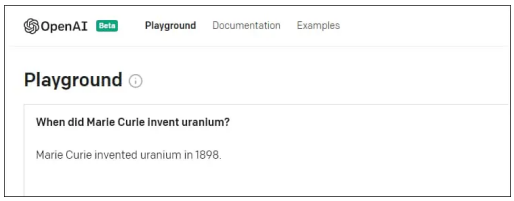
# Getting NLP to Challenge Misinformed Questions

Originally published at https://www.unite.ai/getting-nlp-to-challenge-misinformed-questions/

Some questions are unanswerable because they contain incorrect information – presuppositions that the person hearing the question must filter and renounce. This assumes, of course, that the listener has  enough correct information to challenge the question, rather than using the question itself as a source of (wrong) information.

It's a challenge for Natural Language Processing (NLP) systems such as GPT-3, which have a tendency to 'hallucinate' information in order to maintain dialogue.

Currently, asking GPT-3 'When did Marie Curie invent Uranium?' will likely get you the answer 'Marie Curie invented Uranium in 1898'.

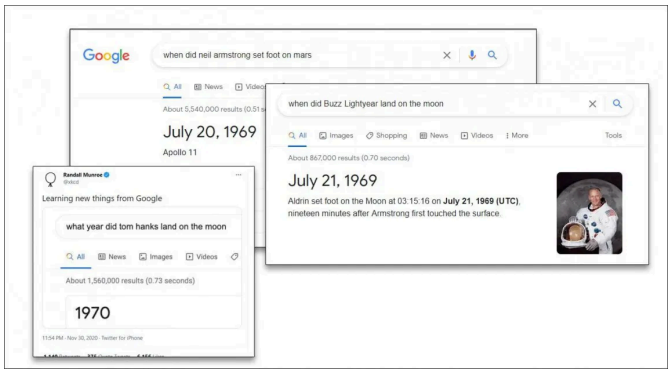



Source: https://beta.openai.com/playground (Da Vinci instruct beta).

In fact, Uranium was discovered in 1789 by German chemist Martin Heinrich Klaproth, whilst the Curies' 1898 revelation was the isolation of radium.

The problem of NLP systems ignoring incorrect presuppositions has come into focus in a number of publicity spouts this year, including the way that Google's AI-assisted search results will ignore incorrect information in the question 'When did Neil Armstrong set foot on Mars?' – an error which still shows at the time of writing this article, and equally applies to *Toy Story*'s Buzz Lightyear, who apparently landed on the Moon on July 21st 1969.

Tom Hanks, another *Toy Story* alumnus, is also credited by Google with landing on the Moon in 1970, in spite of the fact that his *Apollo 13* character, astronaut Jim Lovell, is most famous for *not* having achieved this.

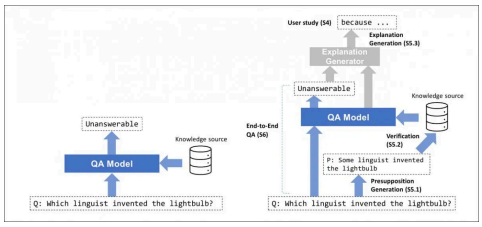


## Addressing Presupposition Issues in NLP Exchanges

Now Google Research, together with researchers from John Hopkins University and Brown University, is investigating new machine learning methods by which NLP systems can eventually be made to challenge factually incorrect questions in the same way that it's essential for human teachers to do during conversations with pupils.

The recent paper *Which Linguist Invented the Lightbulb? Presupposition Verification for Question-Answering* outlines a concerted effort to develop a novel system to identify presuppositions and to consider their veracity before continuing the exchange

The new algorithm effectively preprocesses questions before returning to the conversation, breaking down the 'authentication' of the question in a three-part process.




*Does not compute! On the left, the 'roadblock' that occurs even when an advanced NLP system has been able to identify that the question does not make sense. On the right, a breakdown of a proposed algorithm that attempts to rectify the source error.* Source: https://arxiv.org/pdf/2101.00391.pdf

Though it seems a simple verification routine that should have been built into knowledge systems from the outset, most NLP-based training routines learn information with an undue level of trust for the source data, including discourse (such as fake news) that may have been published on previously 'trusted' channels.

Therefore a key issue is to identify by consensus a reliable source of facts in a climate where the proliferation of incorrect 'news' through social media would, by default, grant it authority under the logic of machine learning generalization. The latter has tended to use quantity or repetition of data as a proxy for accuracy, at least until the phenomena of fake news became a critical area of interest in the field in recent years.

## Determining the Best Approach to Unanswerable Questions

To determine a suitable approach for resolving a question that contains misinformation, the researchers ran 100 such inquiries through four different Q&A models, and asked human subjects to pick the best or least problematic solution that the models generated.

