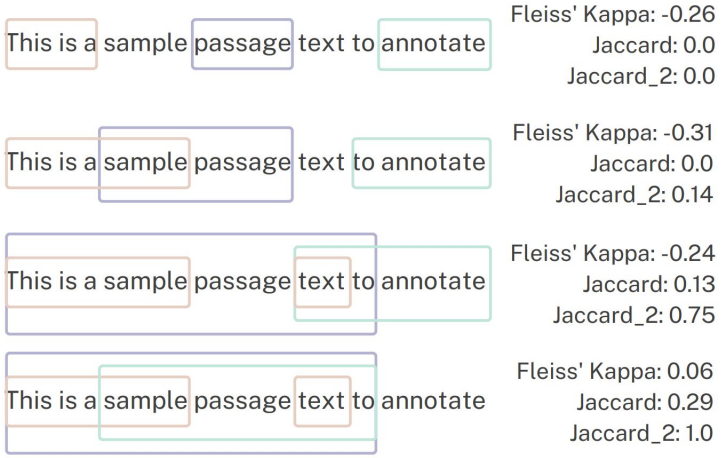


- **Platforms and workers:**
  - Amazon MTurk (regular vs. master workers)
  - Prolific
  - Expert annotators (PhD students)

Evaluation measures

Traditional measures of inter-annotator agreement are insufficient

- Fleiss' Kappa and Krippendorff's Alpha are measures for categorical annotations that rely on a binary notion of agreement
- **Here:** we need to measure the degree to which snippets selected by different workers overlap
    - Inter-annotator agreement: Jaccard similarity (also a less strict variant, k-Jaccard)
    - Similarity against expert annotators: "ROUGE-like" variant of precision and recall



Results

Inter-annotator agreement

Task variant	Annotators	Jaccard	Jaccard_k		
			k = 4	k = 3	k = 2
Paragraph-based	MTurk regular (n=5)	0.02	0.08	0.21	0.48
	MTurk master (n=5)	0.18	0.35	0.53	0.73
	Prolific (n=5)	0.14	0.27	0.44	0.65
	Expert (m=3)	0.25	-	-	0.54
Sentence-based	MTurk regular (n=3)	0.35	-	-	0.71
	MTurk master (n=3)	0.47	-	-	0.76

Comparison to expert annotations

Task variant	Annotators	F1
Paragraph-based	MTurk regular	0.36
	MTurk master	0.54
	Prolific	0.50
Sentence-based	MTurk regular	0.31
	MTurk master	0.41

Main findings

- Relative ordering: MTurk masters > Prolific > MTurk regular
- Paragraph-level > sentence-level (w.r.t. similarity with expert annotations)

⇒ use MTurk and paragraph-based design for the large-scale data collection

Data collection

Setup

Employ a small group of trained crowd workers, selected through a qualification task, and create an extended set of guidelines with help of the annotators

Task consisted of: a detailed description of the problem, examples of correct annotations, a quiz, and 10 query-passage pairs to be annotated

20 workers completed/15 passed

Initial guidelines

Feedback on qualification task

Extended guidelines

Performed in daily batches (1 topic/batch =~46 HITs)

Individual feedback after each submitted batch

General comments/suggestions on a common Slack channel

$0.3 per HIT +$2 bonus for completing within 24h

Resulting dataset: CAsT-snippets

371 queries, top 5 passages per query ⇒ **1855 query-passage pairs** (each annotated by 3 crowd workers)

- Data quality
  - Inter-annotator agreement exceeds even that of expert annotators
  - Similarity with expert annotations is on par with MTurk master workers
- Comparison against other datasets
  - More snippets annotated per input text; also, snippets are longer

Dataset	Input text	Avg. snippets length (tokens)	# snippets per annotation
CAsT-snippets	Paragraph	39.6	2.3
SaaC	Top 10 passages	23.8	1.5
QuaC	Wikipedia article	14.6	1

Challenges identified

Challenges pointed out by the crowd workers that need to be addressed in conversational response generation:

