TRANS-KBLSTM: An External Knowledge Enhanced Transformer BiLSTM model for Tabular Reasoning

Yerram Varun [1*], **Aayush Sharma**[1*], **Vivek Gupta**[2*†]
[1]Indian Institute of Technology, Guwahati ; [2]School of Computing, University of Utah
vgupta@cs.utah.edu; {y.varun, aayushsharma}@iitg.ac.in

Abstract

Natural language inference on tabular data is a challenging task. Existing approaches lack the world and common sense knowledge required to perform at a human level. While massive amounts of KG data exist, approaches to integrate them with deep learning models to enhance tabular reasoning are uncommon. In this paper, we investigate a new approach using BiLSTMs to incorporate knowledge effectively into language models. Through extensive analysis, we show that our proposed architecture, Trans-KBLSTM improves the benchmark performance on INFOTABS , a tabular NLI dataset.

1 Introduction

Understanding tabular or semi-structured knowledge presents a reasoning challenge for modern natural language processing algorithms. Recently, Chen et al. (2020) through TabFact and Gupta et al. (2020) via INFOTABS presented this problem as a natural language inference problem (NLI, Dagan et al., 2013; Bowman et al., 2015, many others), where a model is asked to determine whether a hypothesis is entailed or refuted by a premise, or is unrelated to it (c.f. Table 1). One technique for modeling such tabular reasoning problems is to rely on the success of contextualized representations for the sentential variant of the problem (e.g., Devlin et al., 2019; Liu et al., 2019, etc.). To convert tabular data into a format suitable for these models, they are flattened using heuristics into phrases.

Recently, Neeraja et al. (2021) highlight the significance of adding world knowledge for the tabular inference task (c.f. Table 1). Their approach develops a knowledge addition strategy, namely *KG Explicit*, which expands the keys of a tabular premise with its definitions obtained from Wordnet and Wikipedia articles. These definitions are appended as a suffix to the original input as additional context. With this added additional knowledge, the model outperforms the original baseline. Despite improved effectiveness, knowledge addition has the following drawbacks: (a) **Knowledge Extraction.** *KG Explicit* disambiguates multiple key definitions using the table context, ignoring the hypothesis content entirely. Additionally, the extended definition contains hypothesis-unrelated and unnecessary additional functional terms. All of these factors contribute to erroneous key-sense disambiguation and additional noise. (b) **Knowledge Addition.** *KG Explicit* adds knowledge by appending a suffix definition to existing inputs instead of using more effective semantic representations such as Knowledge Embedding (Graph Embedding or Learned representations). (c) **Knowledge Integration.** Finally, utilizing tokenized input BERT (Devlin et al., 2019) to fuse word-pair relations yields considerably weaker semantic linkages between premise, hypothesis, and the external knowledge.

In this work, we propose a solution to these issues. We drew inspiration from Chen et al. (2018) and utilize relational connections between premise and hypothesis to extract important knowledge relations from ConceptNet (Speer et al., 2017) and

	James Hetfield
Birth Name	James Alan Hetfield
Born	Aug. 3, 1963(age 58), California, U.S.
Genres	Heavy metal, thrash metal, hard rock
Occupation(s)	Musician, Singer
Instruments	Vocals, Guitar
Years active	1978-present
Labels	Warner Bros, Elektra, MegaForce
Hypothesis	James Hetfield was born on the west coast of the USA.
Focused Relation	coast $\xleftarrow{AtLocation}$ california
Human	Entailment
RoBERTa	Neutral
Trans-KBLSTM	Entailment

Table 1: An INFOTABS example demonstrating the need of knowledge augmentation. Predicting the Gold label requires broad understanding of *California* is located on the *Coast*. In the table, for each row the first column represents the keys (unique identifiers) and the second column represents their corresponding values (attributes).

*Equal Contribution †Corresponding Author

Wordnet (Miller (1992)). This enhancement reduces noise in knowledge addition, resulting in improved **Knowledge Extraction**. We embed relational terms in sentences using sentence transformers (Reimers and Gurevych, 2019) to encode semantic representations of the relation, comparable to Gajbhiye et al. (2021), culminating in successful **Knowledge Addition**. Finally, for effective **Knowledge Integration**, we combine these relational embeddings into a word-level language model, using BiLSTM (Hochreiter and Schmidhuber, 1997), and backpropagate using our proposed BiLSTM and transformer architecture together to enhance model inferencing capabilities.

Our proposed model, Trans-KBLSTM, outperforms the earlier baseline, i.e., *KG Explicit* in full as well as limited supervision setting, substantially for some specific categories. Furthermore, knowledge addition via Trans-KBLSTM improve model *lexical*, *multi-row* and *Numerical* reasoning. We also performed a detailed ablation study to understand the importance of each component. Our contributions are as follows:

1. We address the challenges inherent in existing techniques, e.g., KG Explicit, for explicit knowledge addition in tabular reasoning.
2. We investigate a more efficient knowledge extraction method that involves using knowledge embeddings rather than directly appending them to the input.
3. We propose a novel architecture, namely Trans-KBLSTM, for integrating word-level knowledge effectively with BiLSTM's encoders with state-of-the-art transformers such as BERT.
4. Through extensive experiments, analysis and ablation studies, we demonstrate that Trans-KBLSTM improves reasoning for INFOTABS dataset.

The dataset, and associated scripts, are available at https://trans-kblstm.github.io/.

2 Proposed Trans-KBLSTM Model

We highlight the main model components and their implementation details in this section. We begin with a description of the knowledge relations retrieval technique, followed by a discussion of the model architecture's core components.

2.1 External Knowledge Relations Retrieval

It is challenging to retrieve contextually relevant knowledge relations from the knowledge graphs.

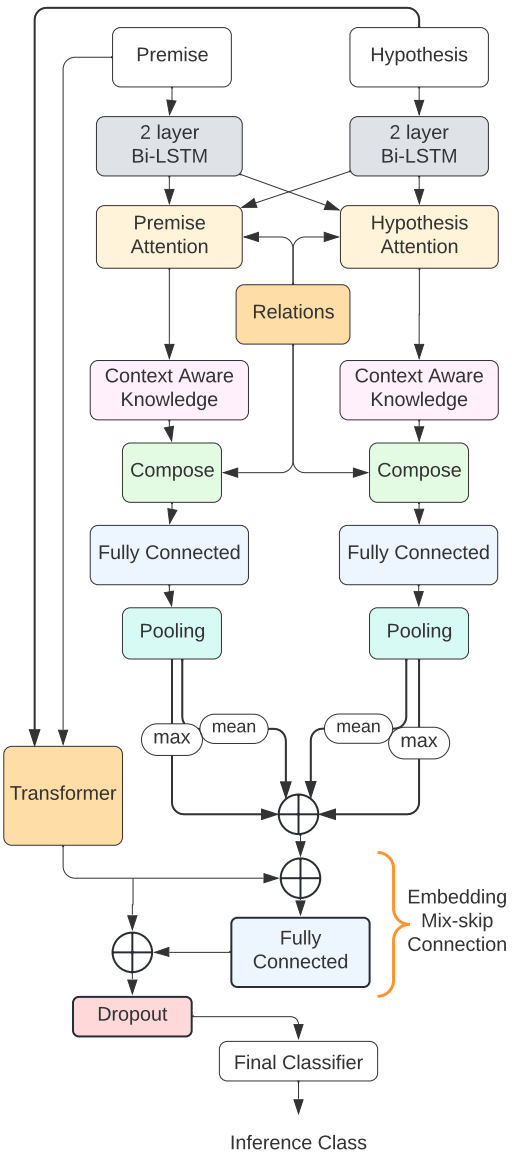

Figure 1: High level flowchart of Trans-KBLSTM.

The challenge is to retrieve task-relevant knowledge relations from massive volumes of noisy Knowledge Graph data. Our method is inspired by Chen et al. (2018), which considers a connection to be significant if the knowledge graph contains the term pair relations.

Relational Connections We define relational connections between two sentences through external relational knowledge between each pair of words in the sentences. The token level relation connections are based on word triples derived from the knowledge graphs.

Relational Connections Retrieval Stop words and punctuation are first removed from the premise and hypothesis. Then, we analyze the knowledge relational connections between the premise and hy-

pothesis token pairs and compute the relationship attention matrix, A^r_{ij}, as follows:

$$A^r_{ij} = \begin{cases} 1 & i^{th} \text{ and } j^{th} \text{ words are } \textbf{related} \\ 0 & i^{th} \text{ and } j^{th} \text{ words are } \textbf{not related} \end{cases}$$

Each knowledge relational triple, consisting of two token terms (one from each premise and hypothesis) and their respective relationship is transformed into a complete grammatical sentence. For instance, the triple {Day, *Antonym*, Night} is transformed into "*Day* is the *opposite* of *Night*". For a complete list of knowledge templates refer to table 5 in Appendix §B. We utilize sentence transformers, as presented in Reimers and Gurevych (2019), to convert the relationship phrase e.g. *"is opposite of"* in the preceding example into high-level semantic representations. The contextual representations denote the relational pair's across relational pairs.