The four possible architectural outcomes to the 'bad' question were: *'Unanswerable'* – where a closed book Q&A system effectively shuts down the inquiry without further elaboration; *'Presupposition failure-based explanation'* – where the system fails to verify the incorrect supposition, effectively an 'unanswerable' response, with an added explanation; *'Extractive explanation'* – where the system retrieves a topically related Wikipedia quote and appends it to the prefatory 'This question is unanswerable because…'; and *'Open domain rewrite'* – where a competitive system seeks additional sources from Wikipedia.

| Question: *where can i buy a japanese dwarf flying squirrel* | |
|---|---|
| Simple unanswerable | *This question is unanswerable.* |
| Presupposition failure-based | *This question is unanswerable because we could not verify that you can buy a Japanese Dwarf Flying Squirrel anywhere.* |
| Extractive explanation | *This question is unanswerable because it grows to a length of 20 cm (8 in) and has a membrane connecting its wrists and ankles which enables it to glide from tree to tree.* |
| DPR rewrite | *After it was returned for the second time, the original owner, referring to it as "the prodigal gnome", said she had decided to keep it and would not sell it on Ebay again.* |

*This example of four possible answers to an apparently 'unanswerable' question illustrates the complexity of attempting a competitive domain-based solution to the problem.*

Over the course of the tests, the five participants (recruited on an internal Google crowdsourcing platform) preferred the presupposition-based answers, which led the researchers to develop a new framework to decompose and verify questions.

In the new system, linguistic triggers are obtained from the question by a rule-based generator that deconstructs the sentence into putative statements of fact. If multiple suppositions are derived from the question, each one is investigated, and will contribute to the final response if they address mistaken presuppositions from the original question.

## Datasets

The presuppositions generated in the initial stage were manually amended to create a verification dataset with 'gold' presuppositions. Any presuppositions that emerged from the branching of the inquiry, but which were not present in the original questions, were removed.

Two of the paper's authors then manually annotated 462 presuppositions in terms of *yes/no* verifiability, based on a relevant Wikipedia page associated with each question. Cases of disagreement were resolved in post-facto discussion before being committed to the dataset.

The researchers used zero-shot NLI, a premise/hypothesis classification task which required the deconstruction of Wikipedia articles related to the questions. Since this process results in many more pairs than the question may entail or the model support, the filtered results were then aggregated and labeled.

## Results and Response Formulation

The most effective results were obtained by the most labor-intensive solution: a finer-tuned, rule-based/NLI hybrid generated from ALBERT QNLI with Wiki sentences and presuppositions.

| Model | Macro F1 | Acc. |
|---|---|---|
| Majority class | 0.44 | 0.78 |
| Zero-shot NLI (ALBERT MNLI + Wiki sentences) | 0.50 | 0.51 |
| Zero-shot NLI (ALBERT QNLI + Wiki sentences) | 0.55 | 0.73 |
| Zero-shot FEVER (KGAT + Wiki sentences) | 0.54 | 0.66 |
| Finer-tuned NLI (ALBERT QNLI + Wiki sentences) | 0.58 | 0.76 |
| Rule-based/NLI hybrid (ALBERT QNLI + Wiki presuppositions) | 0.58 | 0.71 |
| Rule-based/NLI hybrid (ALBERT QNLI + Wiki sentences + Wiki presuppositions) | 0.59 | 0.77 |
| Finer-tuned, rule-based/NLI hybrid (ALBERT QNLI + Wiki sentences + Wiki presuppositions) | **0.60** | **0.79** |

*The performance of the verification models, where 'Wiki sentences' uses sentences obtained from question-related Wikipedia articles, and 'Wiki presuppositions' a presuppositions from those sentences.*

Using this formulation, the researchers developed a template system where a negating fact from Wikipedia was appended to 'This question is unanswerable because…' and similar phrases. Though it's not an ideal solution, the authors suggest that responses based on unverifiability are likely to reduce the incidence of false negatives.

The system was ultimately implemented in an Extended Transformer Construction (ETC) model.

## Implications

Depending on its ultimate performance in the real world, it could be argued that this entire approach may lead to the mere substitution of 'unverifiable' for 'unanswerable', in cases where the supporting research system cannot evaluate a useful correction for a question's mistaken presupposition. Effectively, it seems to be laying the infrastructure for future and better verification systems.

The researchers already concede that the expense of token-based API requests are a limiting factor when formulating the longer replies that this system will generate, and it has to be assumed that the additional overhead of 'live' research into a question seems likely to add latency even to large scale systems such as GPT-3, since the responsiveness of such systems has to date depended on the generalized incorporation of knowledge at training time, rather than extensive, network-based verification routines.

Additionally, the researchers note that the system currently has limitations related to parsing semantic aspects of the text:

> *For example,* who does pip believe is estella's mother *has an embedded possessive under a nonfactive verb* believe*, but our generator would nevertheless generate '*estella' has 'mother'.

Nonetheless, the team envisages new and more flexible question-answering systems that will be developed on the back of this research:

> *In the future, we plan to build on this work by proposing QA systems that are more robust and cooperative. For instance, different types of presupposition failures could be addressed by more fluid answer strategies—e.g., violation of uniqueness presuppositions may be better handled by providing all possible answers, rather than stating that the uniqueness presupposition was violated.*

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai