Does it Make Sense? And Why? A Pilot Study for Sense-Making and Explanation

[ACL 2019]

Cunxiang Wang, Shuailong Liang, Yue Zhang, Xiaonan Li, Tian Gao

Presenters:

Siba Smarak Panigrahi and Sohan Patnaik

Reading Session VIII
Kharagpur Data Analytics Group
IIT Kharagpur

June 20, 2021

Important Notes

- We expect you all to have certain queries regarding the presentation, certain intricate doubts maybe... Please put them in the chat of this meeting (if you feel shy) else unmute and speak.
- Finally, a Google Form [Link] will be released for feedback but most importantly, we ask you to put up name of any AI-related paper to be presented in the upcoming sessions!

Link to GitHub Repo : Click Here
Link to join Slack Workspace : Click Here
Link to KDAG YouTube Channel : Click Here
Link to doc on good AI Blogs/Resources/Topics : Click Here

Outline

- Introduction
- Testset
  - Task
  - Annotation GuideLines
  - Corpus Analysis
- Experiments
- Results and Analysis
- Case Study
- Conclusion

Introduction

- This paper proposes a benchmark dataset to test whether a system can differentiate natural language statements that make sense from those that do not make sense as well as asks to identify the most crucial reason why a statement does not make sense.

- Limitations to prior benchmark datasets:
  - No direct metric to quantitatively measure sense making capability.
  - They do not explicitly identify the key factors required in a sense making process

Introduction

- The first task is to choose from two natural language statements with similar wordings which one makes sense and which one does not make sense.
- The second task is to find the key reason why a given statement does not make sense
- A statement pair classification rather than labelling each statement 'true' or 'false' in the absolute sense is used because it is easy to cite a counterexample for any single 'true' or 'false' statement.

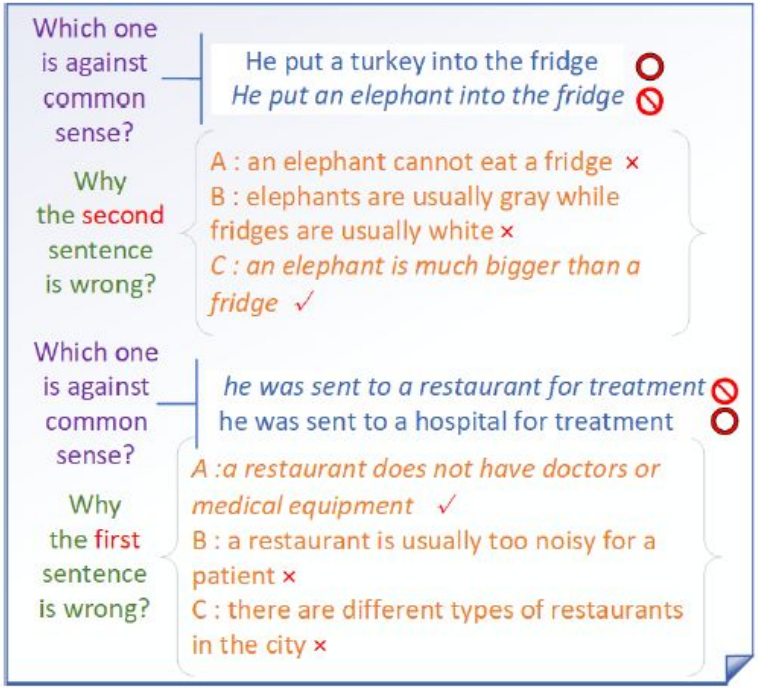

Figure 1: Samples of our dataset

Testset

- Task
  - Formally, each instance in our dataset is composed of **5 sentences: {s1, s2, r1, r2, r3}**. s1 and s2 are two similar sentences which in the same syntactic structure and differ by only few words, but only one of them makes sense while the other does not
  - **Sen-Making**
    - requires the model to identify which one is valid. For the invalid sentence, there are three optional reasons r1, r2 and r3 to explain why the sentence is invalid.
  - **Explanation**
    - requires that the only one correct reason be identified from two other confusing ones.
  - Accuracy score to evaluate both subtasks.

Testset

- Annotation Guidelines
  - **Avoid complex knowledge**; **focus on daily common sense**; **understandable questions**
  - Confusing reasons should **contain more important words** like entities and activities in the against-common-sense statements, for example, the confusing reasons of “he put an elephant into the fridge” should better contain both “elephant” and “fridge”
  - **Confusing reasons to be related to the statements and correct reasons** and aligned with problem context else can be easily captured by BERT (models the sentence contexts explicitly)
  - Three option reasons should be only **related to the incorrect sentence**.
  - **Confusing sentences should be correct themselves**. Otherwise, the models may simply ignore the incorrect options without considering the causal relations
  - Try to make the incorrect statement nearly as long as the correct statement, and the right reason neither too long nor too short among the three reasons.

