

Fact-checking as a conversation

Andreas Vlachos

<http://andreasvlachos.github.io/>



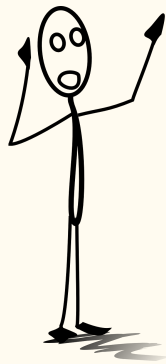
What do fact-checkers do?

The United Kingdom has ten times Italy's number of immigrants.



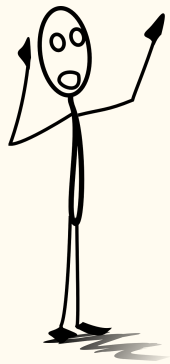
| Country/ Immigration | Italy | UK |
|-------------------------|-------|-------|
| 2014 | 4.92M | 5.05M |
| 2015 | 5.01M | 5.42M |
| 2016 | 5.03M | 5.64M |

FALSE: We find no data to support this claim. The UK does not have "ten times Italy's number of immigrants".



Automated fact-checking

The United Kingdom has ten times Italy's number of immigrants.



| Country/ Immigration | Italy | UK |
|-------------------------|-------|-------|
| 2014 | 4.92M | 5.05M |
| 2015 | 5.01M | 5.42M |
| 2016 | 5.03M | 5.64M |

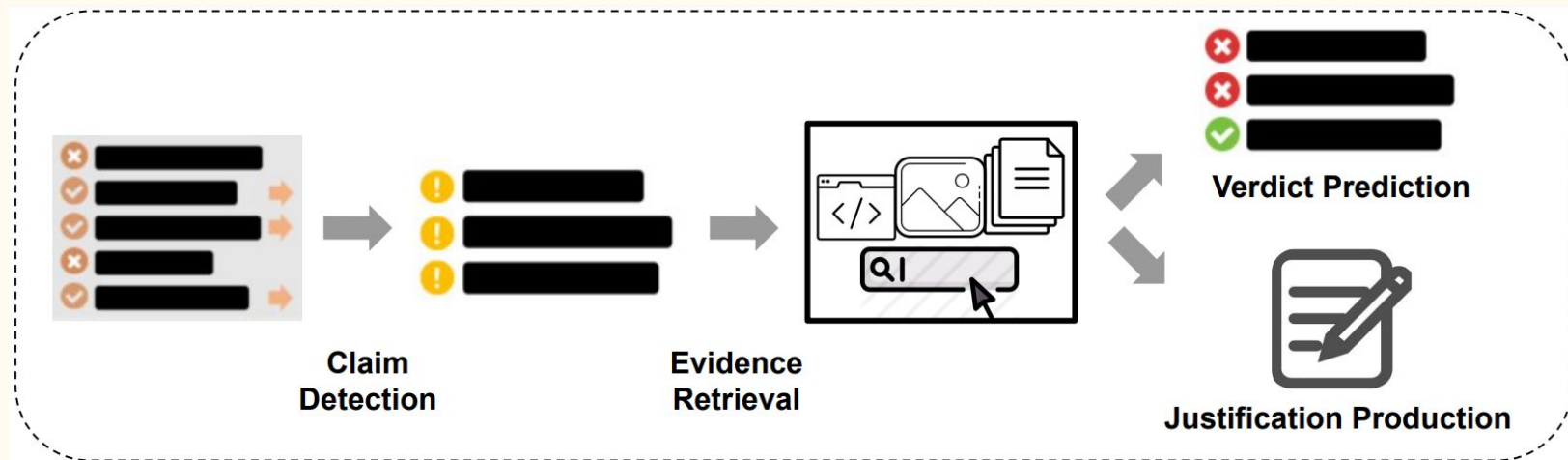
FALSE: We find no data to support this claim. The UK does not have "ten times Italy's number of immigrants".

What do we want from automated fact-checking?

- Verdict justification, a.k.a. algorithmic transparency
 - Can't convince otherwise
 - Need to check their correctness
- Generalization to different domains (economy, health, etc.)
- Learn with (relatively) little data

(Vlachos and Riedel, 2014)

Fact-checking framework (for NLP)



For more details see our surveys (*Thorne and Vlachos, 2018*;
Guo et al., 2022)

New datasets needed

AI successes follow dataset availability (*Wissner-Gross, 2016*)

| Year | Breakthroughs in AI | Datasets (First Available) | Algorithms (First Proposed) |
|---------------------------------------|--|--|--|
| 1994 | Human-level spontaneous speech recognition | Spoken Wall Street Journal articles and other texts (1991) | Hidden Markov Model (1984) |
| 1997 | IBM Deep Blue defeated Garry Kasparov | 700,000 Grandmaster chess games, aka “The Extended Book” (1991) | Negascout planning algorithm (1983) |
| 2005 | Google’s Arabic- and Chinese-to-English translation | 1.8 trillion tokens from Google Web and News pages (collected in 2005) | Statistical machine translation algorithm (1988) |
| 2011 | IBM Watson became the world Jeopardy! champion | 8.6 million documents from Wikipedia, Wiktionary, Wikiquote, and Project Gutenberg (updated in 2010) | Mixture-of-Experts algorithm (1991) |
| 2014 | Google’s GoogLeNet object classification at near-human performance | ImageNet corpus of 1.5 million labeled images and 1,000 object categories (2010) | Convolution neural network algorithm (1989) |
| 2015 | Google’s Deepmind achieved human parity in playing 29 Atari games by learning general control from video | Arcade Learning Environment dataset of over 50 Atari games (2013) | Q-learning algorithm (1992) |
| Average No. of Years to Breakthrough: | | 3 years | 18 years |

Fact Extraction and VERification (FEVER)

Claim:

The Rodney King riots took place in the most populous county in the USA.

SUPPORTED

Evidence:

[wiki/Los Angeles Riots]: The 1992 Los Angeles riots, also known as the Rodney King riots were a series of riots, lootings, arsons, and civil disturbances that occurred in Los Angeles County, California in April and May 1992.

[wiki/Los Angeles County]: Los Angeles County, officially the County of Los Angeles, is the most populous county in the United States.

- 185K claims verified on Wikipedia (Thorne et al., NAACL 2018)
- Enabled progress in system development, still far from solved

Annotation process

1. Pick a random page and sentence from Wikipedia
2. Extract a set of claims (typically shorter/simpler than the sentence)
3. For each claim:
 - a. Generate mutated claims by paraphrasing, substituting words, negations, etc.
 - b. Verify each mutated claim selecting the evidence sentence(s):
 - The claim is **SUPPORTED** by the evidence
 - The claim is **REFUTED** by the evidence
 - **NOT ENOUGH INFORMATION** in Wikipedia to verify it

Annotation process details

- 50 annotators, all native speakers, trained by the authors or more experienced annotators
- Fixed Wikipedia dump to avoid changes in labels
- One annotator constructs the claim, different annotator verifies it
- Dedicated user interfaces were developed for the task
- Guidelines were refined through pilot studies
- Advised to spend 2-3 minutes per claim
- Instructed to avoid using their own world knowledge:
is Canadian” is NOT ENOUGH INFORMATION



Annotation findings

- 0.68 in Fleiss Kappa inter-annotator agreement on 3.4K claims
- 96.12% precision and 74.84% recall in evidence retrieval: measured against annotators who were not time-constrained
- Claims were 7.9 tokens long
- Multi-sentence evidence was chosen for 28.04% of the claims
- Evidence from different pages was chosen for 11.47%
- 7.6% of the mutated claims were excluded due to being too vague/ambiguous
- Final verification by the authors: 91.2% correct on 227 claims.

Results

Unlike previous tasks and datasets, **evidence** matters:

- a correct label with incorrect supporting evidence is wrong
- a simple approach using TF-IDF-based similarity for evidence selection and DecAtt for labeling the claim given the evidence achieved 31.87% acc. (50.91% ignoring evidence)

Fact Extraction and Verification (FEVER) shared task

- EMNLP 2018 workshop with Amazon and Imperial College
- 23 participants, best performance at 64.21%, a year later 68%, now at 76.89%

Fact checking tested by BBC journalists

| Claim | Russia meddled with US elections | Q | BBC (EN) |
|--|----------------------------------|---|----------|
| Supports | | Refutes | |
| Evidence | | Evidence | |
| <p>✓ Not climate change or trade, but why didn't John Podesta give a server that wasn't his to the CIA." Several American intelligence agencies have concluded Russia meddled in last year's US elections, but Mr Trump has as recently as this week said he thinks other countries could have been responsible. [2017-07-07] <small>BBC</small></p> <div><div>correct label ? <input checked="" type="button" value="supports"/> <input type="button" value="refutes"/> <input type="button" value="other"/></div><div>relevant ? <input checked="" type="button" value="yes"/> <input type="button" value="no"/></div></div> | | <p>✗ Vice-presidential nominee Mike Pence quickly stated there will be "serious consequences" if Russia attempts to influence US elections. [2016-07-27] <small>BBC</small></p> <div><div>correct label ? <input type="button" value="supports"/> <input checked="" type="button" value="refutes"/> <input type="button" value="other"/></div><div>relevant ? <input checked="" type="button" value="yes"/> <input type="button" value="no"/></div></div> | |
| <p>✓ Kosachev: "No evidence" Russia meddled in US election [2017-09-26] <small>BBC</small></p> <div><div>correct label ? <input type="button" value="supports"/> <input checked="" type="button" value="refutes"/> <input type="button" value="other"/></div><div>relevant ? <input checked="" type="button" value="yes"/> <input type="button" value="no"/></div></div> | | <p>✗ The new US ambassador to the UK, Woody Johnson, has said Theresa May was right to air her grievances about Russia's potential influence in the US elections. [2017-11-14] <small>BBC</small></p> <div><div>correct label ? <input checked="" type="button" value="supports"/> <input type="button" value="refutes"/> <input type="button" value="other"/></div><div>relevant ? <input checked="" type="button" value="yes"/> <input type="button" value="no"/></div></div> | |
| <p>✓ Trump had come under fire for defending Russia against claims that they meddled in the 2016 US presidential elections. [2018-07-19] <small>BBC</small></p> <div><div>correct label ? <input checked="" type="button" value="supports"/> <input type="button" value="refutes"/> <input type="button" value="other"/></div><div>relevant ? <input checked="" type="button" value="yes"/> <input type="button" value="no"/></div></div> | | <p>✗ The arrest also came days after the justice department charged 12 Russian intelligence officers with hacking Democratic officials in the 2016 US elections. [2018-07-16] <small>BBC</small></p> <div><div>correct label ? <input checked="" type="button" value="supports"/> <input type="button" value="refutes"/> <input type="button" value="other"/></div><div>relevant ? <input checked="" type="button" value="yes"/> <input type="button" value="no"/></div></div> | |

MONITIO: Question generation

Crimea was part of Russia until 1954, when it was given to the Soviet Republic of Ukraine.

When was Crimea part of Russia?

When did Crimea become part of Ukraine?



- Learn to ask questions for different parts of the claim
- Better able to retrieve refuting evidence
- Recent paper (Fan et al., 2020) explored crowd-sourced fact-checking questions; adapt for journalists?
- EU H2020 with Priberam, Deutsche-Welle and Scandinavian Communications

FEVER2

Build it, Break it, Fix it for fact checking:

- **Building:** many systems with source code to help
- **Breaking:** participants create adversarial instances (manually or automatically): half given to fixers, half reserved for evaluation
- **Fixers:** fix the systems using the adversarial instances

Workshop at EMNLP in Hong Kong in November 2019

- Best breaking method was a human/machine hybrid
- See more in Thorne et al. (2019) shared task overview

SOTA architectures for FEVER

Task decomposition remains the same:

- **Document retrieval:** TFIDF/BM25, Wikipedia entity linking, deep passage retrieval
- **Evidence selection:** TFIDF/BM25, claim-sentence classification, multihop approaches
- **Veracity prediction:** Claim-evidence classification (Natural Language Inference/Textual Entailment)

Recent advances:

- Pre-training specialized to handle coreference (Ye et al., 2020)
- Graph attention over the evidence retrieved (Liu et al., 2020)

Verdict justification?

Retrieved evidence is a baseline. However, we also want to know:

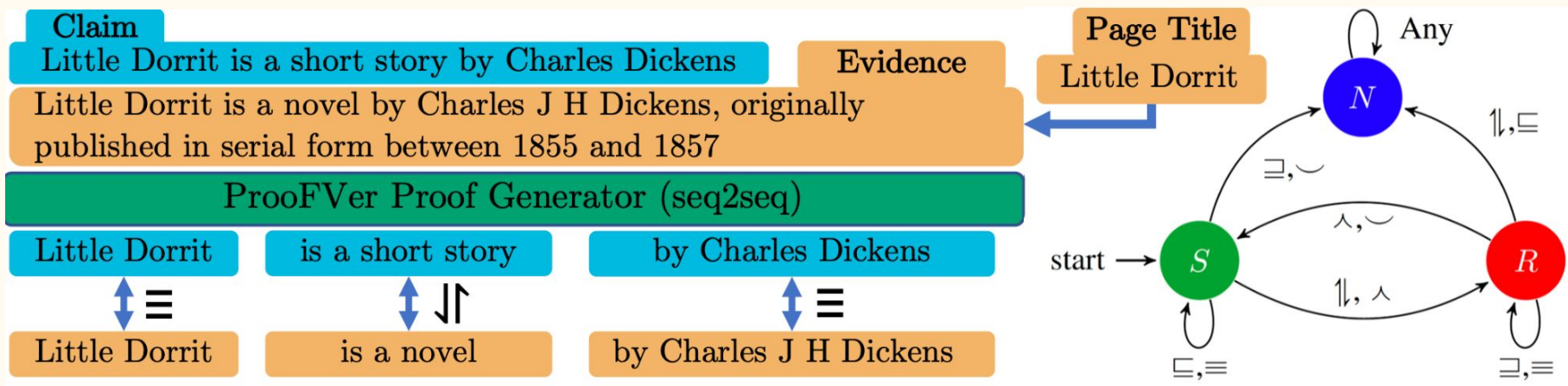
- *How* was the evidence used in the reaching the verdict?
- What were the assumptions/commonsense used?
- What was the reasoning process?

