First paper “A large annotated corpus for learning natural language inference”

Consider the notion of inference that is taken here when collecting the data for the dataset. How does this compare to the notion of inference that is used in the FraCaS dataset?

FraCaS:

- Contains different entailment problems covering multiple semantic phenomena (such as references, quantifiers, etc.)
- Is built to be used as a performance measure for inference systems that use logical relations between premises and hypotheses
- Was built to test systems that rely on logical entailment
- Is written to avoid word sense ambiguity
- Is small
- Is limited in usefulness for the performance evaluation of semantic processing systems on real – world language data

NLI systems that utilize data-driven methods

- need a large dataset
- It contains real-world language - the relationship between pair of sentences is not strictly logical

Why image scene descriptions are used; and why the images are not shown to the describers?

To ensure that the relationship between the produced premises and the provided hypothesis (image caption) is based solely on the available text, not on some assumptions on events and entity co-reference that could be induced by the image.

The paper tests the collected dataset on existing NLI systems. From description and a discussion of results, what do you think are there strengths and weakness of each modelling inference?

- Both neural-based and lexicalized based systems have achieved the same accuracy, while the unlexicalized system achieves the worst.

- The neural-based system, unlike both lexicalized and lexicalized based systems, can learn from unstructured representations of sentence meaning.

- To be able to achieve this performance NLI systems need a larger dataset.

- Yet the performance of all systems is limited to their ability to handle syntactic complexity and semantic problems like word sense disambiguation and compositionality representation.

Points to consider for discussion

Since the annotators are humans, there might be a human bias due to their cultural backgrounds, personal motives, belief about the world and so on. Bias usually affects when we design a model and also when we evaluate a model.