- Only a partial answer is present
- Temporal considerations
  - Spans may need to be excluded given the time constraints in the query
  - Assessing temporal validity can be challenging based on the paragraph alone (without larger context)
- Subjectivity of the passages originating from blogs or comments
- Indirect answers that require reasoning and background knowledge
- Determining the appropriate amount of context to include in each span
  - Balancing between being concise and being self-contained
- Determining whether the evidence or additional information is needed or an entity alone is sufficient as an answer

Summary

- Snippet-level annotations for conversational response generation (information-seeking queries)
- Several measures to ensure high data quality
  - Preliminary study to compare task variants and crowdsourcing platforms
  - Providing feedback and training to annotators throughout the data collection process
  - Incentive structure to engage crowd workers over a period of time and avoid worker fatigue
- Communication with workers also led to various insights regarding challenges in conversational response generation

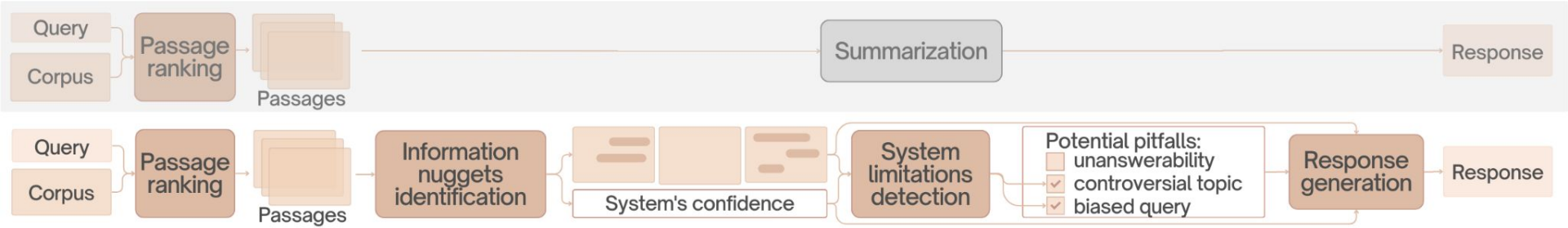
Questions?

Extended version on arXiv: https://arxiv.org/abs/2308.08911

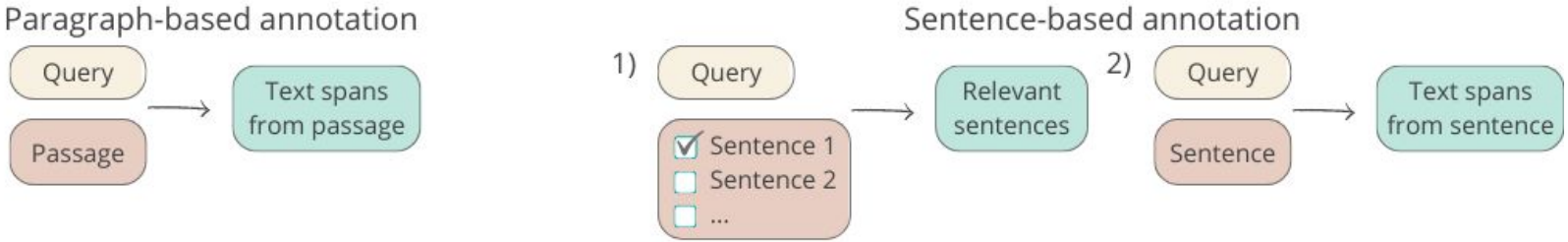
Dataset: https://github.com/iai-group/CAsT-snippets

Overview of our Approach to Conversational Response Generation

"A true teacher would never tell you what to do. But he would give you the knowledge with which you could decide what would be best for you to do."
— Christopher Pike, Sati



Preliminary study



Dataset: TREC CAsT'20 and '22 (top 5 passages according to relevance score for each query)