Current approaches:

- Highlight(attention)-based, e.g. Popat et al. (2018), but not clear if attention is an explanation indeed
- (Evidence) Summarization, e.g. Kotonya and Toni (2020), Atanasova et al. (2020), but does not correspond to reasoning
- *Faithfulness* is lacking in both

Proof System for Fact Verification (ProoFVer)

When comparing the claim with the evidence, we generate the proof directly and infer the verdict from it (*Krishna et al., 2022*)

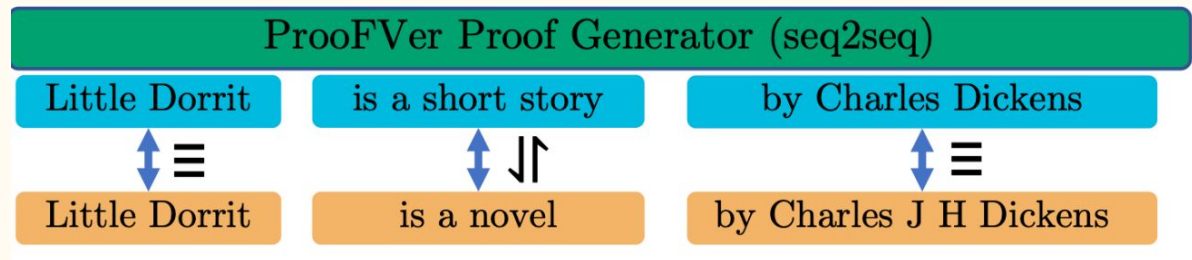


The six operators are from Natural Logic (Angeli and Manning, 2014) indicating negation, equivalence, alternation, etc.

Inference

Inference is a two-stage process (given evidence):

1. Generate the sequence of lexically aligned spans in claim and evidence labeled with NatLog operators (constrained seq2seq):



2. A deterministic finite state automaton for the verdict (fixed)

Training data for step 1 was obtained from FEVER, PPDB, Wikidata, WordNet and manual annotation (2.5% of the cases)

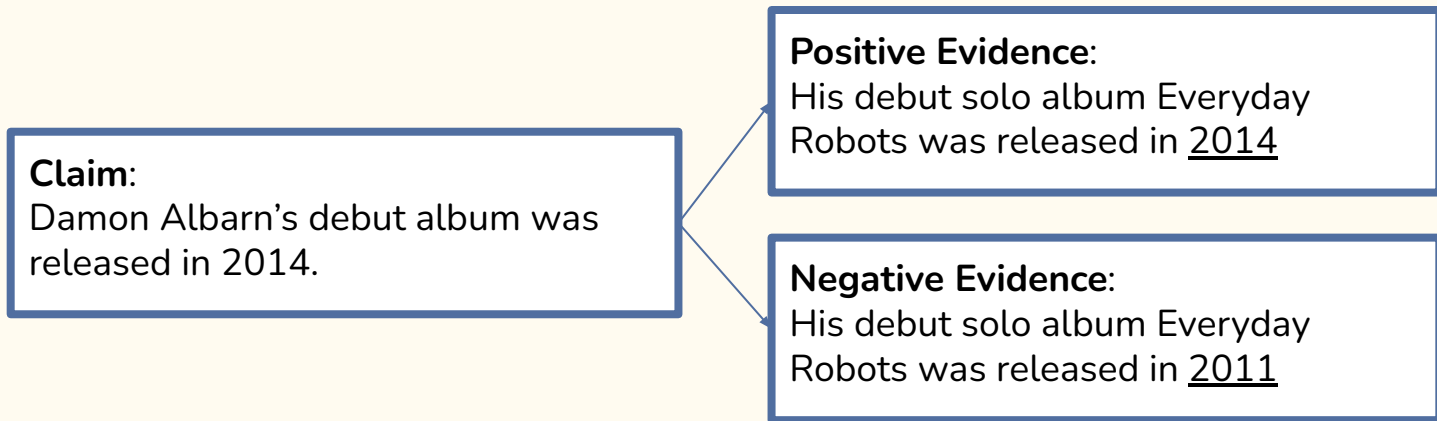
Results (FEVER blind test)

| Model | Label accuracy | Evidence+Label |
|----------------------------|----------------|----------------|
| ProoFVer | 79.25 | 74.37 |
| KGAT (Liu et al. 2020) | 74.07 | 70.38 |
| CorefBERT (Ye et al. 2020) | 75.96 | 72.30 |
| DREAM (Zhong et al. 2020) | 76.85 | 70.60 |

All models use the same evidence retrieval; currently second best

Testing robustness

Use paired symmetric counterfactual data from Schuster et al. (2019) to validate dependence on claim's text alone



Results on symmetric FEVER

| Model | Trained on FEVER | After Fine-Tuning | After Fine-Tuning tested on FEVER |
|----------------------------|------------------|-------------------|-----------------------------------|
| ProofVer | 81.70 | 85.88 | 86.41 |
| KGAT (Liu et al. 2020) | 65.73 | 84.94 | 76.67 |
| CorefBERT (Ye et al. 2020) | 68.49 | 85.45 | 78.79 |

- Robustness of ProofVer also when faced with additional evidence
- Improvements on FEVER2 dataset

Claim ~~Verification~~ Correction

Refuted

Make meaning altering changes to claims so that they are better supported by evidence

“Bullitt was directed by D’Antoni”



Bullitt

From Wikipedia, the free encyclopedia

For other uses, see [Bullitt \(disambiguation\)](#).

Bullitt is a 1968 American neo-noir action thriller film^[4] directed by Peter Yates and produced by Philip D'Antoni. The picture stars Steve McQueen, Robert Vaughn, and Jacqueline Bisset.^[5] The screenplay by Alan R. Trustman and Harry Kleiner was based on the 1963 novel *Mute Witness*,^{[6][7][8][9]} by Robert L. Fish, writing under the pseudonym Robert L. Pike.^{[10][11]} Lalo Schifrin wrote the original jazz-inspired score.



