Code and data!

Similarity or deeper understanding? Analyzing the TED-Q dataset of evoked questions

Matthijs Westera, Jacopo Amidei, Laia Mayol

Universitat Pompeu Fabra Barcelona

COLING 2020 The 28th International Conference on Computational Linguistics

Contributions:

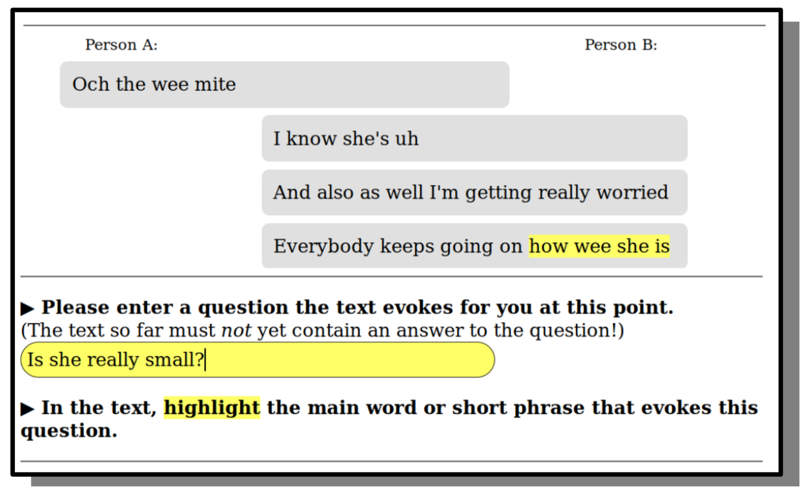
1. We turn the TED-Q dataset into a classification task and compare different notions of similarity.

2. We compare results against an analogous task extracted from the BookCorpus.

TED-Q crowdsource tool

Westera & Rohde (2019)

- Eliciting a question:


Person A: Person B:

Och the wee mite

I know she's uh

And also as well I'm getting really worried

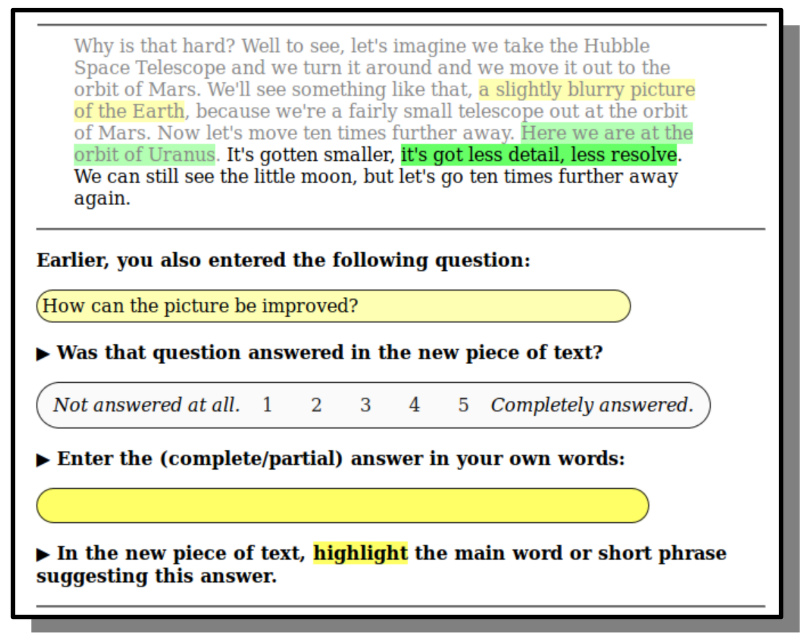
Everybody keeps going on how wee she is

▶ **Please enter a question the text evokes for you at this point.**
(The text so far must *not* yet contain an answer to the question!)

Is she really small?

▶ **In the text, highlight the main word or short phrase that evokes this question.**

- Eliciting an answer to a prior question:


Why is that hard? Well to see, let's imagine we take the Hubble Space Telescope and we turn it around and we move it out to the orbit of Mars. We'll see something like that, a slightly blurry picture of the Earth, because we're a fairly small telescope out at the orbit of Mars. Now let's move ten times further away. Here we are at the orbit of Uranus. It's gotten smaller, it's got less detail, less resolve. We can still see the little moon, but let's go ten times further away again.

Earlier, you also entered the following question:

How can the picture be improved?

