How Do We Answer Complex Question: Discourse Structure of Long-form Answers

Fangyuan Xu, Junyi Jessy Li, Eunsol Choi
Going Beyond Span-based Answer

**Question:** Who lives in the imperial palace in Tokyo?

**Answer:** the Imperial Family

**Question:** Why does salt bring out the flavor in most foods?

**Answer:** Salt does a couple of things that add to the flavor of foods. First off, it makes things salty. That sounds simplistic, but salt is one of the 5 basic tastes, so it tends to taste good simply all by itself. Secondly, salt will lessen sensitivity to other tastes, especially bitter. This is somewhat true of all the tastes, but adding salt will balance out a dish and not let one taste overpower the others. Thirdly, it's been shown to increase that aromatic effects of many types of food. A good deal of your "taste" of a food actually comes from the smell of that food (which is why things tend to taste so bland when you nose is congested, like when you have the flu).

Natural Question (NQ), from Kwiatkowski et al. 2019

Explain Like I’m Five (ELI5), from Fan et al. 2019
Why Long-form Question Answering?

✓ Can handle broader set of questions
Why Long-form Question Answering?

✓ Can handle broader set of questions

✓ Can provide comprehensive answers to factoid questions

+ Address ambiguity [Min et al. 2020, Zhang et al. 2021]

+ Provide source of the knowledge, acknowledge limitations [Fan et al. 2019]
Challenges in Long-form QA

- Evaluation for lengthy, complicated answers is challenging [Krishna et al. 2021]
Challenges in Long-form QA

- Evaluation for lengthy, complicated answers is challenging [Krishna et al. 2021]
- ROUGE is not meaningful ❌
Challenges in Long-form QA

- Evaluation for lengthy, complicated answers is challenging [Krishna et al. 2021]
  - ROUGE is not meaningful ❌
  - Human evaluations are non-trivial 😞
    - Unfamiliar with the question topics
    - High cognitive load to evaluate correctness and fluency of long answer
Goal: Understanding long-form answers

- How is long-form answer structured?
- What purpose does each sentence serve to answer the question?
Goal: Understanding long-form answers

- How is long form answer structured?
- What purpose does each sentence serve to answer the question?
  ✷ Longer term goals:
    ✷ Design fine-grained human and automatic evaluation protocols
    ✷ Structure the answer generation to include appropriate information
What purpose does each sentence serve to answer the question?

- background
- method
- findings

Functional Roles

Kircz, 1991; Liddy, 1991; Mizuta et al, 2006
Why does salt bring out the flavor in most foods?

Salt does a couple of things that add to the flavor of foods. First off, it makes things salty. That sounds simplistic, but salt is one of the 5 basic tastes, so it tends to taste good simply all by itself. Secondly, salt will lessen sensitivity to other tastes, especially bitter. This is somewhat true of all the tastes, but adding salt will balance out a dish and not let one taste overpower the others. Thirdly, it's been shown to increase that aromatic effects of many types of food. A good deal of your "taste" of a food actually comes from the smell of that food (which is why things tend to taste so bland when you nose is congested, like when you have the flu).

Kircz, 1991; Liddy, 1991; Mizuta et al, 2006
Question: Why does salt bring out the flavor in most foods?

Answer: Salt does a couple of things that add to the flavor of foods. First off, it makes things salty. That sounds simplistic, but salt is one of the 5 basic tastes, so it tends to taste good simply all by itself. Secondly, salt will lessen sensitivity to other tastes, especially bitter. This is somewhat true of all the tastes, but adding salt will balance out a dish and not let one taste overpower the others. Thirdly, it’s been shown to increase that aromatic effects of many types of food. A good deal of your "taste" of a food actually comes from the smell of that food (which is why things tend to taste so bland when you nose is congested, like when you have the flu).

This work: developing an ontology of six functional roles for long-form answers.
**Question:** Why does salt bring out the flavor in most foods?