Supported

“Bullitt was produced by D’Antoni”

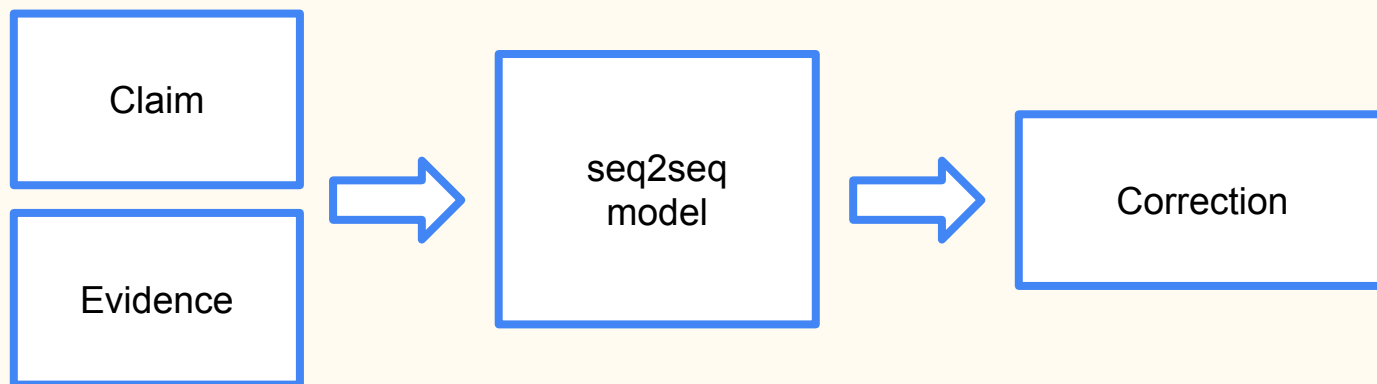
Sub-tasks:

- Find evidence
- Identify falsehoods
- Incorporate information from the evidence into the claim to correct them

Requirements for generating corrections

- 1. Intelligent:** Is the generated correction grammatical and can the meaning be understood? Considers the correction in isolation
- 2. Supported by Evidence:** Same definition as in claim verification. Considers the relation between correction and evidence
- 3. Correcting the Error:** Is the correction addressing an error in the claim? Considers the relation between correction and claim

Supervised encoder-decoder model?



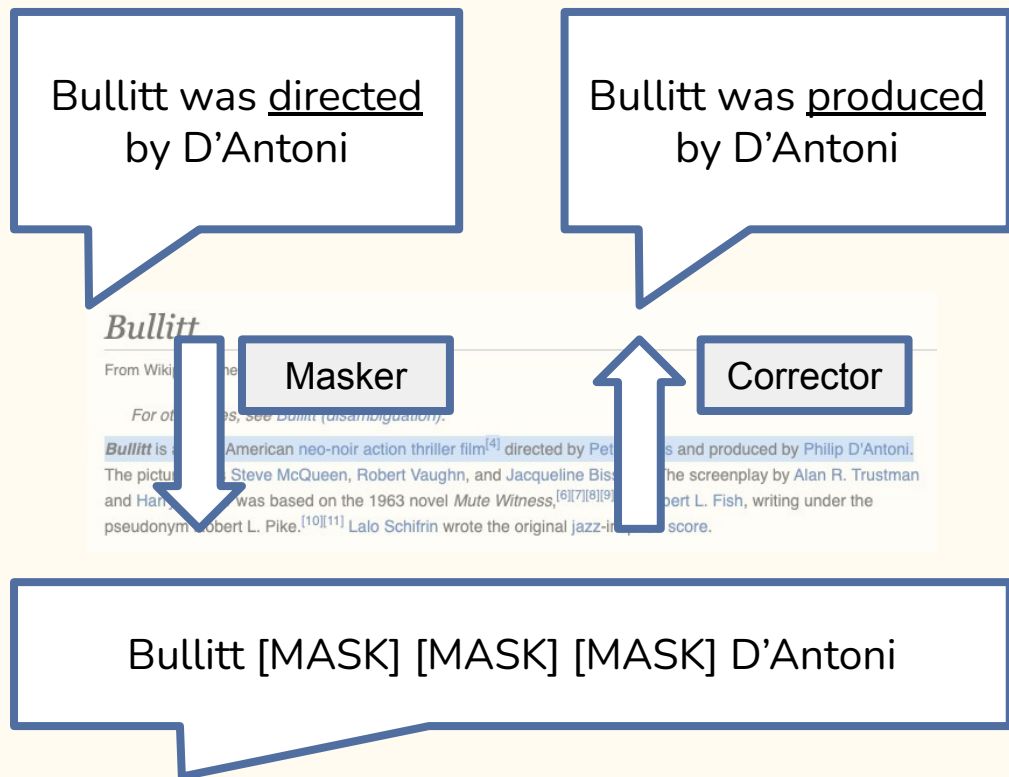
Yes, but we don't have the data!

Datasets typically have claim and evidence, not corrections.

Generating corrections

Step 1: Mask out tokens from claim using retrieved evidence.

Step 2: Use evidence to rewrite the masked sentence into a correction



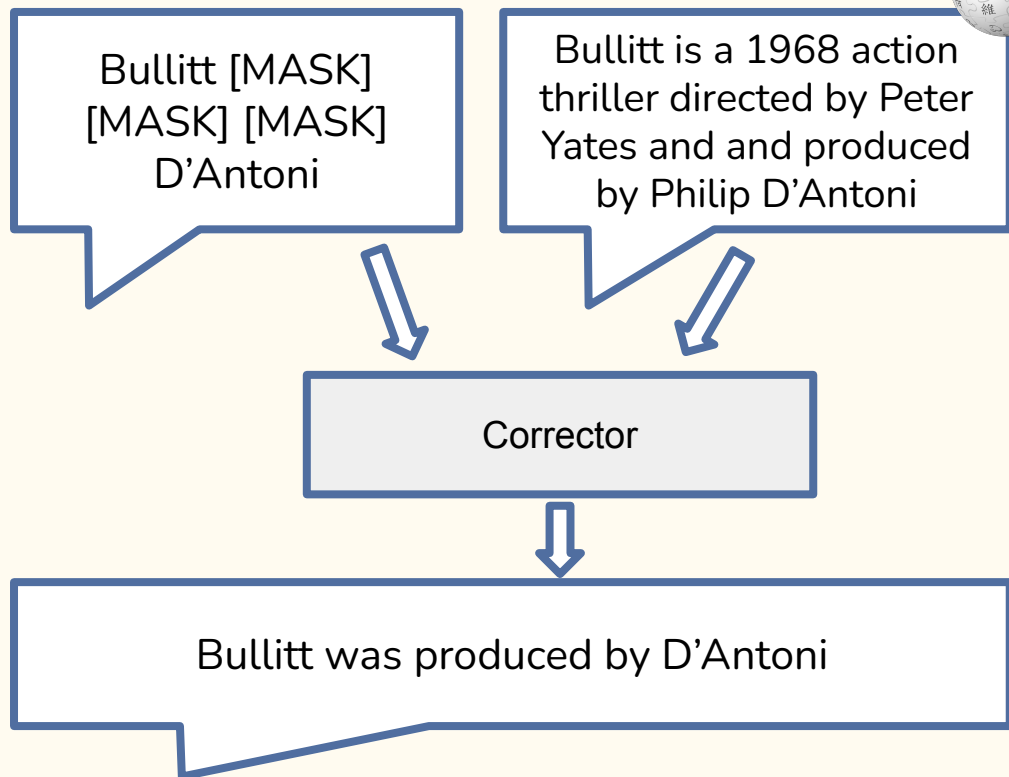
Incorporating evidence into masked claims