Input: query + passage/sentence

Output: snippet-level annotations in passage

Task Variant	Annotator	Time	# workers	Acceptance rate	Cost
Paragraph	MTurk regular	182s	5	50%	$0.36
	MTurk master	63s	5	90%	$0.38
	Prolific	154s	5	79%	$0.51
	Expert	96s	3	-	-
Sentence	MTurk regular	977s	3	72%	$0.43
	MTurk master	305s	3	87%	$0.56

Evaluation Measures

Traditional measures of inter-annotator agreement are insufficient

- Fleiss' Kappa and Krippendorff's Alpha are measures for categorical annotations that rely on a binary notion of agreement
- **Here:** we need to measure the degree to which snippets selected by different workers overlap
  - Inter-annotator agreement: Jaccard similarity (also a less strict variant, k-Jaccard)
  - Similarity against expert annotators: "ROUGE-like" variant of precision and recall

$$J(t) = \frac{|\bigcap_{i=1}^{n} snippets(t, w_i)|}{|\bigcup_{i=1}^{n} snippets(t, w_i)|},$$

$$p_t^{i,j} = \frac{|snippets(t, w_i) \cap snippets(t, e_j)|}{|snippets(t, w_i)|},$$

$$r_t^{i,j} = \frac{|snippets(t, w_i) \cap snippets(t, e_j)|}{|snippets(t, e_j)|}.$$

Results (large-scale data collection)

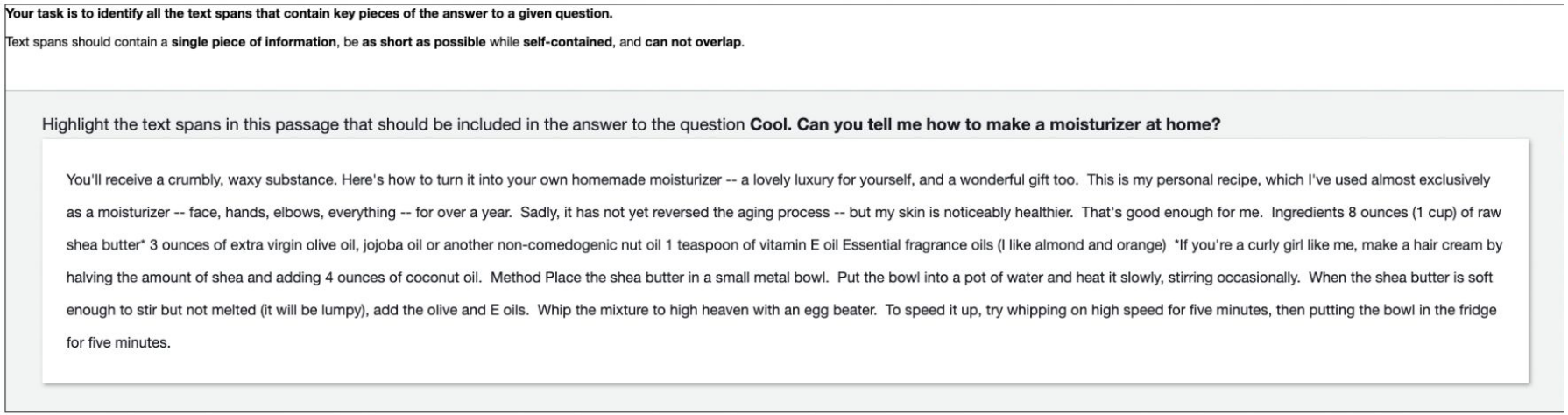
Inter-annotator agreement

Task variant	Annotator	Jaccard	Jaccard_2
Paragraph -based	MTurk regular (n=5)	0.02	0.48
	MTurk master (n=5)	0.18	0.73
	Prolific (n=5)	0.14	0.65
	Expert (m=3)	0.25	0.54
	Large-scale (topics 1,2) (m=3)	**0.38**	**0.62**
	Large-scale (all data) (m=3)	**0.33**	**0.61**
Sentence -based	MTurk regular (n=3)	0.35	0.71
	MTurk master (n=3)	0.47	0.76

Comparison to expert annotations

Task variant	Annotator	F1
Paragraph -based	MTurk regular	0.36
	MTurk master	0.54
	Prolific	0.50
	Large-scale (topics 1,2) (m=3)	**0.54**
Sentence -based	MTurk regular	0.31
	MTurk master	0.41

Amazon MTurk - paragraph-based design



Amazon MTurk - sentence-based design

Instructions:

Choose **all** sentences that contain information that should be included in the answer to the question.