**Answer:**
Salt does a couple of things that add to the flavor of foods. First off, it makes things salty. That sounds simplistic, but salt is one of the 5 basic tastes, so it tends to taste good simply all by itself. Secondly, salt will lessen sensitivity to other tastes, especially bitter. This is somewhat true of all the tastes, but adding salt will balance out a dish and not let one taste overpower the others. Thirdly, it's been shown to increase that aromatic effects of many types of food. A good deal of your "taste" of a food actually comes from the smell of that food (which is why things tend to taste so bland when you nose is congested, like when you have the flu).
Question: Why does salt bring out the flavor in most foods?

Answer:
Salt does a couple of things that add to the flavor of foods.
First off, it makes things salty. That sounds simplistic, but salt is one of the 5 basic tastes, so it tends to taste good simply all by itself.
Secondly, salt will lessen sensitivity to other tastes, especially bitter. This is somewhat true of all the tastes, but adding salt will balance out a dish and not let one taste overpower the others.
Thirdly, it’s been shown to increase that aromatic effects of many types of food.
A good deal of your “taste” of a food actually comes from the smell of that food (which is why things tend to taste so bland when you nose is congested, like when you have the flu).
Question: Why does salt bring out the flavor in most foods?

Answer: Salt does a couple of things that add to the flavor of foods. First off, it makes things salty. That sounds simplistic, but salt is one of the 5 basic tastes, so it tends to taste good simply all by itself. Secondly, salt will lessen sensitivity to other tastes, especially bitter. This is somewhat true of all the tastes, but adding salt will balance out a dish and not let one taste overpower the others. Thirdly, it's been shown to increase that aromatic effects of many types of food. A good deal of your "taste" of a food actually comes from the smell of that food (which is why things tend to taste so bland when you nose is congested, like when you have the flu).

Organizational sentence

which informs the reader how the answer will be structured.

- There are a few reasons. [...]  
- So why does it happen more often when you shower? It's largely because [...]


Question: Why does salt bring out the flavor in most foods?

Answer:
Salt does a couple of things that add to the flavor of foods. First off, it makes things salty. That sounds simplistic, but salt is one of the 5 basic tastes, so it tends to taste good simply all by itself.
Secondly, salt will lessen sensitivity to other tastes, especially bitter. This is somewhat true of all the tastes, but adding salt will balance out a dish and not let one taste overpower the others.
Thirdly, it's been shown to increase that aromatic effects of many types of food.
A good deal of your "taste" of a food actually comes from the smell of that food (which is why things tend to taste so bland when you nose is congested, like when you have the flu).
Question: Why does salt bring out the flavor in most foods?

Answer:
Salt does a couple of things that add to the flavor of foods. First off, it makes things salty. That sounds simplistic, but salt is one of the 5 basic tastes, so it tends to taste good simply all by itself. Secondly, salt will lessen sensitivity to other tastes, especially bitter. This is somewhat true of all the tastes, but adding salt will balance out a dish and not let one taste overpower the others. Thirdly, it’s been shown to increase that aromatic effects of many types of food. A good deal of your “taste” of a food actually comes from the smell of that food (which is why things tend to taste so bland when you nose is congested, like when you have the flu).
**Question:** Why does salt bring out the flavor in most foods?

**Answer:**
Salt does a couple of things that add to the flavor of foods. **First off, it makes things salty.** That sounds simplistic, but salt is one of the 5 basic tastes, so it tends to taste good simply all by itself. **Secondly, salt will lessen sensitivity to other tastes, especially bitter.** This is somewhat true of all the tastes, but adding salt will balance out a dish and not let one taste overpower the others. **Thirdly, it's been shown to increase that aromatic effects of many types of food.** A good deal of your "taste" of a food actually comes from the smell of that food (which is why things tend to taste so bland when you nose is congested, like when you have the flu).
Question: Why does salt bring out the flavor in most foods?

Answer:
Salt does a couple of things that add to the flavor of foods. First off, it makes things salty. That sounds simplistic, but salt is one of the 5 basic tastes, so it tends to taste good simply all by itself. Secondly, salt will lessen sensitivity to other tastes, especially bitter. This is somewhat true of all the tastes, but adding salt will balance out a dish and not let one taste overpower the others. Thirdly, it's been shown to increase that aromatic effects of many types of food. A good deal of your "taste" of a food actually comes from the smell of that food (which is why things tend to taste so bland when you nose is congested, like when you have the flu).
Question: why does sugar turn brown when melted?