Relational Connection Embedding The contextual knowledge connections between premise and hypothesis token pairs are used to generate a relational vector, R_{ijk}. Each marginal vector R_{ij} is the k dimension BERT representation for the *"Relation Connection Sentence"* in the previously described sentential form constructed using the relationship between the i^{th} premise word and the j^{th} hypothesis word. For words whose relations are absent from knowlege source, we initialize the R_{ij} vector with 'zero' values.[1]

2.2 Model Architecture Details

Next, we described several components of our proposed model. Figure 1 describe the high level architecture of the **Trans-KBLSTM** model.

Transformer We encode the premise and hypothesis using RoBERTa(Liu et al., 2019) to generate contextual word embeddings. Consider $P = \{p_i\}_{i=1}^m$ as table premise of length m and $H = \{h_j\}_{j=1}^n$ as hypothesis of length n. We input these premise-hypothesis pairs to RoBERTa as :

$$S = [\text{<s> } P \text{ </s> } H \text{ </s>}] \; ; \; T_r = \text{RoBERTa}(S)$$

Here, T_r denotes the context-aware representations of the premise and hypothesis sentence.

Encoding Premise and Hypothesis The encoder approach is inspired from Chen et al. (2018). We encode the Premise, $P = \{p_i\}_{i=1}^m$ and Hypothesis, $H = \{h_j\}_{j=1}^n$ using bidirectional LSTMs (BiLSTMs). We embed p_i and h_i into d_e dimensional vectors $\left[\mathbf{E}(p_1), ...\mathbf{E}(p_m)\right]$ and $\left[\mathbf{E}(h_1), ...\mathbf{E}(h_n)\right]$ using embedding matrix $\mathbf{E} \in \mathbb{R}^{d_e \times |V|}$, where $|V|$ is the Vocabulary size and $\mathbf{E}$ can be initialized with pretrained embeddings. We feed the premise-hypothesis pairs into BiLSTM encoders (Hochreiter and Schmidhuber (1997)) to generate context-aware hidden states p^s and h^s.

$$p^s = \text{BiLSTM}(\mathbf{E(p)}, i) \; ; \; h^s = \text{BiLSTM}(\mathbf{E(h)}, i)$$

$$p^s \in \mathbb{R}^{m \times l_k} \text{ and } h^s \in \mathbb{R}^{n \times l_k}$$

Here, l_k is the LSTM hidden state size. Following that we apply embedding dropout (Gal and Ghahramani (2016)) to enhance variation and prevent overfitting (Zaremba et al. (2014)).

Premise and Hypothesis Attention Module To assess the contribution of external knowledge to the premise (and hypothesis), we utilize the Multi-Head dot-product attention (Vaswani et al., 2017) across knowledge representations and premise-hypothesis encoding. We calculate premise hypothesis relation values by normalizing relational connection embedding (R_{ijk}) with respect to column-axis (1), to obtain $R^{prem}_{jk} \in \mathbb{R}^{n \times k}$ which is the average premise relation for every hypothesis word.

$$R^{prem}_{jk} = \sum_{i=1}^{m} \frac{R_{ijk}}{m}$$

To apply dot product attention, we then reduce the dimension of the relation matrix to BiLSTM hidden state dimension, i.e., l_k.

$$R^r_{jk} = F^r_P(R^{prem}_{jk}) \quad \in \mathbb{R}^{n \times l_k}$$

where, F^r_P is a single layer neural network.

To highlight the importance of premise and its relations to hypothesis we utilise the premise attention head. The context-aware hypothesis hidden state h^s is used as queries, premise hidden state is used as keys and reduced premise hypothesis relation values are used as values. The attention function can be defined as follows:

$$\text{Attention}(h^s, p^s, R^r_{jk}) = \text{softmax}(\frac{h^s p^{sT}}{\sqrt{l}})R^r_{jk}$$

where, the multi-head attention is defined:

$$\begin{aligned} h^{att}_p &= \text{MH}(h^s, p^s, R^r_{jk}) \\ &= \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^o \end{aligned}$$

[1] Experiment with non-zero random initialization ref §3.3.

Here, $\text{head}_i = \text{Attention}(h^s W_i^q, p^s W_i^k, R_{jk}^r W_i^v)$ and W_i^q, W_i^k, and W_i^v are projection matrices and i is the number of attention heads. The output $h_p^{att} \in \mathbb{R}^{n \times l_k}$ is a context matrix that is attention-weighted according to the strength of the premise and its relationships to each of the hypothesis words. We also extract P^{att}, the premise multi-head attention attention weights. In hypothesis attention module, we use hypothesis attention head to highlight the importance of hypothesis and its relations to premise. Similar to the premise attention module, we calculate[2] $p_h^{att} \in \mathbb{R}^{m \times l_k}$, attention-weighted context matrix measuring the importance of premise and relations to each of the hypothesis. We also extract H^{att}, the hypothesis multi-head attention attention weights.

Context Aware External Knowledge ExBERT (Gajbhiye et al., 2021) uses a mixture model to weigh the balance of external relations and premise-hypothesis during inference. We construct attention-weighted external knowledge relations using Multi-head attention weights obtained in the attention modules.

$$P^{CE} = \sum_{k=1}^{h} P_{ij}^{att} R_{ijk} \;\; ; \;\; H^{CE} = \sum_{k=1}^{h} H_{ij}^{att} R_{ijk}$$

Composition Layer p^s encodes the individual word representations of the premise while p_h^{att} is the context representation of the premise aligned to the hypothesis. We can obtain word-level inference information for each word in the premise by composing them together with attention weights and context-aware external knowledge. We can do the same calculation for hypothesis, h^s and h_p^{att}:

$$p^m = G_P([p^s; p_h^{att}; p^s - p_h^{att}; p^s * p_h^{att}; \sum_{j=1}^{n} P_{ij}^{CE}])$$

$$h^m = G_H([h^s; h_p^{att}; h^s - h_p^{att}; h^s * h_p^{att}; \sum_{j=1}^{n} H_{ij}^{CE}])$$

Here, G_P and G_H are 2-layer neural networks with Dropout and ReLU activation (Agarap (2018)) that compose the knowledge relations and premise-hypothesis contextual vectors into a unified knowledge aware context vector.

Pooling Layer The pooling layer creates fixed-length representations from the knowledge-aware premise and hypothesis context vectors.

$$p_{mean} = \text{MeanPool}(p^m) \; ; \; p_{max} = \text{MaxPool}(p^m)$$

$$h_{mean} = \text{MeanPool}(h^m) \; ; \; h_{max} = \text{MaxPool}(h^m)$$

[2] More details can be found in section A in §Appendix

Embedding mix-skip connection To effectively integrate transformer embeddings with representations from premise and hypothesis, we introduce an Embedding mix-skip connection, where the embeddings are concatenated and passed through a fully connected layer with a skip connection to transformer embeddings. Skip connections, introduced by He et al. (2016), provides a shortcut to gradient flow and preserve the context between layers.

$$f = [p_{mean}, p_{max}, h_{mean}, h_{max}]$$

$$f' = T_r + F_c([T_r, f])$$

Here, F_c is a two-layer neural network with dropout and ReLU activation. Finally, f' is passed through a classification layer to obtain the inference class.

3 Experiment and Analysis

Our experiments study the following questions.

RQ1: Is our proposed model competent in using external knowledge sources effectively to enhance performance across INFOTABS evaluations sets?

RQ2: How effective is our approach in settings with little supervision? How much supervision is necessary to outperform benchmark models?

RQ3: (a) Which reasoning types is our proposed model most effective at boosting? (b) Is our approach equally effective across all domains, that is, across all table categories? (c.f.§C)

RQ4: How does the model component choices impact performance? (a) To what extent are skip connections, (b) knowledge embeddings, (c) additional MNLI (Williams et al., 2018) pre-finetuning, and (d) a bigger pre-trained model beneficial?

3.1 Experimental setup

Here, we discuss the datasets, external knowledge sources, and the models used in the experiments.

Datasets. We use INFOTABS , a tabular Language inference dataset introduced by Gupta et al. (2020) for all our experiments. The dataset is diverse in categories and keys and requires background knowledge and semantic understanding of the text. Examples in INFOTABS are labeled with three types of inference: entailment, neutrality, and contradiction, based on their relation with premise tables. Along with the standard development set and test set (dubbed α_1), the dataset includes two adversarial test sets: a contrast set dubbed α_2 that

is lexically similar to α_1 but contains fewer hypotheses, and a zero-shot set dubbed α_3 that contains long tables from various domains with little key overlap with the training set.