Testset

- Corpus Analysis
  - Average length of two statements in the SenMaking task are both 8.26.
  - Average length of true reasons is 7.63, and 7.77 for confusing reasons'
  - We find that incorrect statements have much the same negative different words compared with correct statements! (What is different word?)

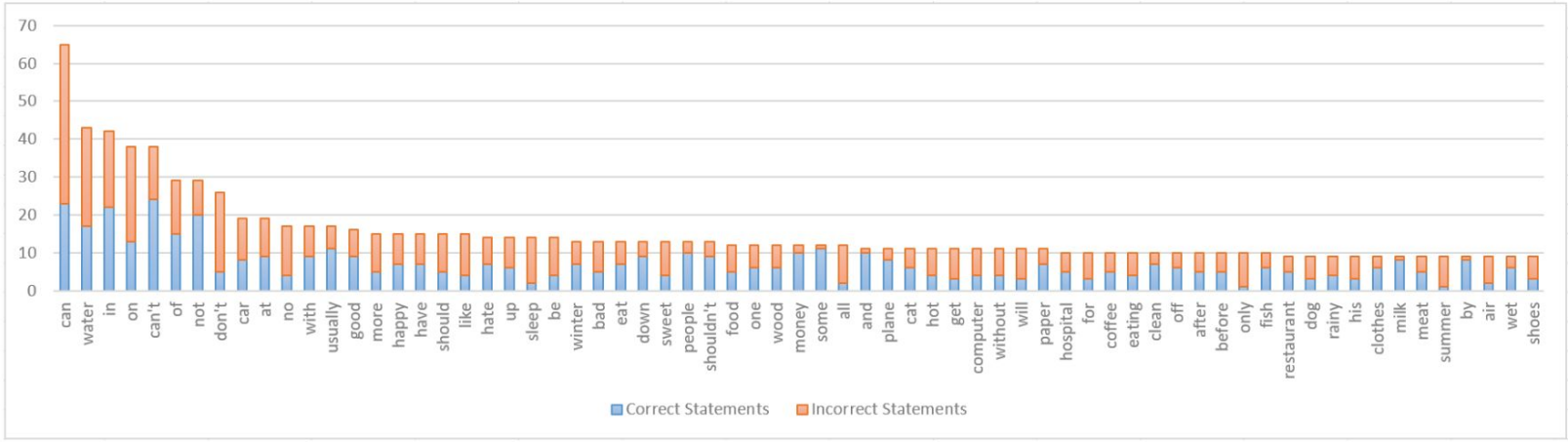

Figure 2: Number of 'Different Words'

Experiments

- For the **sense making task**,
  - calculate score of both statements, then choose the one with lower scores as correct one.

$$score = (p_{w_1} * p_{w_2} * \ldots * p_{w_n})^{(-1/n)} = (\prod_{i=1}^{n} P(w_i \mid w_1 \ldots w_{i-1} w_{i+1} \ldots w_n))^{-1/n}$$

- For **explanation**,
  - concatenate the statement with the each reason
  - use the three concatenated sentences to calculate scores.
  - Example: “he put an elephant into the fridge” + its optional reasons = “he put an elephant into the fridge is against common sense because an elephant cannot eat a fridge”
- For **human evaluation**,
  - If more than half testees do one sample wrong (either Sen-Making or Explanation), we will rewrite or abolish the sample; else keep it and record it in results

Results and Analysis

For Sen-Making, (fine-tuned) ELMo does better than BERT; However, BERT beats ELMo in Explanation.

ELMo cannot handle the causal relationship between the incorrect statements and the reasons.

In contrast, BERT is significantly better than random guess in Explanation. This intuition comes from the fact that BERT has been trained on Next Sentence Prediction, which assists to handle the logic relationship between two sentences.

Model	Sen-Making	Explanation
Random	50.0%	33.3%
ELMo	69.4%	33.4%
BERT	70.1%	45.6%
fine-tuned ELMo	74.1%	34.6%
Human Performance	99.1%	97.8%

Table 1: Experimental Results

Case Study

- Case study that shows why fine-tuning can help model identify common sense and but cannot help in much inference.
- Consider the example, “New York is located in the northeastern part of USA” vs “the USA is located in the northeastern part of New York”.
- ELMo is incorrect for both Sense Making and Explanation task for this pair
- After fine tuning and training on a corpora that has sentences “New York is a city” and “the USA is a country”, the ELMo can figure out the incorrect sentence but still cannot pick out the correct reason.
- It is difficult for language models to capture common sense as LSTM language models are not suited to make multi-step inference.

Conclusion

- The authors provide a benchmark for
  - directly evaluating whether a system has the capability of sense making and explanation
  - evaluating models trained over the large raw text
  - a common sense database
- Results show that sense making remains a technical challenge for such models, whereas inference is a key factor that is missing

Thank You!