▶ **Was that question answered in the new piece of text?**

Not answered at all. 1 2 3 4 5 *Completely answered.*

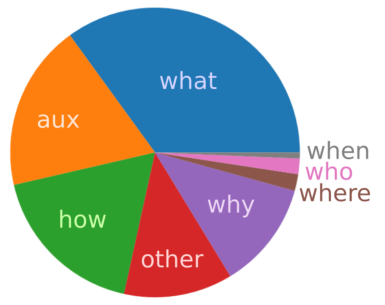
▶ **Enter the (complete/partial) answer in your own words:**

▶ **In the new piece of text, highlight the main word or short phrase suggesting this answer.**

The TED-Q dataset

Westera, Mayol & Rohde 2020 (LREC)

Elicitation phase:		Comparison phase:	
TED talks:	6	question pairs:	4516
words:	6975	participants/pair:	6
probe points:	460	participants:	163
participants/probe:	5+	judgments:	30412
participants:	111	RELATED mean:	1.21
questions:	2412	RELATED std:	0.79
answers:	1107	Agreement (AC_2):	.46
ANSWERED mean:	2.50		
ANSWERED std:	1.51		

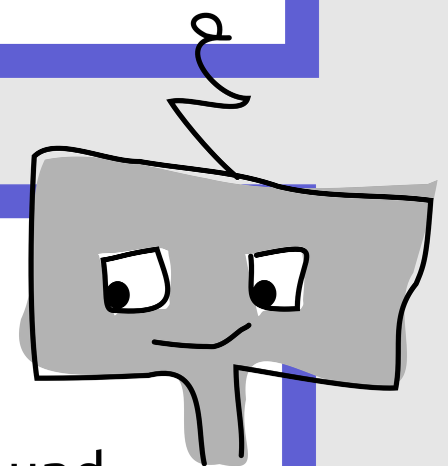


Main finding: Discourse structure tends to be more implicit when questions are better anticipated.

Task definition

Two classification tasks: *'evoked here or not?'*.

- From **TED-Q**: 4.8K items, half positive:
  - **+**: context (3 sents) + question it evoked.
  - **−**: context + random question evoked 3-4 sents away.
- From **BookCorpus**: 3.8M items, from written dialogues extracted from 11K books by quotation extraction.
  - **+/−**: Same settings as for TED-Q.



Models

1. Random decision forests based on:

- LEMMAOVERLAP: proportion of question lemmata also found in the context. (using www.spacy.io)
- GLEU: Based on matching n-grams. Wu et al. (2016)
- MEANCOS: mean cosine similarity, by GLoVe, between question words and context words. Pennington et al. (2014)
- BERTSCORE: question/context token match based on BERT embeddings, F_1-score variant. Zhang et al. (2019) Devlin et al. (2019)
- ALLSIMS: All of the above in a single model.

More semantic.

2. Fine-tuned BERT-base as our most powerful model.

Recent datasets similar to TED-Q

- Choi et al. (2018): **QuAC**: 100K Qs from unscripted dialogue.
- Rao & Daumé III (2018): 75K clarification Qs from StackExchange.
- Riester (2019) expert annotation of 'questions under discussion'.
- Pyatkin et al. (2020): **QADiscourse**, crowdsourced Q-A pairs.
- Ko et al. (2020): **Inquisitive**, 19K questions evoked by news.

Acknowledgements

We thank the anonymous reviewers for their comments. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 715154) and from the Spanish State Research Agency (AEI) and the European Regional Development Fund (FEDER, UE) (project PGC2018-094029-A-I00). This paper reflects the authors' view only, and the EU is not responsible for any use that may be made of the information it contains.

Main results

- How do the different models fare on the TED-Q-based task?

Matthew's Correlation Coefficient (MCC)

false positives

Models trained on TED-Q:

	LEMMAOVERLAP	GLEU	MEANCOS	BERTSCORE	ALLSIMS	BERT
TED-Q	0.47 (60%)	0.24 (32%)	0.14 (49%)	0.33 (30%)	0.46 (52%)	**0.55** (61%)

very superficial notion does well!

BERT does better (as expected)

- What about the BookCorpus-based task? Can it be used for pre-training?

Maybe!

Models trained on BookCorpus:

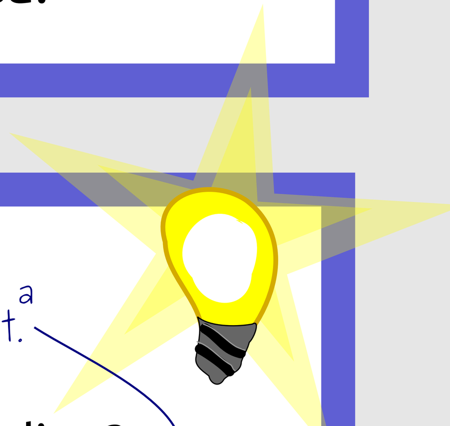
tested on	LEMMAOVER.	GLEU	COSSIM	BERTSCORE	ALLSIMS	BERT
TED-Q	**0.46** (63%)	0.28 (9%)	0.12 (25%)	0.29 (54%)	0.42 (72%)	0.43 (84%)
BookCorpus	0.18 (13%)	0.06 (27%)	0.03 (21%)	0.14 (32%)	0.17 (30%)	**0.38** (44%)

the most superficial notion seems to generalize best.

BERT better again, but task seems harder.

Further analysis

- 195 TED-Q items annotated by 2 experts (MCC=.55/.60, κ=.66).
- BERT's errors often involve general questions that fit multiple places.
- Smaller context yields higher scores; models trained on smaller context perform worse when given the full context, but not vice versa.



Conclusions

Arguably not (just) a crowdsource artefact.

- BERT best, close to human. *Some* deeper understanding?
- LemmaOverlap better than more syntactic/semantic notions.
- Predicting explicit questions harder than implicit questions.

Makes sense!