Table Representation. To represent tables, we utilize Neeraja et al. (2021) *Better Paragraph Representation* (BPR) technique in conjunction with *Distracting Row Removal* (DRR). The BPR technique turns its rows into sentences using a universal template, enabling it to be used as the input for a BERT-style model. We utilize the DRR approach to reduce the premise table by identifying the most relevant premise sentence. For finding the most relevant rows, we use cosine similarity over fastText embeddings (Bojanowski et al. (2017)) and word alignment with the specified hypothesis. We select the top four aligned table rows from each premise table with hypotheses.

Knowledge Sources. We utilize ConceptNet, as introduced by Speer et al. (2017) to extract external commonsense knowledge to create relational occurrences. We notice that 85% of premise-hypothesis pairings contain at least one relationship in the ConceptNet database. To supplement the coverage, we also use Wordnet (Miller, 1992), to extract additional lexical word relations, namely *Synonyms*, *Antonyms*, *Hypernyms*, *Hyponyms* and *Co-Hyponyms*. After combining the two knowledge databases and removing duplicates, the number of non-zero relational connection pairings increases to 90%. We create an English directional single word relations dataset by merging ConceptNet and Wordnet. The combined KG source contains 11.2 million relation triples. For example in the table 1, the relational occurrence { "coast" ← "California"} extracted from Conceptnet, provide the necessary world knowledge required for correct inference.

Word Embeddings. We utilize pre-learned word embeddings to initialize the BiLSTM encoders. The premise and hypothesis words are embedded in 300-dimensional vectors using Glove embeddings [3], introduced by Pennington et al. (2014). Glove is a collection of 400,000-word embeddings learned using the Wikipedia, Common crawl, and Twitter datasets. We realize that the GloVe vocabulary covers 85.6% of the terms in INFOTABS dataset.[4]

[3] We also investigate fastText embeddings for representation, but it has only 77.4 % coverage of all tokens. [4] Due to limited supervision, we found that freezing word embedding during the BiLSTM training is beneficial. For the remaining unseen tokens, we initialized with zero vectors.

Models. To evaluate we compare our model with INFOTABS (Gupta et al., 2020) and Knowledge-INFOTABS (Neeraja et al., 2021) baselines, specifically we employ the following methods:

- **RoBERTa**. The original RoBERTa baseline of INFOTABS . We append and encode premise-hypothesis pairs with BPR with DRR representation and generate an inference label with the RoBERTa classification head.
- **KG Explicit.** Knowledge-INFOTABS introduced this baseline. The baseline uses the same RoBERTa classifier as the INFOTABS , except that the premise end is augmented with extracted premise row key definitions from Wordnet and Wikipedia sources before encoding and classifying using RoBERTa. Additionally, prior to appending, the method employs key sense disambiguation to assure that only relevant hypothesis context-related definitions are added. For example, for a table with category *"Person"* and key *"Spouse"*, the definition of *"Spouse"* from Wikipedia, i.e., *"Spouse is defined as a spouse is a significant other in a marriage, civil union, or common-law marriage."* is appended as a suffix.
- **Tok-KTrans**. We utilize Wordnet to expand premise hypothesis pairs with word relations in Tokens added transformers before encoding and classifying using RoBERTa. We extend the tokenizer by including relational tokens and appending the relationships with the following format - {<*KNW*> *[premise_word*$_1$ *: hypothesis_word*$_1$ *;* <*relation*$_1$> *] [premise_word*$_2$ *: hypothesis_word*$_2$ *;* <*relation*$_2$>*]* ...}. For example, The table *Jallikattu* contains a key ***Mixed Gender*** with a value *NO*. The hypothesis, *Jallikattu is a single **sex** sport* contradicts the premise table. We append the relation {<*KNW*> *[gender : sex ;* <*SYN*> *]*} as suffix to input prior to the RoBERTa classification.
- **Trans-KBLSTM.** This is our proposed model as described in the §2. For details on model training and hyper-parameters, refer to Appendix §G.

3.2 Results and Analysis

This section summarizes our findings concerning the research questions.

Full Supervision Setting. To assess the effectiveness of our method Trans-KBLSTM (i.e. RQ1), we train baseline and our model Trans-KBLSTM with 100% of training data. Table 2 shows the perfor-

mance (accuracy) for all models. We observe that Trans-KBLSTM outperform[5] all other baselines. On development, α_1 , and α_3 Trans-KBLSTM outperform 0.75 - 0.95 % with 100% training data.

Model	Dev	$\alpha 1$	$\alpha 2$	$\alpha 3$
w/o Knowledge	77.30	76.44	70.49	69.05
Tok-KTrans	78.17	76.19	70.75	69.77
KG Explicit	78.97	77.84	71.13	69.58
Trans-KBLSTM	**79.92**	**79.62**	**72.10**	**70.21**

Table 2: Performance in terms of accuracy with full supervision. **w/o Knoweldge** represent RoBERTa INFOTABS (Gupta et al., 2020) baseline, **KG Explicit** represent Knowledge-INFOTABS (Neeraja et al., 2021) baseline, **Tok-KTrans** is the token appended transformers and **Trans-KBLSTM** represent our proposed model. Reported number are average over three random seeds with standard deviation of 0.27 (w/o KG), 0.69 (Tok-KTrans), 0.23 (KG Explicit) and 0.36 (Trans-KBLSTM). All improvements are statistically significant with Student's t-test $p < 0.05$ except α_2 with KG Explicit.

Limited Supervision Setting. To ensure that our model works effectively in low-resource scenarios (i.e., RQ2), we analyze models trained under limited supervision. We randomly sampled {1, 2, 3, 5, 10, 15, 20, 25, 30, 50, and 100} data in an incremental method [6]. We experimented three times using random seeds for sampling/training to account for sample variability.

Figure 2 shows the accuracy for all models. We observe a huge performance improvement with Tran-KBLSTM over other baseline models for low data regimes. All improvements are statistically significant with Student's t-test $p < 0.05$ except dev results with 3% and 5%. For precise numbers and standard deviation plots, see Table refer Table 8 in the Appendix §D. Additionally, as the training supervision increases, the performance margin across models narrows. This improvement can be attributed to the fact that the model's reasoning ability increases when more training data is added, resulting in more accurate predictions without explicitly necessitating external knowledge addition. As a result, adding external knowledge may not be as beneficial if there is adequate supervision.

Reasoning Analysis To investigate the reasoning behind a model's prediction (i.e., RQ3(a)), INFOTABS adapted the set of reasoning categories from GLUE (Wang et al. (2018a)) for tabular premises. Thus, we also analyze performance across several reasoning types on the development set of INFOTABS . We utilized the reasoning annotated instances from INFOTABS for our analysis. Figure 3 show the performance across various reasoning types on the development set for 1% and 3% of INFOTABS development set. Trans-KBLSTM model shows improvements in several reasoning types including *"Lexical"*, *"Multi-Row"*, and *"KCS"*.

- *Lexical Reasoning* involves inferencing through words independent of context, where the word falls. Since we add relational connections between words which include synonyms, antonyms, etc. lexical reasoning ability of the model enhances. For example, in the table *"Chibuku Shake"*, the key *"Ingredients"* contains *"Sorghum"* and *"Maize"* while the hypothesis requires us to infer about *Corn* as an ingredient in the Chibuku shake. The relation {"corn" $\xleftarrow{Synonym}$ "Maize"} helps the model in making the correct prediction. For details refer to table 13 in Appendix §E.
- *Multi-Row Reasoning* involves making an inference using multiple rows of the table. When the reasoning involves multiple rows, the model needs to extract the relevant rows and rightly focus on selected related connected phrases. The relational connections that we propose between premise and hypothesis tokens establish these extractions and connections and thus enhancing the multi-row reasoning ability of the Trans-KBLSTM model. For example in a "Person" table relations such as { "born" $\xleftrightarrow{RelatedTo}$ "young" ; "born" $\xleftrightarrow{RelatedTo}$ "child" ; "child" $\xleftrightarrow{RelatedTo}$ "age" ; "year active" $\xleftrightarrow{Co-Hyponym}$ "child" } help in connected both the born, child and year active keys with the concern hypothesis. For details refer to table 12 in Appendix §E.
- *Knowledge and Common Sense Reasoning.* This reasoning is related to the World Knowledge and Common Sense category from GLUE-Benchmark (Wang et al., 2018b), which is quoted as "...the entailment rests not only on correct disambiguation of the sentences, but also, application of extra knowledge, whether factual knowledge about world affairs or more commonsense knowledge about word meanings or social or physical dynamics." Knowledge databases like ConceptNet contain many knowledge relations capable of enhancing these reasoning type. For example, in a "Country" table relations such

[5] reaches maximum in 6-7 epochs while Neeraja et al. (2021) takes 14-15 epochs [6] Higher % include all instances from lower %, i.e. a 20% includes all instances from a 10% samples.

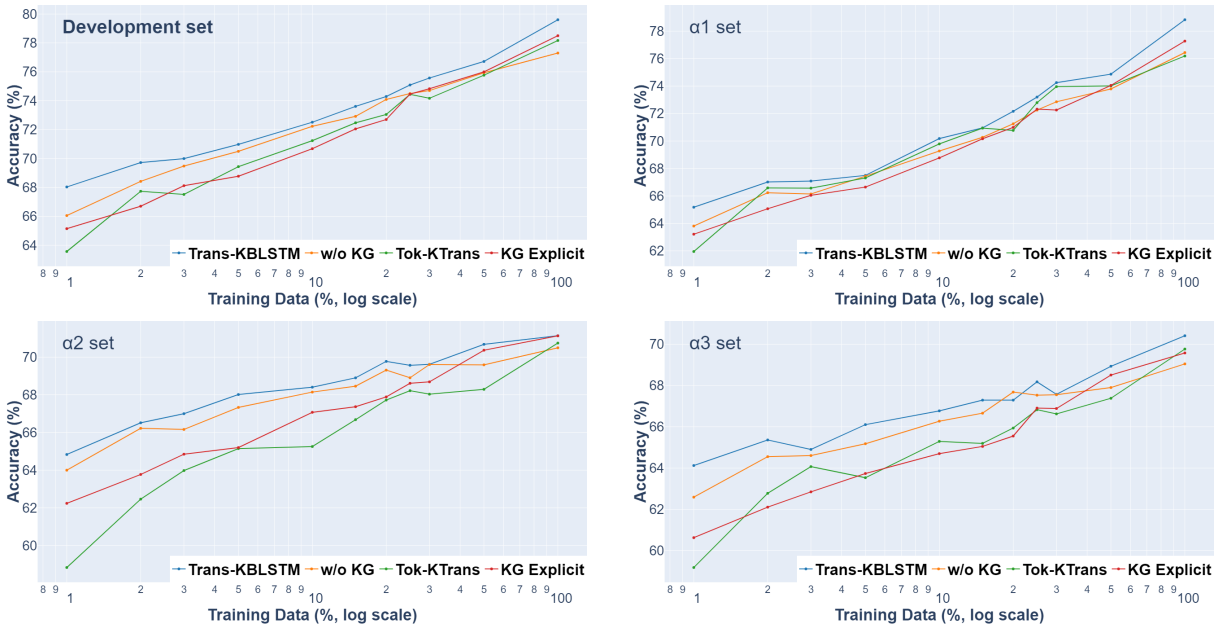

Figure 2: Performance in terms of accuracy in limited supervision setting. **w/o KG** represent RoBERTa INFOTABS (Gupta et al., 2020) baseline, **KG Explicit** represent Knowledge-INFOTABS (Neeraja et al., 2021) baseline, **Tok-KTrans** is the token appended transformers and **Trans-KBLSTM** represent our proposed model. Reported results are average over 3 random seed runs with average standard deviation of 0.233 (w/o KG), 0.49 (KG Explicit), 0.50 (Tok-KTrans) and 0.30 (Trans-KBLSTM). All the improvements are statistically significant with Student's t-test $p < 0.05$ of one-tailed Student t-test.

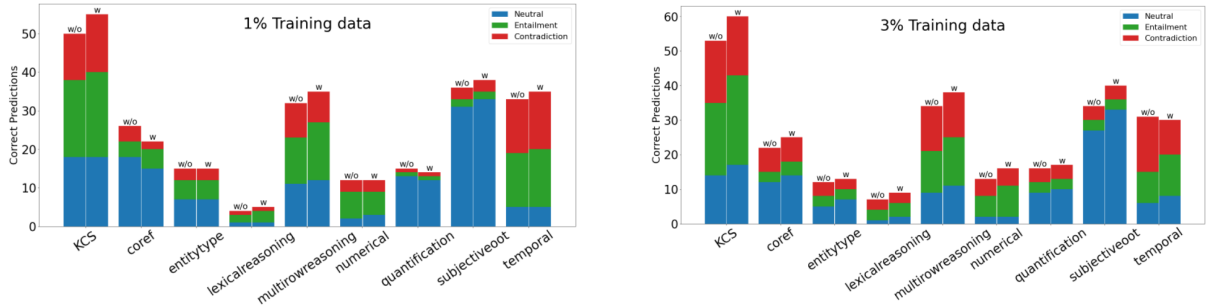

Figure 3: Number of correct model predictions across various reasoning types. **w/o** represents without knowledge (KG) i.e. original RoBERTa models and **w** represents Trans-KBLSTM model with explicitly added relational connection knowledge (KG).

as { "kingdom" $\xleftrightarrow{IsA}$ "monarchy" ; "democracy" $\xleftrightarrow{RelatedTo}$ "Government" } add additional information necessary for inference. For details refer to table 14 in Appendix §E.

Improvement across Inference Labels. In our analysis, we observe a performance improvement across the Entailment and Neutral labels, but only a negligible increase, for example, in instances labeled with the Contradiction label. Contradictory label prediction requires noise-free, contextually relevant knowledge to ascertain the negation. External knowledge addition with minimal noise can lead to the predicted Neutral or Entailment label. Additional ways for relational connection trimming may be explored in future studies.

3.3 Ablation Study

We perform ablation studies (i.e., RQ4) to understand the importance of individual model components further. The ablation study was conducted to ascertain the significance of (a) Trans-KBLSTM Skip Connection, (b) Knowledge Relations, (c) Implicit KG addition via. MNLI pre-training (Embeddings), and (d) Transformer Model Param Size. (e) Independent Component training.

Effect of Skip Connections. We study the significance of embedding skip connection and the knowledge relations (i.e., RQ4(a,b)). The knowledge relations are initialized with random vectors to examine model performance variations.

Table 3 shows the Trans-KBLSTM performance with several ablations. We observe that adding knowledge and the introduction of skip connection improve the model performance. The addition of knowledge to the model improves the performance on Dev, $\alpha 1$, and $\alpha 2$ sets. The inclusion of knowledge improves performance the most for De-

velopment, α_2 , and α_3 sets, whereas the addition of skip connection improves performance substantially in α_1 set. The performance improvement in α_3 set demonstrates that using external information benefits zero-shot settings (i.e., cross-domain transfer learning). The improved performance by the addition of skip connection demonstrates that effective knowledge integration significantly impacts model performance.

Ablations	Dev	α_1	α_2	α_3
Trans-KBLSTM	**67.55**	**65.16**	**64.00**	**63.38**
- Skip Connect	65.72	62.83	60.00	61.55
- KB	60.44	61.88	56.94	55.55
- (KB + Skip Connect)	60.11	61.50	55.94	57.38

Table 3: Ablation study performance on stratified 1% split of dataset. We systematically eliminate model components in order to evaluate the performance improvement.

Implicit Knowledge Addition. We examine the effect of implicit knowledge addition (i.e., RQ4(b)) in Trans-KGLSTM model. Thus, similar to the KG Implicit baseline of Knowledge-INFOTABS (Neeraja et al., 2021), we supplement implicit knowledge using the MNLI via data augmentation. To ensure a fair comparison, we compare the two Trans-KBLSTM RoBERTa-based classifiers, one with and the other without MNLI data pre-training.

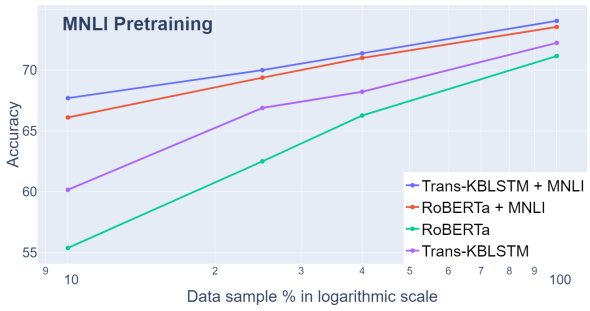

Figure 4: Performance improvement with MNLI pre-training across various models.

We observe an improvement in performance for all percentages of train data after pre-training using MNLI data. Pre-training enables the model to acquire domain-specific information, hence enhancing its performance. There is a more significant gain in performance for non-pre-trained than for MNLI pre-trained models, suggesting that external information addition is more beneficial for models without any implicit knowledge. In comparison, our approach uses relational connections to augment the model's knowledge in the phase, final training avoiding the computational, time, and economic cost of large MNLI pre-training.

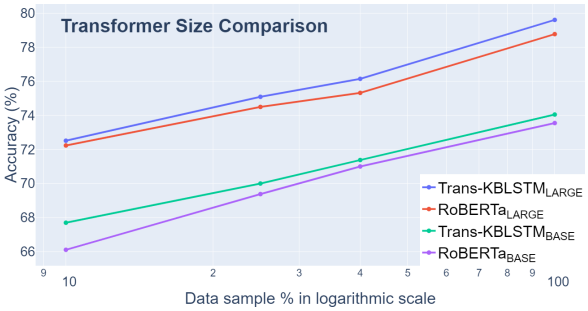

Figure 5: Improvement in model performance across varying models sizes.