Seq2seq model

Trained to recover missing tokens, conditioned on evidence

Training: reproduce the (unmasked) claim, only possible if claim was supported by the evidence, otherwise partly

Masker: random in training, heuristic in testing



Human Evaluation on FEVER

| Model | Evidence | Intelligible | Supported | Corrected |
|---|-----------|--------------|-----------|-----------|
| Supervised (oracle) | Retrieved | 98% | 65% | 49% |
| Ours | Retrieved | 89% | 58% | 40% |
| [Shah et al 2020] (tokens from NLI interpretation for masking) | Gold | 32% | 11% | 5% |
| BERT (no evidence) | N/A | 30% | 20% | 15% |

For more details see our paper (*Thorne and Vlachos, 2021*)

Beyond text-based verification

FEVEROUS: Fact Extraction and VERification Over Unstructured and Structured information (*Aly et al., NeurIPS datasets 2021*)

- 87K claims with evidence (text and tables)
- All of Wikipedia, not just the introductory sections
- Evidence+veracity needs to be correct
- Not Enough Information must be accompanied by evidence
- Shared task outcomes:
 - 12 teams, accuracies from 2% to 27% (our baseline: 18%)
 - Results are on our website: <http://fever.ai/task.html>

FEVEROUS examples

Claim: In the 2018 Naples general election, Roberto Fico, an Italian politician and member of the Five Star Movement, received 57,119 votes with 57.6 percent of the total votes.

Evidence:

Page: wiki/Roberto_Fico
e₁(Electoral history):

2018 general election: Naples -Fuorigrotta

| Candidate | Party | Votes |
|-----------------|--------------|--------|
| Roberto Fico | Five Star | 61,819 |
| Marta Schifone | Centre-right | 21,651 |
| Daniela Iaconis | Centre-left | 15,779 |

Verdict: Refuted

Claim: Red Sundown screenplay was written by Martin Berkeley; based on a story by Lewis B. Patten, who often published under the names Lewis Ford, Lee Leighton and Joseph Wayne.

Evidence:

Page: wiki/Red_Sundown
e₁(Introduction):

| Red Sundown | |
|---------------|-----------------|
| Directed by | Jack Arnold |
| Produced by | Albert Zugsmith |
| Screenplay by | Martin Berkeley |
| Based on | Lewis B. Patten |
| ... | |

Page: wiki/Lewis_B._Patten
e₂(Introduction): He often published under the names Lewis Ford, Lee Leighton and Joseph Wayne.

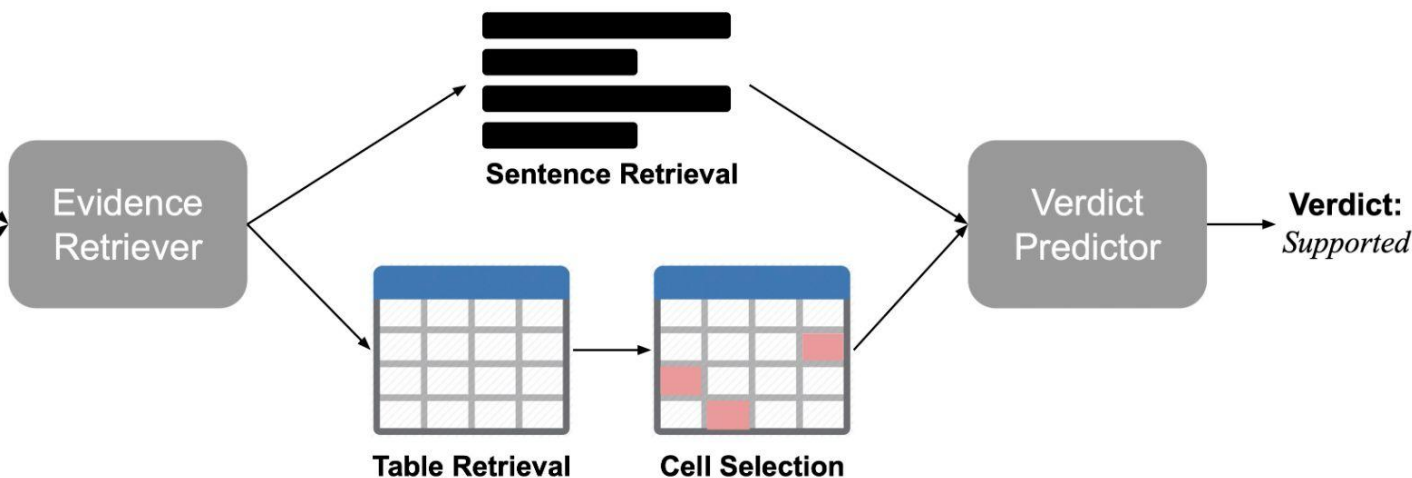
Verdict: Supported

FEVEROUS baseline

Claim: *Red Sundown*
screenplay was written by
Martin Berkeley; based
on a story by Lewis...



Retrieval Corpus

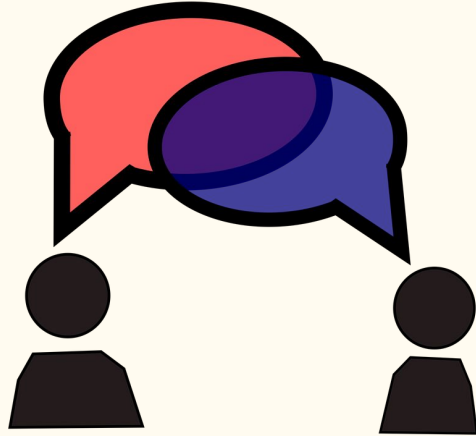


- Wikipedia pages retrieved by entity matching and TF-IDF
- Sentences and tables are retrieved with TF-IDF
- Cells are picked with a fine-tuned RoBERTa sequence labeler
- Veracity is predicted with fine-tuned RoBERTa on concatenated claim and all evidence

Next steps

- Neurosymbolic evidence retrieval
- Neurosymbolic inference for tables and text
- Fact-checking real-world claims
- Multilingual fact-checking

Misinformation and fact-checking are social



- Online social media helps us see outside the bubble but also increase polarisation (Bail et al., 2018)
- What correlates with **more constructive conversations**?
- How can we **intervene** to make them happen?

