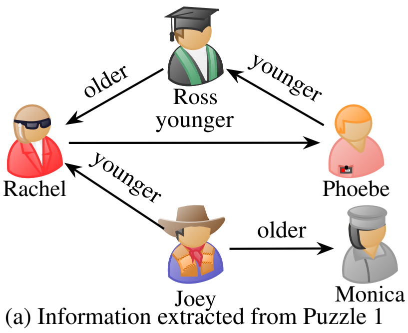
A Corpus for Precise Natural Language Inference

We propose a corpus for tasks related to natural language inference. The corpus contains logical puzzles in natural language from two domains: comparing puzzles and truth-telling puzzles. For instance:

Example 1 (Comparison puzzle) *Ross is older than Rachel who is younger than Phoebe. Joey is older than Monica but younger than Rachel. Phoebe is younger than Ross. Who is the tallest? Is Phoebe older than Monica? Is Ross younger than Joey? (Figure 1)*

First, note that the text in the queries does not explicitly appear within the text of the puzzle. Since reasoning is mandatory to answer, approaches based only on machine learning may not suffice. Solutions based on reasoning or hybrid approaches will be required by this puzzle-based corpus. Second, note that some background knowledge is required: *older* and *younger* are transitive relations, or the definition of concept *tallest*. Third, good puzzles have two properties: (i) each piece of information is necessary and (ii) no unnecessary information is provided. These properties make puzzles interesting candidates for machine comprehension tasks. Fourth, since the solution of the puzzle is clear, one can void the cases of mislabelling the text due to subjective annotation or human biases. Recall the many troublesome annotations with the SICK dataset identified by Kalouli et al. [Kalouli et al.2017], in which 611 pairs out 9,840 does not make sense. Fifth, there is wide range of logical puzzles with available solutions[1] that can be collected and adapted for building such puzzle based corpus, starting with simple ones (e.g. comparison puzzles) and continuing with more complex ones (e.g. zebra puzzles). Sixth, the existing work on automatic puzzle generation combined with the work on natural language generation can be used to automatically create such puzzle-based benchmarks for precise inference tasks. One example are the 382 knights and knaves puzzles popularised by Raymond Smullyan and automatically generated by Lau and Chan[2]. The complexity of these puzzles depends on the number of the individuals ranging from 2 to 9 individuals. The above aspects can be discussed during the workshop to clarify the best way for delivering a puzzle-based benchmark for question answering.



(a) Information extracted from Puzzle 1

Question	Theorem in FOL	Answer
Is Ross the tallest ?	$tallest(Ross)$	Entailment
Is Monica the shortest ?	$shortest(Monica)$	Entailment
Is Phoebe older than Monica ?	$older(Phoebe, Monica)$	Entailment
Is Monica younger than Joey ?	$younger(Monica, Joey)$	Entailment
Is Rachel the tallest ?	$tallest(Rachel)$	Contradiction
Is Phoebe the shortest ?	$shortest(Phoebe)$	Contradiction
Is Monica older than Phoebe ?	$older(Monica, Phoebe)$	Contradiction
Is Ross younger than Joey?	$younger(Ross, Joey)$	Contradiction

(b) Sample of atomic question with 4 predicates and 5 individuals

Figure 1: Question answering for unambiguous comparison puzzles

Automatically solving such text-based puzzles requires several technical challenges. Consider the following puzzle with two friends:

[1] Consider for instance many sites like https://www.brainzilla.com/logic/zebra/, https://www.ahapuzzles.com/logic/logic-puzzles, https://www.mathsisfun.com/puzzles/logic-puzzles-index.html to name a few

[2] https://philosophy.hku.hk/think/logic/knights.php

Figure 2: A truth-telling puzzle with two characters

Example 2 (Truth telling puzzle) *In the Central Perk cafe there is this particular behaviour: married people don't lie, while single people always lie. While Joey and Chandler were sitting on the sofa a woman is approaching them and asked: "Are you married?" Joey promptly replied: "We are both single!" Can the woman figure out weather the two friends are married or not? (Figure 2)*

When translating the puzzle into some logical formalism (e.g. First Order Logic [Groza and Nitu2022] using NLTK [Perkins2014] and Prover9 [McCune2005] or Description Logics using for instance FRED [Gangemi et al.2017]), a machine comprehension tool should handle various technical challenges including: (i) recognising the named entities (e.g. Joey, Chandler); (ii) coreference resolution (e.g. "We are both single"); (iii) automatically computing the domain size for model finders (e.g. Mace4), (iv) reducing the interpretation models to a single one (e.g. adding the unique name assumption, adding relevant background knowledge, closing the world, removing isomorphic models).

For each puzzle one can generate a large set of questions, as exemplified in the right part of Figure 1. Each puzzle can be associated with the entire set of atomic questions that can be generated based on the relations and individuals occurring in the text [Szomiu and Groza2021]. If the puzzle has a unique solution, there should be only pairs labelled as "entailment" and "contradiction". To obtain unknown pairs, one can remove some clues from the puzzle. The resulted ambiguous puzzle will have several interpretation models, in which the statements can be proved as true (entailment), false (contradiction), or unknown (if they appear true or false in the computed models.

In line with the work of Pease et al. [Pease et al.2020], we propose a corpus for reasoning tasks. The puzzle-based corpus may benefit from the huge number of puzzle available that provide unique solutions. Apart from the comparison and truth-telling puzzles exemplified here, the corpus can be extended with various types of logical puzzles [Groza2021].

References

Aldo Gangemi, Valentina Presutti, Diego Reforgiato Recupero, Andrea Giovanni Nuzzolese, Francesco Draicchio, and Misael Mongiovì. 2017. Semantic web machine reading with FRED. *Semantic Web*, 8(6):873–893.

Adrian Groza and Cristian Nitu. 2022. Question answering over logic puzzles using theorem proving. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, pages 871–874.

Adrian Groza. 2021. *Modelling Puzzles in First Order Logic*. Springer International Publishing.

Aikaterini-Lida Kalouli, Livy Real, and Valeria De Paiva. 2017. Textual inference: getting logic from humans. In *IWCS 2017-12th International Conference on Computational Semantics—Short papers.*

William McCune. 2005. Prover9. *University of New México.*

Adam Pease, Paulo Santos, and Alexandre Rademaker. 2020. A corpus of spatial reasoning problems. In *5th Conf. on Artificial Intelligence and Theorem Proving, September 13-19, 2020, Aussois, France.*

Jacob Perkins. 2014. *Python 3 text processing with NLTK 3 cookbook*. Packt Publishing Ltd.

Roxana Szomiu and Adrian Groza. 2021. A puzzle-based dataset for natural language inference. *arXiv preprint arXiv:2112.05742.*