Effect of Transformer Size. We substitute RoBERTa$_{LARGE}$ with RoBERTa$_{BASE}$ to study the effect of transformer size on performance (i.e. RQ4 (d)) of INFOTABS test sets. We pre-train both the transformers model using the MultiNLI dataset for all percentages. The performance is depicted in Figure 5. We see an increase in performance as the model's size increases, especially for external knowledge addition, i.e., Trans-KBLSTM model.

Independent Training. We examine the effect of training transformer and KBLSTM components independently. For independent training, we first train RoBERTa$_{LARGE}$ transformer model on INFOTABS . Then we utilize these weights to initialize the transformer component of Trans-KBLSTM. Finally, we trained the KBLSTM component of Trans-KBLSTM on INFOTABS while keeping these pre-trained transformer weights frozen (constant). Table 4 shows the results of training Trans-KBLSTM with different regimes. We observe that training the components together shows a more significant improvement in performance than training the KBLSTM component independently. Joint training of transformer and KBLSTM generates representations in the same embedding space, enhancing external knowledge integration.

Ablations	Dev	α_1	α_2	α_3
RoBERTa$_{LARGE}$	77.30	76.44	70.49	69.05
+ KBLSTM (Independent)	79.22	78.38	71.00	69.22
+ KBLSTM (Joint Train)	**79.92**	**79.62**	**72.10**	**70.21**

Table 4: Joint/Independent training performance on INFOTABS dataset. First row shows results of training only RoBERTa$_{LARGE}$ model without knowledge. Second row shows results of training KBLSTM independently after freezing RoBERTa$_{LARGE}$ parameters. Third row shows the results of our proposed approach i.e. Joint-training of RoBERTa$_{LARGE}$ and KBLSTM.

4 Comparison with Related Work

Recently, several papers have been published focusing on NLP tasks involving semi-structured Tabular data. Examples include tabular NLI (Gupta et al., 2020), and fact verification (Chen et al. (2020); Aly et al. (2021); Zhang and Balog (2019)). The use of external knowledge into Tabular data was first explored by Neeraja et al. (2021) through *KG-Explicit* model described in §3.1. We aim to improve on this benchmark through this extensive study.

Knowledge Integration. Traditional approaches to integrating external knowledge into deep learning models do not use contextual embeddings from pre-trained language models. The Knowledge-based Inference Model (KIM) (Chen et al., 2018) incorporates lexical relations (such as antonyms and synonyms) into the premise and hypothesis representations using attention and composition units. Lin et al. (2017) provides a method to mine and exploit commonsense knowledge by defining inference rules between elements under different kinds of commonsense relations, with an inference cost for each rule. KG-Augmented Entailment System (KES) (Kang et al., 2018) augments the NLI model with external knowledge encoded using graph convolutional networks. ConseqNet (Wang et al., 2019) concatenates the output of the text-based model and the graph-based model and then feeds it to a classifier. Lin et al. (2019) uses LSTMs and a novel knowledge-aware graph network module named KagNet to achieve state-of-the-art performance on CommonSenseQA. Bi-CAM (Gajbhiye et al., 2020) models incorporate knowledge from ConceptNet and AristoTuple KGs (Dalvi Mishra et al., 2017) by factorized bilinear pooling to improve performance on NLI Datasets.

Incorporating external knowledge into language models has been extensively explored in recent times. Approaches similar to the Tok-KTrans baseline described in §3.1 where external knowledge is added at input level were explored in Chen et al. (2021); Xu et al. (2021); Mitra et al. (2019). At the representational level, the model understands these external knowledge additions and interacts with these representations using multi-head attention modules (Chang et al., 2020). Other approaches include, pretraining on external knowledge corpus to inject knowledge (Wang et al., 2021; Peters et al., 2019; Umair and Ferraro, 2021), better knowledge representations (Bauer et al., 2021), modifications to multi-head attention in pre-trained language models (Li and Sethy, 2019; Haihong et al., 2019), designing relation-aware tasks (Xia et al., 2019) and integration of knowledge through multi-head attention (Gajbhiye et al., 2021).

Closely Related Work. Li et al. (2019) finds that when explicit knowledge is added in the form of word-pair information, models such as Chen et al. (2018) improve performance. However, such models necessitate the use of classic *seq2seq* architectures such as BiLSTM to integrate word-level knowledge. In our proposed approach, external knowledge is separately added to the premise and hypothesis using a multi-head attention dot product. To encode the contextual relationships between premise and hypothesis, we use a pre-trained language model, RoBERTa (Liu et al., 2019). We combine the LM embeddings (Gajbhiye et al., 2021) and BiLSTM embeddings using a skip connection which preserves the premise-hypothesis relational context and integrates knowledge effectively.

5 Conclusion and Future Work

In this paper, we introduce Trans-KBLSTM, a novel architecture to integrate external knowledge into tabular NLI models. Trans-KBLSTM is shown to improve reasoning on the INFOTABS dataset. The performance advantage is particularly pronounced in low-data regimes. The reasoning study demonstrates that the model enhances lexical, numerical, and multiple-row reasoning. Ablation experiments demonstrate the critical nature of each component in the model's design. We believe that our findings will be valuable to researchers working on the integration of external knowledge into deep learning architectures. Performance of the proposed architecture on more datasets can be explored in future studies. Looking forward, the application of this architecture to other NLP tasks that can benefit from external knowledge enhanced relational connections between sentence pairs, such as question answering and dialogue understanding.

Acknowledgement

We thank members of the Utah NLP group for their valuable insights and suggestions at various stages of the project; and reviewers their helpful comments. Additionally, we appreciate the inputs provided by Vivek Srikumar and Ellen Riloff. Vivek Gupta acknowledges support from Bloomberg's Data Science Ph.D. Fellowship.

References

Abien Fred Agarap. 2018. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. The fact extraction and VERification over unstructured and structured information (FEVEROUS) shared task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 1–13, Dominican Republic. Association for Computational Linguistics.

Lisa Bauer, Lingjia Deng, and Mohit Bansal. 2021. ERNIE-NLI: Analyzing the impact of domain-specific external knowledge on enhanced representations for NLI. In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 58–69, Online. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Ting-Yun Chang, Yang Liu, Karthik Gopalakrishnan, Behnam Hedayatnia, Pei Zhou, and Dilek Hakkani-Tur. 2020. Incorporating commonsense knowledge graph in pretrained models for social commonsense tasks. In *Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 74–79, Online. Association for Computational Linguistics.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Diana Inkpen, and Si Wei. 2018. Neural natural language inference models enhanced with external knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2406–2417, Melbourne, Australia. Association for Computational Linguistics.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. Tabfact: A large-scale dataset for table-based fact verification. In *International Conference on Learning Representations*.

Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2021. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. *arXiv preprint arXiv:2104.07650*.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. Recognizing Textual Entailment: Models and Applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–220.

Bhavana Dalvi Mishra, Niket Tandon, and Peter Clark. 2017. Domain-targeted, high precision knowledge extraction. *Transactions of the Association for Computational Linguistics*, 5:233–246.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159.

Amit Gajbhiye, Noura Al Moubayed, and Steven Bradley. 2021. Exbert: An external knowledge enhanced bert for natural language inference. In *Artificial Neural Networks and Machine Learning – ICANN 2021*, pages 460–472, Cham. Springer International Publishing.

Amit Gajbhiye, Thomas Winterbottom, Noura Al Moubayed, and Steven Bradley. 2020. Bilinear fusion of commonsense knowledge with attention-based nli models. In *International Conference on Artificial Neural Networks*, pages 633–646. Springer.

Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Vivek Gupta, Maitrey Mehta, Pegah Nokhiz, and Vivek Srikumar. 2020. INFOTABS: Inference on tables as semi-structured data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2309–2324, Online. Association for Computational Linguistics.

E Haihong, Wenjing Zhang, and Meina Song. 2019. Kb-transformer: Incorporating knowledge into end-to-end task-oriented dialog systems. In *2019 15th International Conference on Semantics, Knowledge and Grids (SKG)*, pages 44–48. IEEE.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Dongyeop Kang, Tushar Khot, Ashish Sabharwal, and Eduard Hovy. 2018. AdvEntuRe: Adversarial training for textual entailment with knowledge-guided examples. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2418–2428, Melbourne, Australia. Association for Computational Linguistics.

Alexander Hanbo Li and Abhinav Sethy. 2019. Knowledge enhanced attention for robust natural language inference. *CoRR*, abs/1909.00102.