Task:

Question: **How much would making my own deodorant cost?**

- [ ] Before You Start, You'll Need Coconut oil (or 1/2 as much of a liquid oil if you are allergic to coconut oil) shea butter , cocoa butter or mango butter (or a mix of all three) beeswax (pastilles)
- [ ] Optional: Vitamin E oil baking soda (Omit this if you have sensitive skin and just use extra arrowroot) organic arrowroot powder or non-gmo cornstarch 2-3 capsules of high quality probiotics that don't need to be refrigerated (I love Bio Kult brand)- optional
- [ ] Optional: Essential oils of choice – I used about 20 drops of lavender essential oil Deodorant Bar Ingredients ½ cup coconut oil ½ cup shea butter , cocoa butter or mango butter (or a mix of all three equal to 1 part) ½ cup + 1 tsp beeswax 1 teaspoon Vitamin E oil – optional 3 tablespoons baking soda
- [ ] (Omit this if you have sensitive skin and just use extra arrowroot or cornstarch) 1/2 cup organic arrowroot powder 2-3 capsules of high quality probiotics that don't need to be refrigerated (optional)
- [ ] Optional: Essential oils of choice – I used about 20 drops of lavender essential oil and also like citrus and frankincense Deodorant Bar
- [ ] Instructions Combine coconut oil, shea (or other) butter, and beeswax in a double boiler, or a glass bowl over a smaller saucepan with 1 inch of water in it.

Submit

Your task is to identify all the text spans that contain key pieces of the answer to a given question.

Text spans should contain a **single piece of information**, be **as short as possible** while **self-contained**, and **can not overlap**.

Highlight the text spans in this sentence that should be included in the answer to the question **I remember Glasgow hosting COP26 last year, but unfortunately I was out of the loop. What was the conference about?**

If countries cannot agree on sufficient pledges, in another 5 years, the emissions reduction necessary will leap to a near-impossible 15.5% every year.

Prolific

paragraph-based design

Snippet annotation task 1

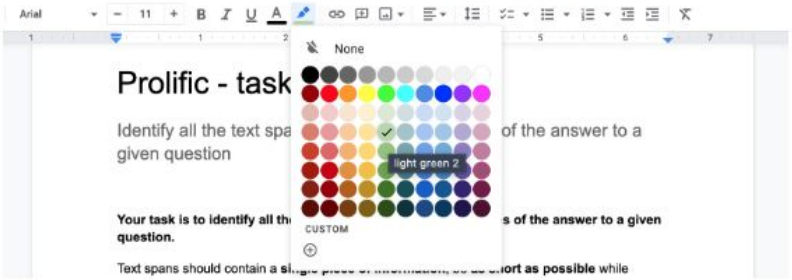
Identify all the text spans that contain key pieces of the answer to a given question

Instructions

Your task is to identify all the text spans that contain key pieces of the answer to a given question.

Text spans should contain a **single piece of information**, be **as short as possible** while **self-contained**, and **can not overlap**.

In **each** passage highlight the chosen text spans using green text highlight:



Do not edit the text of the passages!

Task

Question 1:

I remember Glasgow hosting COP26 last year, but unfortunately I was out of the loop. What was the conference about?

Passage 1:

The initial pledges of 2015 are insufficient to meet the target, and governments are expected to review and increase these pledges as a key objective this year, 2021. The updated Paris Agreement commitments will be reviewed at the climate change conference known as COP 26 in Glasgow, UK in November 2021. This conference will be the most important ...