Answer:
Sugar is a molecule made of a particular arrangement of carbon, oxygen, and hydrogen atoms. When you heat up the sugar, this adds energy to the system that can be used to fuel chemical reactions that rearrange the bonds between these atoms. Some of these compounds are dark brown in color and add caramel flavors to the sugar. Similar reactions occur in meat when you cook it and get a tasty brown crust.
Question: why does sugar turn brown when melted?

Answer:
Sugar is a molecule made of a particular arrangement of carbon, oxygen, and hydrogen atoms. When you heat up the sugar, this adds energy to the system that can be used to fuel chemical reactions that rearrange the bonds between these atoms. Some of these compounds are dark brown in color and add caramel flavors to the sugar. Similar reactions occur in meat when you cook it and get a tasty brown crust.
**Question:** why does sugar turn brown when melted?

**Answer:**
Sugar is a molecule made of a particular arrangement of carbon, oxygen, and hydrogen atoms. When you heat up the sugar, this adds energy to the system that can be used to fuel chemical reactions that rearrange the bonds between these atoms. Some of these compounds are dark brown in color and add caramel flavors to the sugar. Similar reactions occur in meat when you cook it and get a tasty brown crust.
Question: why does sugar turn brown when melted?

Answer:
Sugar is a molecule made of a particular arrangement of carbon, oxygen, and hydrogen atoms.
When you heat up the sugar, this adds energy to the system that can be used to fuel chemical reactions that rearrange the bonds between these atoms.
Some of these compounds are dark brown in color and add caramel flavors to the sugar.
Similar reactions occur in meat when you cook it and get a tasty brown crust.
Question: Were major news outlets established with political bias or was it formed over time?

Answer:
This is impossible due to the problem of “anchoring.”
Consider a world where people on the right want the tax rate to be 1% lower and people on the left want the tax rate to be 1% higher. The news outlet reports both sides with equal respect. Then one side decides to play a little trick, and declare that they want the tax rate to change by 90% instead of 1%. [...]
**Question:** Were major news outlets established with political bias or was it formed over time?

**Answer:**
This is impossible due to the problem of “anchoring.” Consider a world where people on the right want the tax rate to be 1% lower and people on the left want the tax rate to be 1% higher. The news outlet reports both sides with equal respect. Then one side decides to play a little trick, and declare that they want the tax rate to change by 90% instead of 1%.[…]
Miscellaneous

• Acknowledge limitation of the answer / specify the scope of the answer
  • Q: Why are there such drastic differences in salaries between different countries?
  • A: I’m mostly talking tech.

• Provide the original source of the answer
  • “The person who installed my heat pump explained this to me …”

• Express sentiment about the question, or other answers
  • “Good God, the amount of misinformation upvoted is hurting…”

• Refer to other answers in the platform
  • “/raskhistorians has a few excellent discussions about this”
Functional Roles

• Organizational Sentence
• Answer Summary
• Answer
• Example
• Auxiliary Information
• Miscellaneous
Datasets

Three types of long-form answers
ELI5: Why are skyscraper windows still washed by hand?

I worked on a window-washing robot that cleaned acres of rooftops over a huge commercial greenhouse. Worked great, except when it didn't, and would either break down completely or just get lost and start climbing the wrong parts of the structure. Then repair techs and manual window washers still have to be employed. I think this ends up being a cost/benefit problem where the reliability of our robots and price of implementation isn't quite at the point where it makes this commercially viable for skyscrapers. For what it's worth, I think the Twin Towers actually used a washer robot on the upper floors to limited success.

Fan et al. 2019
ELI5: Why are skyscraper windows still washed by hand?