Fact checking as a conversation



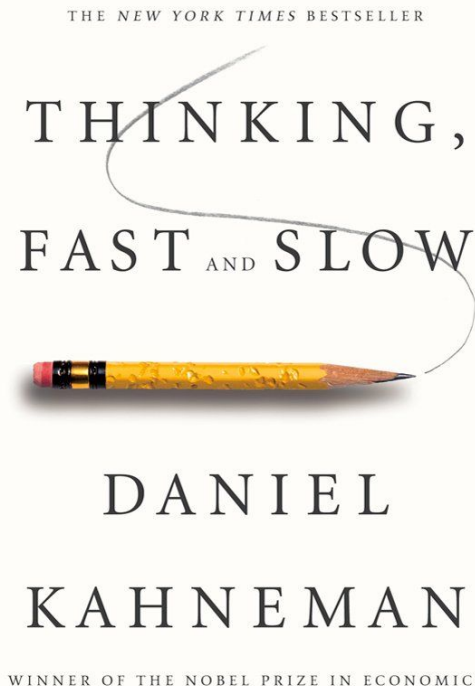
WikiTRIBUNE



- Wikipedia: most successful large-scale online conversation
- Success not straightforward to replicate
- How can we make it happen again?

Let's take a look at reasoning

- Dual system
 - System 1: Fast, biased
 - System 2: Slow, rational
- Various cognitive biases:
 - Recency bias
 - Confirmation bias
 - etc.



"[A] masterpiece . . . This is one of the greatest and most engaging collections of insights into the human mind I have read." —WILLIAM EASTERLY, *Financial Times*

Wason (1968) selection task

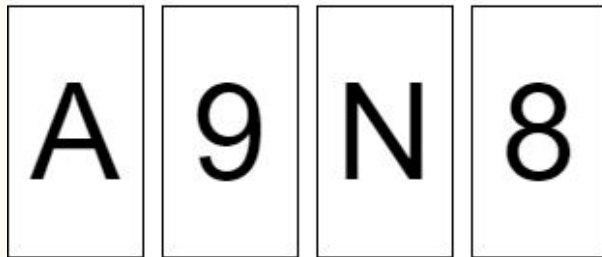
What do you think?

Individuals' success rate: 10-20%

Small groups success rate?

80%! What makes groups work?

Each of the 4 cards bellow has letter on one side and a number on the other. Which card(s) do you need to turn to test the rule:
All cards with vowels on one side, have an even number on the other.

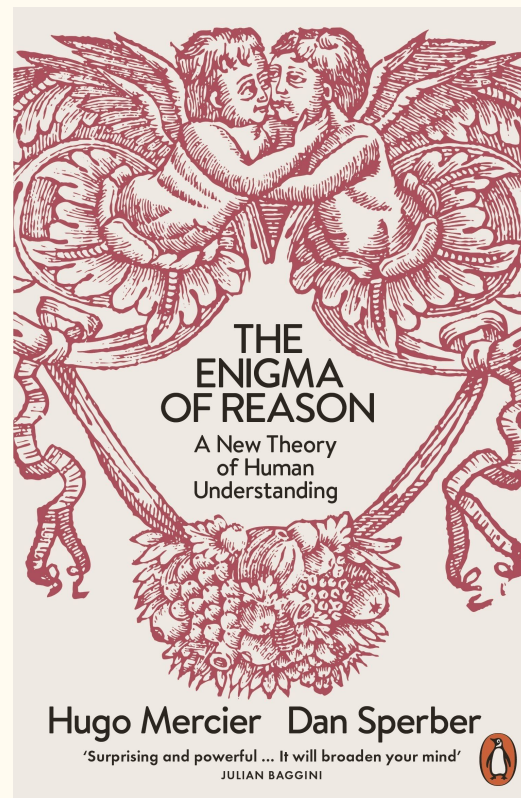


With a little help from my friends

Reasoning has evolved in the context of communication, not in isolation:

- arguments are made to help us justify ourselves and convince each other
- we are bad judges of our own arguments but good for the others
- Scientists are no different!

Can we help groups work better?



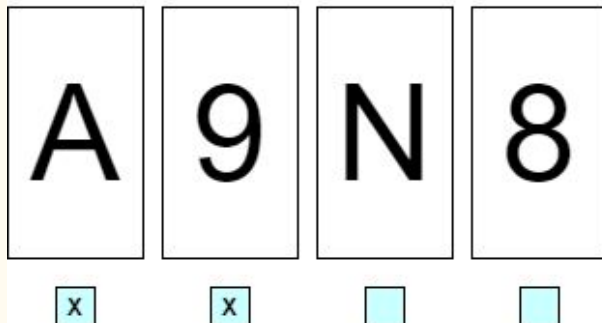
Deliberation Enhancing Bots (DEliBots)

- Develop conversational agents that make conversations better!
- A different kind of dialogue agent:
 - Unlike chatbots, they help users accomplish a task
 - Unlike task-oriented bots (e.g. restaurant booking), they don't know or give the answer

Let's look at some data

Each of the 4 cards bellow has letter on one side and a number on the other. Which card(s) do you need to turn to test the rule:

All cards with vowels on one side, have an even number on the other.



Beaver: What do you think?

Cat: I think A and 8

Duck: I thought A and 8 too, but we may be wrong

Cat: @Duck, well we need A for sure

Beaver: What if we don't turn 8 at all?

Duck: Yes! We don't care what is behind the even numbers

Cat: This may be right, but we may need to check the odds

Duck: So, A and 9?

Beaver: Yes

Data collection

Initial experiments with local volunteers, once stable used MTurk

- 500 groups, 2-5 persons (avg 3.16) (smaller group, fewer ideas)
- each group member submits responses at onboarding
- the group deliberates and members submit again
- no need for the group consensus but bonus for correct response

Onboarding success rate: 11%

Success rate after deliberation: 33%

In 43.8% of the groups with the correct solution, no participant had chosen it initially!

Annotating deliberation

How do we improve deliberation?

Ask questions/probes for:

- moderation
- solutions
- reasons

Hypothesis: **probing for reasoning** makes a difference

Beaver: What do you think?

moderation

Cat: I think A and 8

Duck: I thought A and 8 too, but we may be wrong

Cat: @Duck, well we need A for sure

Beaver: What if we don't turn 8 at all?

reason

Duck: Yes! We don't care what is behind the even numbers

Cat: This may be right, but we may need to check the odds

Duck: So, A and 9?

solution

Beaver: Yes

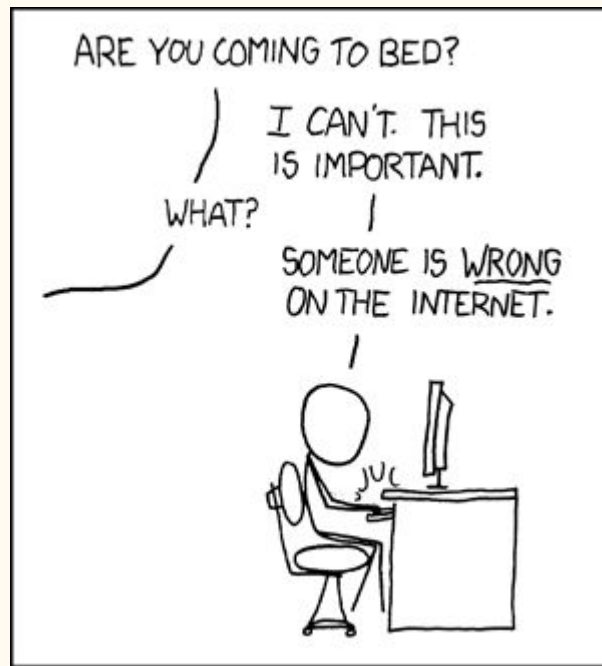
Annotation findings

- Three annotators (two NLPers, one psychologist)
- Inter-annotator agreement (Kappa): 75% on annotating probes, 71% on determining probing type
- Key correlations:
 - **Probing for reasoning** correlates positively but weakly; how we probe (choice of language) is likely to matter.
 - **Conversation length** correlates positively but weakly
 - Strongest correlation is with **group consensus**; agrees with previous work that small group discussion is better than wisdom of crowds (Navajas et al., 2018)

Next steps

- Identify the utterances that result in improved solutions
- Build DeliBots!
- Evaluation with humans in the loop
- Data and paper here: <https://www.delibot.xyz/delidata/>

Getting real: Wikipedia vs much of the web

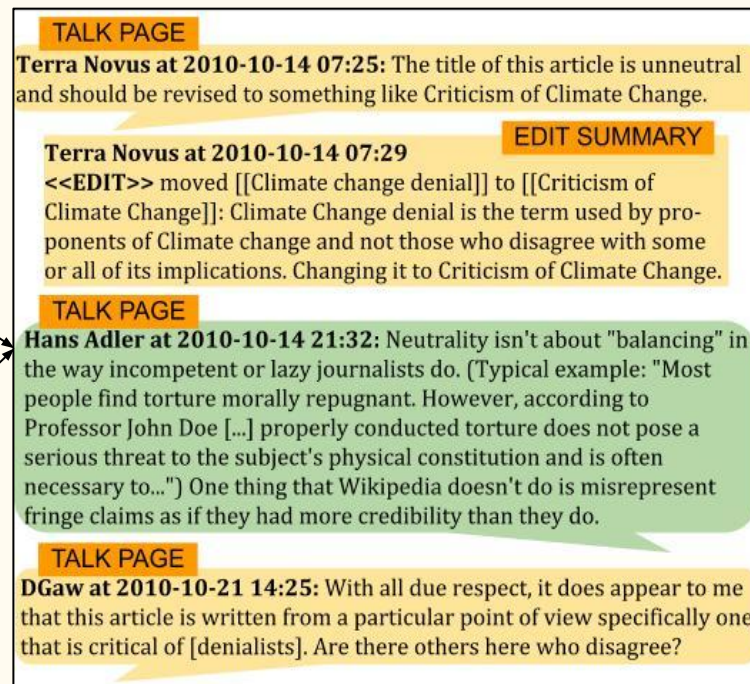
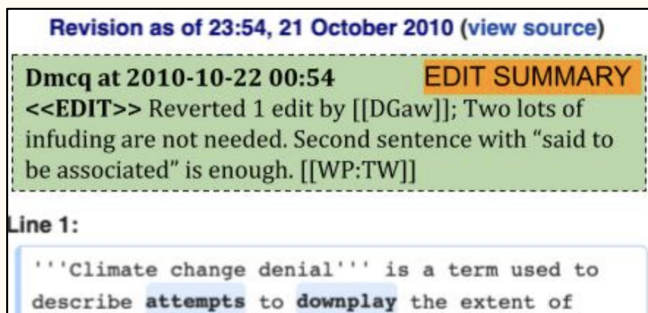


WikiDisputes (De Kock and Vlachos, EACL2021)

- A corpus of 7 425 disagreements on Wikipedia Talk pages




WikiConv: A Corpus of the Complete Conversational History of a Large Online Collaborative Community Hua et al., 2018



Predicting escalation

Welcome to the dispute resolution noticeboard (DRN)

This is an *informal* place to resolve small **content disputes** as part of [dispute resolution](#). It may also be used as a tool to direct certain discussions to more appropriate forums, such as [requests for comment](#), or other noticeboards. You can ask a question on the [talk page](#). This is an early stop for most disputes on Wikipedia. You are [not required to participate](#), however, the case filer must participate in all aspects of the dispute or the matter will be considered failed. Any editor may volunteer! Click this button  to add your name! You don't need to volunteer to help. Please feel free to comment below on any case. **Be civil and remember; Maintain Wikipedia policy: it is usually a misuse of a talk page to continue to argue any point that has not met policy requirements.** "Editors must take particular care adding *information about living persons* to any Wikipedia page. This may also apply to some [groups](#).

Noticeboards should not be a substitute for talk pages. Editors are expected to have had extensive discussion on a talk page (not just through edit summaries) to work out the issues before coming to DRN.

Do you need assistance?

Would you like to help?

Request dispute resolution

Become a volunteer

Shortcuts

WP:DRN

WP:DR/N

- + Escalation labels:
- 201 Escalated
 - 7224 Not escalated*
- *sub-sampled to correct for length imbalance

Predicting escalation

Feature-based models

- **Toxicity:** Wulczyn et al. (2017)
- **Sentiment:** Liu et al. (2005)
- **Politeness:** Zhang et al. (2018)
- **Collaboration:** Niculescu and Danescu- Niculescu-Mizil (2016)

+ **Gradients** to consider how feature values change throughout conversation

| Model | PR-AUC |
|------------------------------|--------|
| Baselines | |
| Random | 0.121 |
| Bag-of-words | 0.213 |
| Feature-based models | |
| Toxicity | 0.140 |
| Sentiment | 0.150 |
| Politeness | 0.232 |
| + <i>gradients</i> | 0.275 |
| Collaboration | 0.261 |
| + <i>gradients</i> | 0.269 |
| Politeness and collaboration | 0.255 |
| + <i>gradients</i> | 0.281 |

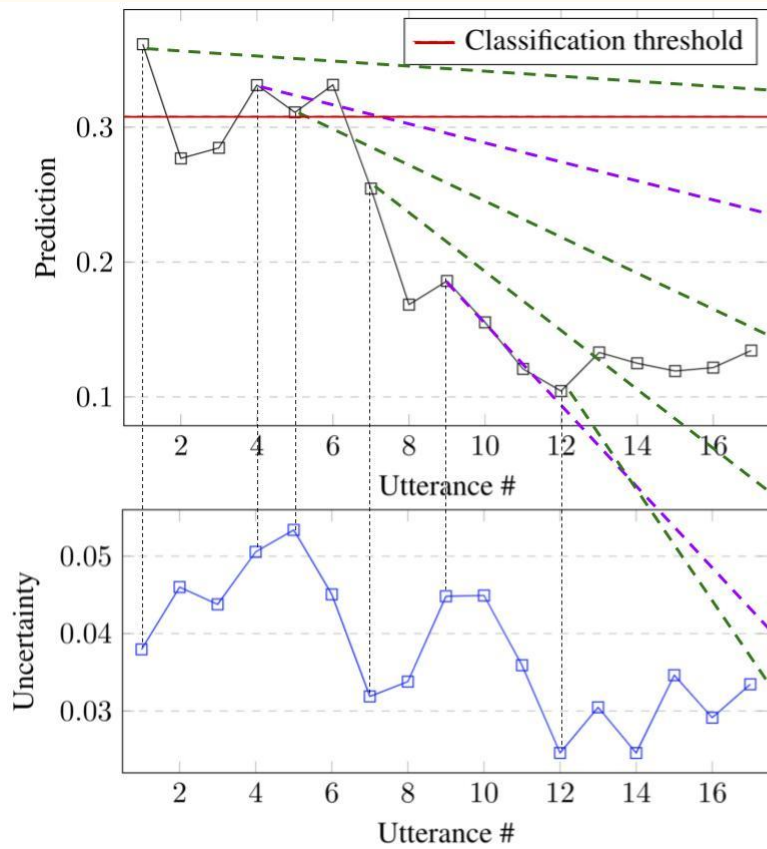
Predicting escalation

Neural networks

- **GloVe embeddings:** Pennington et al. (2014)
 - **LSTM-based:** Hochreiter and Schmidhuber (1997)
 - **Hierarchical attention network:** Yang et al. (2016)
- Representing **structure** helps
- **Edit summaries** help

| Model | PR-AUC |
|------------------------------|--------------|
| Baselines | |
| Random | 0.121 |
| Bag-of-words | 0.213 |
| Feature-based models | |
| Toxicity | 0.140 |
| Sentiment | 0.150 |
| Politeness | 0.232 |
| + <i>gradients</i> | 0.275 |
| Collaboration | 0.261 |
| + <i>gradients</i> | 0.269 |
| Politeness and collaboration | 0.255 |
| + <i>gradients</i> | 0.281 |
| Neural models | |
| Averaged embeddings | 0.243 |
| LSTM | 0.263 |
| HAN | 0.373 |
| + <i>edit summaries</i> | 0.400 |

A cherry picked example from our model



Ggugvunt at 2006-08-25 15:09: The rest of this article is quite well written and reasonably NPOV, but the caption to the Canada picture reads like polemic to me.

Frogsprog at 2006-08-25 15:12: I would start by saying I am "very" annoyed by american editors refusing to accept information that makes their country look unpopular.

Ggugvunt at 2006-08-25 15:20: First of all, you need to calm down a bit. Secondly, if you think the picture is so violent and strong, let it speak for itself.

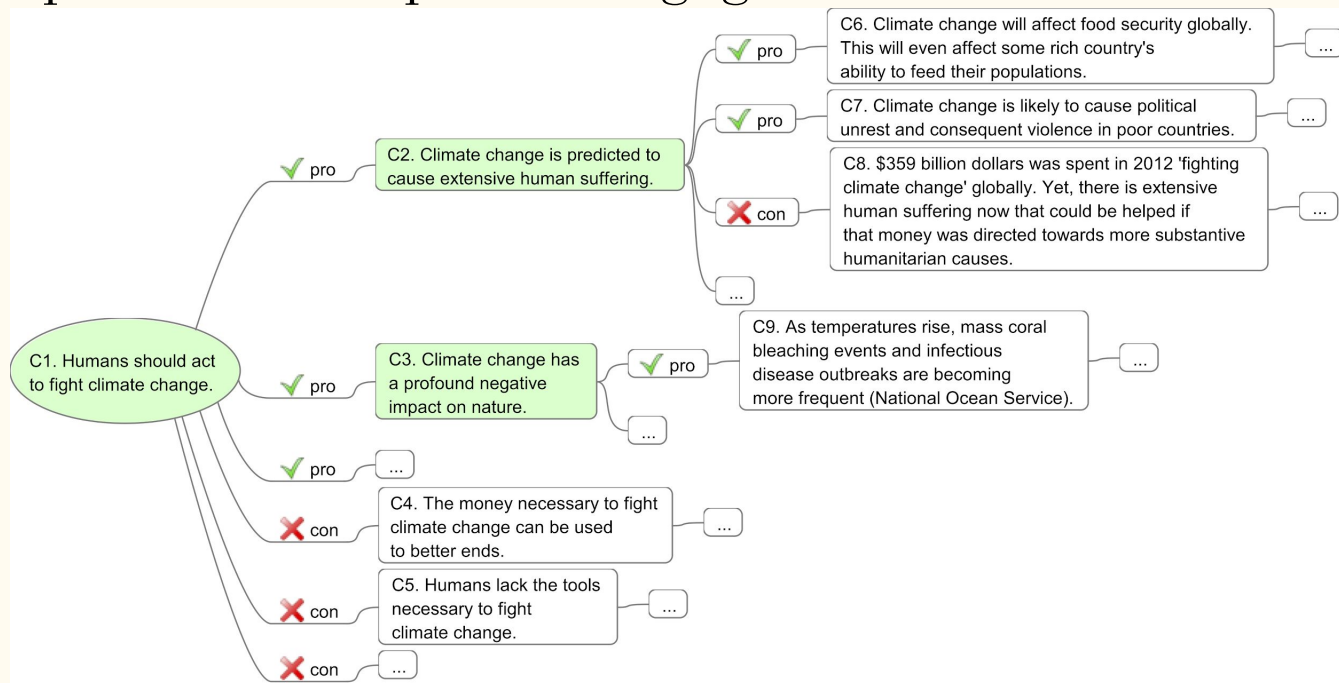
Ggugvunt at 2006-08-25 15:33: Listen - my only issue is this: the caption sounded out of place with the reasonable tone of the rest of the article. Don't jump to conclusions - please. Can I revert it back to the non-adverbs version and remove the NPOV marker now?

Frogsprog at 2006-08-25 18:11: OK, a compromise, I will re-add the adverb "violently" but leave out strong

Ggugvunt at 2006-08-25 18:39: Much better! Thank you!

How do we encourage them? Opening Minds Up

- Joint project with Open University, Sheffield and Toshiba
- Develop bots that help users engage with the “other side”





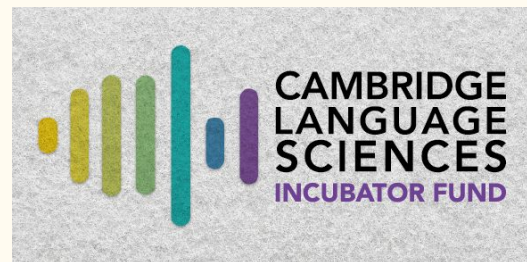
Thanks to the funding agencies



European Research Council



monitio



Questions?

andreas.vlachos@cst.cam.ac.uk