Tianda Li, Xiaodan Zhu, Quan Liu, Qian Chen, Zhigang Chen, and Si Wei. 2019. Several experiments on investigating pretraining and knowledge-enhanced models for natural language inference. *arXiv preprint arXiv:1904.12104*.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. KagNet: Knowledge-aware graph networks for commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2829–2839, Hong Kong, China. Association for Computational Linguistics.

Hongyu Lin, Le Sun, and Xianpei Han. 2017. Reasoning with heterogeneous knowledge for commonsense machine comprehension. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2032–2043, Copenhagen, Denmark. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692. Version 1*.

George A. Miller. 1992. WordNet: A lexical database for English. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Arindam Mitra, Pratyay Banerjee, Kuntal Kumar Pal, Swaroop Mishra, and Chitta Baral. 2019. How additional knowledge can improve natural language commonsense question answering? *arXiv preprint arXiv:1909.08855*.

J. Neeraja, Vivek Gupta, and Vivek Srikumar. 2021. Incorporating external knowledge to enhance tabular reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2799–2809, Online. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. pages 43–54.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4444–4451. AAAI Press.

Mohammad Umair and Francis Ferraro. 2021. Transferring semantic knowledge into language encoders. *arXiv preprint arXiv:2110.07382*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018a. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*,

pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Xiaoyan Wang, Pavan Kapanipathi, Ryan Musa, Mo Yu, Kartik Talamadupula, Ibrahim Abdelaziz, Maria Chang, Achille Fokoue, Bassem Makni, Nicholas Mattei, and Michael Witbrock. 2019. Improving natural language inference using external knowledge in the science questions domain. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7208–7215.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021. Kepler: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jiangnan Xia, Chen Wu, and Ming Yan. 2019. Incorporating relation knowledge into commonsense reading comprehension with multi-task learning. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2393–2396.

Yichong Xu, Chenguang Zhu, Ruochen Xu, Yang Liu, Michael Zeng, and Xuedong Huang. 2021. Fusing context into knowledge graph for commonsense question answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1201–1207, Online. Association for Computational Linguistics.

Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. 2014. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*.

Shuo Zhang and Krisztian Balog. 2019. Autocompletion for data cells in relational tables. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, pages 761–770, New York, NY, USA. ACM.

A Hypothesis Attention Module

In Hypothesis attention module, we calculate hypothesis relation values by normalizing R_{ijk} with respect to row-axis(2), to generate $R^{hyp}_{ik} \in \mathbb{R}^{m \times k}$ which is the average hypothesis relation for every premise word.

$$R^{hyp}ik = \sum i = 1^n \frac{R_{ijk}}{n}$$

We reduce the dimension by applying the dot product attention.

$$R^r_{ik} = F^r_H(R^{hyp}_{ik}) \in \mathbb{R}^{m \times l_k}$$

F^r_N can again be a single layer neural network. We then use the Hypothesis attention head to highlight the importance of the hypothesis and its relations to the premise. The context-aware premise hidden state p^s is used as queries, the hypothesis hidden state is used as keys, and reduced hypothesis premise relation values are used. The attention function can be defined as follows:

$$\text{Attention}(p^s, h^s, R^r_{ik}) = \text{softmax}(\frac{p^s h^{sT}}{\sqrt{l}})R^r_{ik}$$

Then the multi-head attention is as follows:

$$\begin{aligned} p^{att}_h &= \text{MH}(p^s, h^s, R^r_{ik}) \\ &= \text{Concat}(\text{head}_1, ..., \text{head}_h)W^o \end{aligned}$$

where, head$i$ = Attention$(p^s W^q_i, h^s W^k_i, R^r ikW^v_i)$ and W^q_i, W^k_i, and W^v_i are projection matrices and i is the number of attention heads. The output $p^{att}_h \in \mathbb{R}^{m \times l_k}$ is an attention-weighted context matrix measuring the importance of premise and relations to each of the hypothesis. We calculate $p^{att}_h \in \mathbb{R}^{m \times l_k}$, attention-weighted context matrix measuring the importance of premise and relations to each of the hypothesis. We also extract H^{att}, the attention weights of the hypothesis multi-head attention.

B Knowledge Relations to Sentence Conversion

We create templates to convert knowledge relations in ConceptNet & WordNet to natural language sentences. These templates resemble natural English text, which can be encoded using pretrained language models. The templates can be seen in table 5.

KB Relation	Natural Language
Antonym	is opposite of
AtLocation	is at location
CapableOf	is capable of
Causes	causes
CausesDesire	causes desire to
CreatedBy	is created by
DefinedAs	is defined as
DerivedFrom	is derived from
Desires	desires
DistinctFrom	is distinct from
Entails	entailes
EtymologicallyDerivedFrom	is etymologically derived from
EtymologicallyRelatedTo	is etymologically related to
ExternalURL	external url
FormOf	is a form of
HasA	has a
HasContext	has context
HasFirstSubevent	has first subevent
HasLastSubevent	has last subevent
HasPrerequisite	has prerequisite
HasProperty	has property
HasSubevent	has subevent
InstanceOf	is an instance of
IsA	is a
LocatedNear	is located near
MadeOf	is made of
MannerOf	is manner of
MotivatedByGoal	is motivated by goal
NotCapableOf	is not capable of
NotDesires	does not desire
NotHasProperty	does not have property
PartOf	is part of
ReceivesAction	receives action
RelatedTo	is related to
SimilarTo	is similar to
SymbolOf	is a symbol of
Synonym	is same as
UsedFor	is used for
dbpedia/capital	has capital
dbpedia/field	has field
dbpedia/genre	has genre
dbpedia/genus	has genus
dbpedia/influencedBy	is influenced by
dbpedia/knownFor	is known for
dbpedia/language	has language
dbpedia/leader	has leader
dbpedia/occupation	‘ has occupation
dbpedia/product	has product
Hypernym	is hypernym of
Hyponym	is hyponym of
Co-Hyponym	is co-hyponym of

Table 5: ConceptNet and Wordnet Relations with their Natural language templates

C Domain Analysis

To understand the models performance across tabular domains (i.e. RQ3(b)), we analyse domain-wise table category results. We evaluate the twelve major categories contained in the INFOTABS datasets. All remaining categories are grouped together in the *"Other"* category. Table summarizes the performance of models (trained with 2% and 5% INFOTABS train data)[7] on the INFOTABS development set across several categories.

Category	1%		3%		10%	
	w/o KG	w KG	w/o KG	w KG	w/o KG	w KG
Album	65.87	65.87	73.81	**76.98**	77.78	73.02
Animal	60.49	**66.67**	75.31	66.67	67.9	**72.84**
City	64.05	**64.71**	56.21	**61.44**	63.4	**64.71**
Country	56.48	54.63	56.48	55.56	60.19	**62.96**
Food & Drinks	69.44	**70.83**	72.22	**73.61**	83.33	79.17
Movie	61.11	**63.89**	63.89	63.89	70	**73.89**
Musician	62.57	**69.88**	73.1	**74.56**	75.73	**76.9**
Organization	61.11	58.33	55.56	**66.67**	69.44	**72.22**
Painting	80.25	80.25	75.31	**77.78**	77.78	**80.25**
Person	57	**62.96**	62.35	**67.28**	74.9	**75.72**
Sports	65.08	**73.02**	61.9	**71.43**	68.25	**69.84**
Others	63.89	**65.28**	66.67	**70.84**	63.89	61.11
TOTAL	62	**65.83**	65.88	**68.61**	72.27	**73.22**

Table 6: Accuracy (%) across different categories observed in the Development set (Others (<10%) includes the categories, University, Awards, Event, Book and Aircraft), trained on 1%, 3% and 5% samples of the data. **w/o KG** represents RoBERTa and **w KG** represents Trans-KBLSTM model.

As the supervision increases from 1% to 10%, we observe an increasing accurate prediction trend across the categories. Our proposed model shows significant improvements in *"Musician"* and *"Sports"* categories. We attribute these huge gains to two main reasons: (a) . Under minimal supervision, knowledge relations enable the model to concentrate on relevant context, thus helping in establishing premise rows and hypothesis tokens connections. For example refer to table 10 in Appendix §E. (b) and the acquisition of additional knowledge enhances the models' overall world knowledge and common sense reasoning capability. E.g. in the table 1, the understanding of the *California* is located at the *coast*.

Additionally, we observe that our proposed model performs poorly in a few categories. This part comprises instances from *"Album"*, *"Food & Drinks"*, and *"University"*. This can be attributed to the noisy addition of knowledge. Sometimes knowledge relations give out the relational context that might not be needed. For example refer to table 11 in Appendix §E. Additional knowledge filtering may be addressed in future studies. For domain analysis results of models trained on 2% and 5% training data, refer to table 7.

D Limited Supervision

We present detailed results on limited supervision experiments. All the reported numbers are aver-

[7] For details results on other percentages refer to Appendix §C Table 7.