I worked on a window-washing robot that cleaned acres of rooftops over a huge commercial greenhouse. Worked great, except when it didn't, and would either break down completely or just get lost and start climbing the wrong parts of the structure. Then repair techs and manual window washers still have to be employed. I think this ends up being a cost/benefit problem where the reliability of our robots and price of implementation isn't quite at the point where it makes this commercially viable for skyscrapers. For what it's worth, I think the Twin Towers actually used a washer robot on the upper floors to limited success.

WebGPT: How do our minds estimate the passage of time?

Studies involving the brain's perception of time passage have shown a few things. One suggests the way we perceive time depends on changes in what we see. When the speed of what we see changes, our perception of time will change accordingly [2]. Another study shows the brain reacts to time passing more quickly if we are busy or enjoying something. On the contrary, when bored, the brain will decipher this as time passing more slowly [3]. It is believed we have two different systems for processing time, one for circadian rhythm and another for time passage awareness [1].

1. Your Brain Has 2 Clocks (www.scientificamerican.com)
2. Physics explains why time passes faster as you age (qz.com)
3. A stopwatch on the brain's perception of time (www.theguardian.com)
ELI5: Why are skyscraper windows still washed by hand?

I worked on a window-washing robot that cleaned acres of rooftops over a huge commercial greenhouse. Worked great, except when it didn’t, and would either break down completely or just get lost and start climbing the wrong parts of the structure. Then repair techs and manual window washers still have to be employed. I think this ends up being a cost/benefit problem where the reliability of our robots and price of implementation isn’t quite at the point where it makes this commercially viable for skyscrapers. For what it’s worth, I think the Twin Towers actually used a washer robot on the upper floors to limited success.

WebGPT: How do our minds estimate the passage of time?

Studies involving the brain’s perception of time passage have shown a few things. One suggests the way we perceive time depends on changes in what we see. When the speed of what we see changes, our perception of time will change accordingly [2]. Another study shows the brain reacts to time passing more quickly if we are busy or enjoying something. On the contrary, when bored, the brain will decipher this as time passing more slowly [3]. It is believed we have two different systems for processing time, one for circadian rhythm and another for time passage awareness [1].

1. Your Brain Has 2 Clocks (www.scientificamerican.com)
2. Physics explains why time passes faster as you age (qz.com)
3. A stopwatch on the brain’s perception of time (www.theguardian.com)

NQ: what does it mean to be a subject matter expert

A subject matter expert (SME) or domain expert is a person who is an authority in a particular area or topic. The term domain expert is frequently used in expert systems software development, and there the term always refers to the domain other than the software domain. …The development of accounting software requires knowledge in two different domains: accounting and software. Some of the development workers may be experts in one domain and not the other. A SME should also have basic knowledge of other technical subjects.
Annotation

Two-stage: (1) validity check (2) sentence-level role annotation

Three-way annotated by undergrad linguistic students (Fleiss kappa = 0.45)

<table>
<thead>
<tr>
<th>Data</th>
<th>Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELI5</td>
<td>411 (2674)</td>
</tr>
<tr>
<td>WebGPT</td>
<td>98 (551)</td>
</tr>
<tr>
<td>NQ</td>
<td>131 (698)</td>
</tr>
<tr>
<td>Total</td>
<td>542 (3,372)</td>
</tr>
</tbody>
</table>

Data statistics: # answer paragraphs ( # sentences)
Role Distribution

![Role Distribution Chart]
~50% sentences serve roles other than directly answering the questions.
Role Distribution

Distribution varies across different types of long-form answers.
Role Distribution

More examples & miscellaneous in ELI5
More auxiliary information in WebGPT/NQ
This is a hard question to answer. Of course, since we occupy finite space, our properties must be finite. But the exceedingly complex structure of the brain and other systems [...] makes it hard to calculate the amount of "data" that we can store. We don't have a way to measure resolution of life or the quality of everyday noises. [...] Most of our memories are vague recollections. [...] However, we can estimate average capacity, although it has no clear meaning in our terms of cognition. That number would be 2.5 petabytes. [...] Even so, this number varies as well from person to person. [...]
This is a hard question to answer. Of course, since we occupy finite space, our properties must be finite. But the exceedingly complex structure of the brain and other systems [...] makes it hard to calculate the amount of "data" that we can store. We don’t have a way to measure resolution of life or the quality of everyday noises. [...] Most of our memories are vague recollections, [...]. However, we can estimate average capacity, although it has no clear meaning in our terms of cognition. That number would be 2.5 petabytes [...] Even so, this number varies as well from person to person, [...].

**Human-written answer**

No. Biologically the brain is made up of different regions called neurons. For example, your olfactory bulb is in the brain stem which in turn is in the hypothalamus which is part of the forebrain. [...] The neurons for your hearing are in the auditory cortex [...] Technically your brain has about 1 billion neurons [...] Technically you can say that a person has about a billion neurons, but that number is in reality only around 0.0500 to 1 billion. [...]
Model-generated Answers

Annotated 114 answers from Krishna et al. 2021
Model-generated Answers

Annotated 114 answers from Krishna et al. 2021

More disagreement among annotators (kappa = 0.31 v.s. 0.45)
Model-generated Answers

Annotated 114 answers from Krishna et al. 2021

Different role distribution between human and model-generated answers.
Model-generated Answers

Annotated 114 answers from Krishna et al. 2021

Less examples & summary
More auxiliary information
Question: Can the capacity of our brains be roughly measured in bytes?

This is a hard question to answer. Of course, since we occupy finite space, our properties must be finite. But the exceedingly complex structure of the brain and other systems [...] makes it hard to calculate the amount of "data" that we can store. We don't have a way to measure resolution of life or the quality of everyday noises. [...] Most of our memories are vague recollections, [...]. However, we can estimate average capacity, although it has no clear meaning in our terms of cognition. That number would be 2.5 petabytes [...] Even so, this number varies as well from person to person.[...]

Human-written answer

No. Biologically the brain is made up of different regions called neurons. For example, your olfactory bulb is in the brain stem which in turn is in the hypothalamus which is part of the forebrain. [...] The neurons for your hearing are in the auditory cortex [...]. Technically your brain has about 1 billion neurons [...]. TL;DR: Technically you can say that a person has about a billion neurons, but that number is in reality only around 0.0500 to 1 billion. [...]
This is a hard question to answer. Of course, since we occupy finite space, our properties must be finite. But the exceedingly complex structure of the brain and other systems […] makes it hard to calculate the amount of data we can store in a way to measure resolution of life or the quality of everyday noises. […] Most of our memories are vague recollections, […] However, we can estimate average capacity, although it has no clear meaning in our terms of cognition. That number would be 2.5 petabytes […] Even so, this number varies as well from person to person. […] Analysing discourse structure reveals a gap between human-written and model-generated answers.
Automatic Discourse Analysis

Task: Given a question $q$ and its long form answers consisting of sentences $s_1$, $s_2$, … $s_n$, assign each sentence $s_n$ one of the six roles.
Automatic Discourse Analysis

Task: Given a question $q$ and its long form answers consisting of sentences $s_1$, $s_2$, … $s_n$, assign each sentence $s_n$ one of the six roles.

Data: Train: ELI5; Evaluation: ELI5/WebGPT, NQ, ELI5_MODEL

Model:

- Classification model: We use [CLS] token from RoBERTa and encodes each sentence separately to predict the role.

- Seq2Seq model: We use T5 model to encode the entire answer paragraph and output the roles sequentially.
Role prediction - result

Result on in-domain test data
Role prediction - result

Trained classifier has comparable performance to human annotators.

Result on in-domain test data

- Majority
- RoBERTa
- T5-large
- Human
Role prediction - result

Model performance degrades on OOD data (including model-generated answers).

Accuracy on in-domain and out-of-domain data
Role prediction - result

Model performance degrades on OOD data (including model-generated answers).

Accuracy on in-domain and out-of-domain data
Takeaways

• Complex structure of long-form answers! About half of the sentences in long-form answers serve roles other than directly answering the question.

• Discourse analysis reveals the gap between human-written and model-generated answers.

• We hope our work inspire more discourse-level evaluation and modelling for long-form QA!

Data and code available at our website

Thank you!