Incorporating External Knowledge to Enhance Tabular Reasoning

J. Neeraja[(1)], Vivek Gupta[(2)], Vivek Srikumar[(2)]

(1) IIT Guwahati; (2) University of Utah

1. Tabular Inference Problem

- Inference task where premises are tabular in nature
- Given a premise table determine hypothesis is true (entailment), false (contradiction), or undetermined (neutral), i.e. tabular natural language inference.

New York Stock Exchange	
Type	Stock exchange
Location	New York City, New York, U.S.
Founded	May 17, 1792; 226 years ago
Currency	United States dollar
No. of listings	2,400
Volume	US$20.161 trillion (2011)

H1: NYSE has fewer than 3,000 stocks listed.
H2: Over 2,500 stocks are listed in the NYSE.
H3: S&P 500 stock trading volume is over $10 trillion.

- Example InfoTabS dataset (Gupta et al., 2020), **H1**: entailed ; **H2**: contradictory ; **H3**: neutral

2. Motivation

1. Recent work mostly focuses on **building** sophisticated **neural models**.
2. How will models designed for the **raw text adapt for tabular data**?
3. How to **represent data** and **incorporate knowledge** into these model?
4. Can better **pre-processing** of **tabular information** enhance table comprehension?

3. Challenges

1. Poor Table Representation
2. Missing Lexical Knowledge
3. Presence of Distracting Information
4. Missing Domain Knowledge

Main Question
Can we fix the above problems by changing how tabular information is provided to a standard model?

4. Poor Table Representation

- Using universal template → Most sentences are ungrammatical or non-sensible

✗ The Founded of New York Stock Exchange are May 17, 1792; 226 years ago.

Better Paragraph Representation

- Entity specific templates : use value entity types DATE, MONEY or CARDINAL or BOOL

✓ New York Stock Exchange was founded on May 17, 1792; 226 years ago.

- Add category information.

New York Stock Exchange is an **organization**.

More grammatical and meaningful sentences

5. Missing Lexical Knowledge

- Limited training data → affects interpretation of hypernym words such as *fewer*, *over* and negations.

Implicit Knowledge Addition

Can pre-training on large NLI dataset help?

1. Pre-training with MNLI data
2. Then, fine-tune on InfoTabS

Exposes model to diverse lexical constructions. Representation is better tuned for the NLI task.

6. Distracting Information Issue

- Only select rows are relevant for a given hypothesis. E.g. ***No. of listings*** is enough for H1 and H2.
- Due to BERT tokenization limit, useful rows in the longer tables cropped.

Distracting Row Removal

- Select only rows relevant to hypothesis.
- Use Alignment based retrieval algorithm with fastText vectors (Yadav et al. (2019, 2020))

E.g. for H1 H2, new prune table :

New York Stock Exchange	
No. of listings	2,400

7. Missing Domain Knowledge

- For H3, we need to interpret ***Volume*** in financial context.

✓ In capital markets, volume, is the total number of a security that was traded during a given period of time.

rather than

✗ In thermodynamics, volume of a system is an extensive parameter for describing its phase state.

Explicit Knowledge Addition

- Add explicit information to enrich keys.
- This improves model's ability to disambiguate meaning of keys.

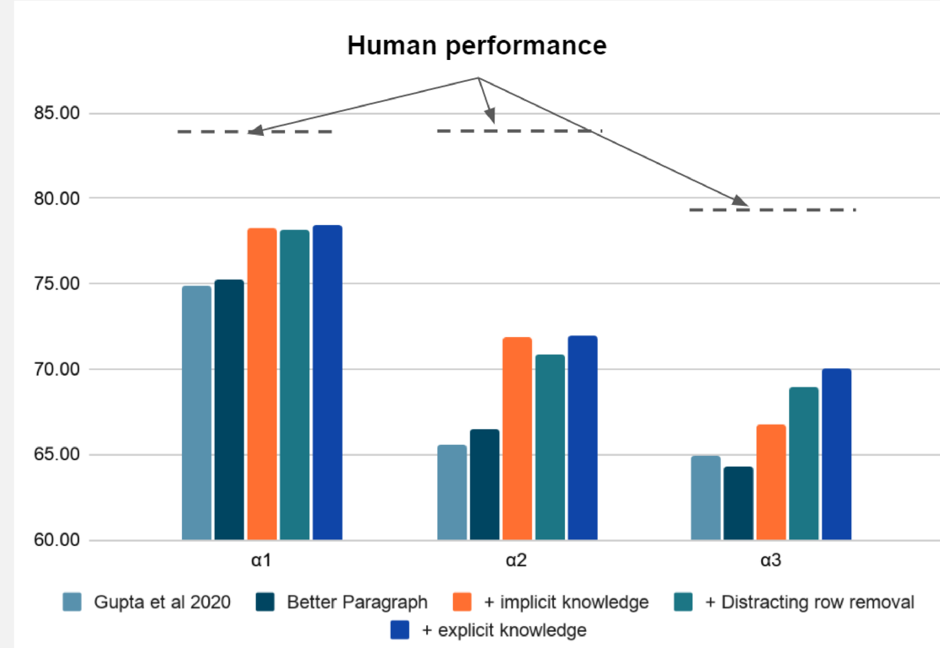
Approach

1. Use **BERT** on wordnet examples to find key embeddings
2. Get key embeddings from premise using **BERT**
3. Find the **best match** and add it definition to premise.

Add to the table in the end for H3

Volume: total number of a security that was traded during a given period of time.

8. Experimental Results



1. Significant improvement in adversarial α_2 and α_3 dataset
2. **Ablation Study**: All changes are needed, knowledge addition being the most important.

9. Conclusion

1. Proposed pre-processing lead to **significant improvements**
2. Propose approach beneficial for adversarial α_1 and α_2 dataset
3. Solutions applicable to *question answering* and *generation* problems with both the *tabular* and *textual inputs*
4. Proposed modifications should be **standardized across other table reasoning tasks**

Data and Software:
https://infotabs.github.io

10. References

- Gupta et. al. *INFOTABS: Inference on Tables as Semi-structured Data.* ACL'20.
- Yadav et. al. *Alignment over heterogeneous embeddings for question answering.* NAACL'19.
- Yadav et. al. *Unsupervised Alignment-based Iterative Evidence Retrieval for Multi-hop Question Answering.* ACL'20.