Category	2%		5%	
	w/o KG	w KG	w/o KG	w KG
Album	68.25	67.46	72.22	**73.81**
Animal	65.43	64.20	72.84	69.14
City	55.56	**58.17**	60.13	**61.44**
Country	58.33	**62.96**	61.11	**68.52**
Food&Drinks	69.44	66.67	75.00	73.61
Movie	58.33	**65.00**	65.56	65.56
Musician	68.42	**71.64**	71.35	**76.32**
Organization	58.33	**61.11**	66.67	66.67
Painting	66.67	59.26	75.31	**76.54**
Person	61.32	60.49	68.72	67.08
Sports	66.67	**69.84**	61.90	**68.25**
Others	62.50	**66.67**	63.89	**65.28**
TOTAL	63.11	**64.44**	68.22	**69.50**

Table 7: Accuracy (%) across different categories observed in the Development set (Others (<10%) includes the categories, University, Awards, Event, Book and Aircraft), trained on 2% and 5% samples of the data. **w/o KG** represents RoBERTa baseline and **w KG** represents Trans-KBLSTM

age over three seed runs with a standard deviation of 0.233 (w/o KG), 0.49 (KG Explicit), 0.5 (Tok-KTrans), and 0.30 (Trans-KBLSTM). All the improvements are statistically significant with $p <$ 0.05 of one-tailed Student t-test.

E Qualitative Examples

Table 10, 11, 12, 13, and 14 present examples to supplement the results presents in Section 3.

F Additional Results Reasoning Analysis

Table 15 detailed results of performance across reasoning keys for models trained on 1%, 3%, 5% and 10% data.

G Training and Hyperparameters Details

Trans-KBLSTM is implemented in PyTorch (Paszke et al., 2019) using Huggingface (Wolf et al., 2020) implementation of RoBERTa (Liu et al., 2019). We pretrain the transformer components on MultiNLI dataset (Williams et al., 2018) for fair comparison with the Knowledge-INFOTABS baseline of Neeraja et al. (2021). We use AdaGrad optimizer (Duchi et al., 2011) with an initial learning rate of 1e-4 for RoBERTa and 1e-3 for non-RoBERTa i.e. LSTM parameters with a scheduler. The batch size is slected from {3,4, 5}. All the hyper-parameters are fined tuned on the development set of INFOTABS .

% Train	Model	Dev	$\alpha 1$	$\alpha 2$	$\alpha 3$
1%	w/o KG	66.05	63.81	64.00	62.59
	KG Explicit	65.15	63.22	62.24	60.63
	Tok-KTrans	63.57	61.96	58.83	59.18
	Trans-KBLSTM	**68.03**	**65.18**	**64.83**	**64.12**
2%	w/o KG	68.42	66.24	66.22	64.55
	KG Explicit	66.70	65.07	63.77	62.11
	Tok-KTrans	67.74	66.59	62.46	62.78
	Trans-KBLSTM	**69.72**	**67.02**	**66.51**	**65.36**
3%	w/o KG	69.48	66.14	66.16	64.61
	KG Explicit	68.12	66.05	64.85	62.85
	Tok-KTrans	67.52	66.57	63.98	64.07
	Trans-KBLSTM	**70.00**	**67.09**	**67.00**	**64.90**
5%	w/o KG	70.50	67.44	67.33	65.18
	KG Explicit	68.78	66.65	65.20	63.74
	Tok-KTrans	69.44	67.31	65.14	63.53
	Trans-KBLSTM	**70.98**	**67.50**	**68.01**	**66.11**
10%	w/o KG	72.23	69.27	68.14	66.27
	KG Explicit	70.68	68.77	67.07	64.70
	Tok-KTrans	71.24	69.79	65.25	65.29
	Trans-KBLSTM	**72.51**	**70.18**	**68.40**	**66.77**
15%	w/o KG	72.92	70.27	68.46	66.66
	KG Explicit	72.05	70.16	67.37	65.05
	Tok-KTrans	72.47	70.94	66.68	65.20
	Trans-KBLSTM	**73.61**	**70.96**	**68.90**	**67.29**
20%	w/o KG	74.09	71.25	69.31	67.68
	KG Explicit	72.70	70.99	67.89	65.55
	Tok-KTrans	73.05	70.77	67.72	65.94
	Trans-KBLSTM	**74.29**	**72.16**	**69.77**	67.29
25%	w/o KG	74.50	72.25	68.90	67.53
	KG Explicit	74.46	72.32	68.61	66.91
	Tok-KTrans	74.44	72.79	68.22	66.83
	Trans-KBLSTM	**75.09**	**73.20**	**69.57**	**68.18**
30%	w/o KG	74.70	72.86	69.61	67.55
	KG Explicit	74.83	72.26	68.69	66.89
	Tok-KTrans	74.17	73.96	68.03	66.63
	Trans-KBLSTM	**75.57**	**74.25**	**69.62**	**67.57**
50%	w/o KG	75.93	73.79	69.59	67.90
	KG Explicit	75.99	74.05	70.36	68.51
	Tok-KTrans	78.44	76.38	70.66	70.38
	Trans-KBLSTM	**76.71**	**74.86**	**70.68**	**68.93**
100%	w/o KG	77.30	76.44	70.49	69.05
	KG Explicit	78.97	77.84	**71.13**	69.58
	Tok-KTrans	78.17	76.19	70.75	69.77
	Trans-KBLSTM	**79.73**	**78.92**	**71.62**	**70.21**

Table 8: Shows the results of of our experiments, where we train under limited supervision setting. **w/o KG** Original RoBERTa baseline, **KG Explicit** KG-Explicit knowledge addition, **Tok-KTrans** Token appended transformers, **Trans-KBLSTM** Proposed model. We train these models on data samples 1, 2, 3, 5, 10, 15, 20, 25, 30, 50, 100 %s. For full results, see appendix.

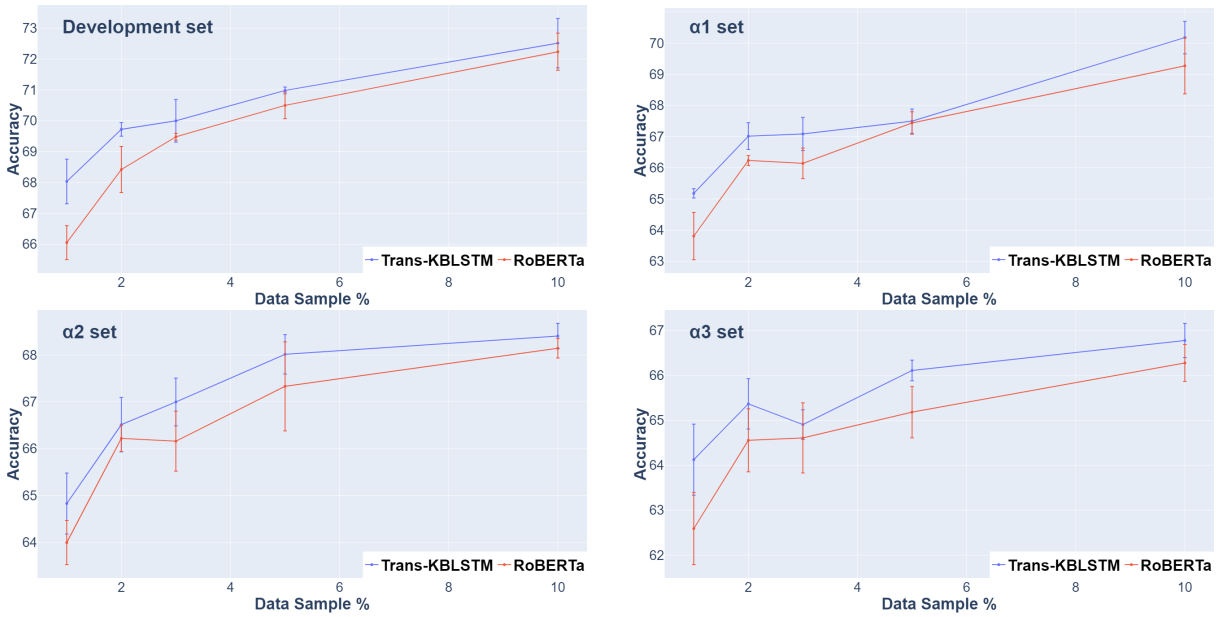

Figure 6: The figures show error bar plots of limited supervision training on 1,2,3,5,10 and 15% of data. for Trans-KBLSTM and RoBERTa baseline. We notice that the error overlap increases with increase in supervision. The improvements are higher under low-data regimes.

Hyperparameter	Value
LSTM Max Length	200
LSTM layers	2
LSTM learning rate	1e-3
LSTM Hidden state size	128
Word Embedding Dimension	300
RoBERTa Hidden state size	768
RoBERTa learning rate	1e-4
# Attention heads	4
Embedding Spatial Dropout	0.3
Dropout (Final classification)	0.2

Table 9: Enlists the hyperparameters used while training the baselines and proposed model on INFOTABS

Joe Budden Premise	
Premise	Joe Budden was Born on (1980-08-31) August 31, 1980 (age 38) in New York, New York. The Origin of Joe Budden are Jersey City, New Jersey. The Years active of Joe Budden are 1999-present. The Labels of Joe Budden are Mood Muzik, EMPIRE (current), Desert Storm, Def Jam, Amalgam Digital, and E1 (former)
Hypothesis	Joe Budden started his career in his twenties.
Focused Relation	age $\xleftrightarrow{Co-Hyponym}$ twenties
Gold Label	Contradiction
Prediction	
RoBERTa	Neutral
Trans-KBLSTM	Contradiction

Table 10: In the absence of knowledge, the model is unable to understand the word *twenties* and concludes that the information is not present in the text. However, addition of knowledge re-enforces the connection between *age* and *twenties* thereby producing correct label

Crooked Teeth Premise	
Premise	The Released of Crooked Teeth are May 19, 2017. The Studio of Crooked Teeth are Steakhouse Studios, North Hollywood, CA. The Genre of Crooked Teeth are Hard rock, nu metal, and rap rock. The Label of Crooked Teeth are Eleven Seven.
Hypothesis	The album Crooked Teeth took over a year to make.
Focused Relation	genre $\xleftrightarrow{Co-Hyponym}$ make || metal $\xrightarrow{RelatedTo}$ make || rap $\xrightarrow{Hypernym}$ make
Gold Label	Neutral
Prediction	
RoBERTa	Neutral
Trans-KBLSTM	Contradiction

Table 11: The baseline prediction correctly predicts the gold label. Our proposed model gets confused with semantically irrelevant relations and hence concludes the statement as contradiction.

Jeff Bridges Premise	
Premise	The Born of Jeff Bridges are December 4, 1949 (age 69) Los Angeles, California, U.S.. The Years active of Jeff Bridges are 1951-present. The Children of Jeff Bridges are 3. The Family of Jeff Bridges are Beau Bridges (brother), and Jordan Bridges (nephew).
Hypothesis	Jeff Bridges started his career as a young child.
Focused Relations	born $\xrightarrow{RelatedTo}$ young born $\xrightarrow{RelatedTo}$ child child $\xrightarrow{RelatedTo}$ age active $\xrightarrow{Co-Hyponym}$ child
Gold Label	Entailment
Prediction	
RoBERTa	Contradiction
Trans-KBLSTM	Entailment

Table 12: The inference of the hypothesis requires the model to focus on 1^{st} and 2^{nd} sentences at the same time. The original model gets confused due to mention of *age 69* and *young* and concludes contradiction. The focused relations develop appropriate connections to the first two sentences and enable better understanding to the model.

Chibuku Shake Premise	
Premise	The Type of Chibuku Shake shake are Opaque Beer. The Alcohol by volume of Chibuku Shake shake are 3.3% to 4.5%. The Colour of Chibuku Shake shake are Tan-pink to white. The Ingredients of Chibuku Shake shake are Sorghum, and Maize.
Hypothesis	Corn is an ingredient found in a Chibuku Shake.
Focused Relations	corn $\xleftrightarrow{Synonym}$ maize
Gold Label	Entailment
Prediction	
RoBERTa	Entailment
Trans-KBLSTM	Entailment

Table 13: The inference of the given hypothesis requires the knowledge of Synonymy between *Corn* and *Maize*

Hashemite Kingdom of Jordan Premise	
Premise	The Legislature of Hashemite Kingdom of Jordan are Parliament. The Religion of Hashemite Kingdom of Jordan are 95% Islam (official), 4% Christianity, and 1% Druze, Baha'i. The Government of Hashemite Kingdom of Jordan are Unitary parliamentary constitutional monarchy. The Monarch of Hashemite Kingdom of Jordan is Abdullah II.
Hypothesis	Hashemite Kingdom of Jordan does not have any democracy.
Focused Relation	Kingdom $\xleftrightarrow{IsA}$ Monarch
Gold Label	Contradiction
Prediction	
RoBERTa	Neutral
Trans-KBLSTM	Contradiction

Table 14: The focused external knowledge relation connects the *Monarchy* in premise to *Kingdom* in hypothesis.

Percent(%)	Reasoning keys	Entailment			Neutral			Contradiction		
		B.L	KtLSTM	.	B.L	KtLSTM	.	B.L	KtLSTM	.
1%	KCS	64.52	**70.97**	31	85.71	85.71	21	50.00	**62.50**	24
	coref	50.00	**62.50**	8	81.82	68.18	22	30.77	15.38	13
	entitytype	83.33	83.33	6	87.50	87.50	8	50.00	50.00	6
	lexicalreasoning	40.00	**60.00**	5	33.33	33.33	3	25.00	25.00	4
	multirowreasoning	60.00	**75.00**	20	68.75	**75.00**	16	52.94	47.06	17
	nameidentity	0.00	0.00	2	0.00	**100.00**	2	100.00	100.00	1
	negation	0.00	0.00	0	0.00	0.00	0	66.67	**83.33**	6
	numerical	63.64	54.55	11	66.67	**100.00**	3	42.86	42.86	7
	quantification	25.00	25.00	4	100.00	92.31	13	16.67	16.67	6
	subjectiveoot	33.33	33.33	6	75.61	**80.49**	41	50.00	50.00	6
	temporal	73.68	**78.95**	19	45.45	45.45	11	56.00	**60.00**	25
3%	KCS	67.74	**83.87**	31	66.67	**80.95**	21	75.00	70.83	24
	coref	37.50	**50.00**	8	54.55	**63.64**	22	53.85	53.85	13
	entitytype	50.00	50.00	6	62.50	**87.50**	8	66.67	50.00	6
	lexicalreasoning	60.00	**80.00**	5	33.33	**66.67**	3	75.00	75.00	4
	multirowreasoning	60.00	**70.00**	20	56.25	**68.75**	16	76.47	76.47	17
	nameidentity	50.00	**100.00**	2	100.00	100.00	2	100.00	100.00	1
	negation	0.00	0.00	0	0.00	0.00	0	100.00	100.00	6
	numerical	54.55	**81.82**	11	66.67	66.67	3	71.43	71.43	7
	quantification	75.00	75.00	4	69.23	**76.92**	13	66.67	66.67	6
	subjectiveoot	50.00	50.00	6	65.85	**80.49**	41	66.67	66.67	6
	temporal	47.37	**63.16**	19	54.55	**72.73**	11	64.00	40.00	25
5%	KCS	87.10	83.87	31	71.43	**90.48**	21	66.67	62.50	24
	coref	75.00	62.50	8	68.18	**81.82**	22	30.77	30.77	13
	entitytype	83.33	83.33	6	87.50	87.50	8	83.33	83.33	6
	lexicalreasoning	60.00	**80.00**	5	33.33	**66.67**	3	50.00	50.00	4
	multirowreasoning	85.00	85.00	20	68.75	**81.25**	16	58.82	**76.47**	17
	nameidentity	100.00	100.00	2	50.00	**100.00**	2	100.00	0.00	1
	negation	0.00	0.00	0	0.00	0.00	0	100.00	66.67	6
	numerical	72.73	**90.91**	11	100.00	100.00	3	71.43	**85.71**	7
	quantification	75.00	50.00	4	92.31	**100.00**	13	33.33	16.67	6
	subjectiveoot	66.67	33.33	6	73.17	**87.80**	41	50.00	50.00	6
	temporal	94.74	84.21	19	36.36	**63.64**	11	56.00	52.00	25
10%	KCS	74.19	**80.65**	31	95.24	90.48	21	70.83	70.83	24
	coref	50.00	**75.00**	8	77.27	77.27	22	46.15	23.08	13
	entitytype	66.67	**83.33**	6	87.50	87.50	8	100.00	83.33	6
	lexicalreasoning	80.00	80.00	5	66.67	66.67	3	25.00	**75.00**	4
	multirowreasoning	80.00	80.00	20	81.25	81.25	16	76.47	70.59	17
	nameidentity	50.00	50.00	2	100.00	100.00	2	100.00	100.00	1
	negation	0.00	0.00	0	0.00	0.00	0	83.33	**100.00**	6
	numerical	81.82	**100.00**	11	100.00	100.00	3	71.43	71.43	7
	quantification	50.00	50.00	4	84.62	**92.31**	13	33.33	33.33	6
	subjectiveoot	33.33	**50.00**	6	82.93	**87.80**	41	50.00	33.33	6
	temporal	78.95	**89.47**	19	63.64	63.64	11	68.00	68.00	25

Table 15: The above numbers represent accuracy on development dataset across different reasoning types with varying percentage of data. The third number indicates the number of examples corresponding to the reasoning type and label.