An overview of Natural Language Inference Data Collection

NLI is considered by many researchers to be the crux of computational semantics. This paper is about the datasets created for this. In particularly, we discuss the most common NLI resources : 1) fail to capture the wealth of inferential mechanisms present in NLI, 2) seem to be driven by the dominant discourse in the field at the time of their creation. NLI platform must satisfy both in terms of the range of inference patterns found in reasoning with NL and the range of the data collection mechanisms needed to acquire this range of inferential patterns.

1. Merits and drawbacks of NLI datasets collection of inference

1.1 The FraCaS test suite

The FraCaS test suite was built in the mid 1990's by the FraCaS project, an EU project aimed at developing a general framework for computational semantics. The data set consists of 346 problems each containing one or more statements and one yes/no-question. All the types of inferences are annotated.

Merits: even though it contains only 346 examples, it covers a lot of inference cases and it now has the advantage of some multilinguality environments.

Drawbacks: the examples are mostly logical inferences, the data are constructed and the dataset is very small especially with today's standards.

1.2 Recognizing Textual Entailment

The Recognizing Textual Entailment (RTE) challenges were first developed in 2004 as a means to test textual entailment, i.e. relations between a premise text and a hypothesis text. It has tripartite classifications -- hypothesis entailed, negation entailed or no entailment.

Merits: The RTE uses natural text and furthermore make a first step of including entailments that require presupposed information, including non-logical presuppositional inferences.

Drawbacks: RTE platforms have been difficult for NLI systems due to the three entailment classifications. A amount of world knowledge needs to be taken into consideration which is beyond the linguist abilities, however, there is no clear annotation that will distinguish different kinds of inference. RTE datasets are still very small (less than 1000 pairs) compared with approaches that relying on large datasets such as Deep Learning approaches (DL)

1.3 Stanford Natural Language Inference

The subjects are given a caption of a picture and are then asked to provide: 1) an alternate true caption 2) an alternate possibly true caption and 3) an alternate false caption. The dataset constructed out of this process contains 570k inference pairs, making SNLI two orders of magnitude bigger than the previous datasets.

Merits: The SNLI platform is extremely useful, and given the current state of affairs in computational linguistics, it is the only one usable for approaches using DL, which are predominant at the moment.

Drawbacks: Situational reasoning can be also claimed to be a drawback of SNLI. and similarly to earlier platforms, SNLI seems to capture only a fraction of the range of phenomena associated with NLI.

1.4 Some other NLI platforms

The SICK dataset (Sentences Involving Compositional Knowledge) is a dataset created to test compositional distributional semantics (DS) models. The PPDB (Paraphrase Database) relation extraction dataset is primarily a dataset on paraphrase. The VQA (Visual Question Answering) corpus6 contains open-ended questions (and answers) about images.

2 What do we need?

Presuppositional inference and non-logical inference as given by implicatures and enthymemes is important. Presuppositional inference is to some extent included implicitly in the FraCaS test suite but it is treated as logical inference implicitly. The data set should be grounded in real data which includes reasoning in interactive dialogue settings and not just text , so that it does not just represent armchair intuitions of linguistic experts as the FraCaS test suite. Connecting language and vision recently and exploring inference in this context are important.

3 How to get it?

(i) finding corresponding examples in corpora consisting of text from different genres or the web; (ii) by conducting experiments in which subjects are asked to evaluate or construct inferences using both textual and visual context. In the latter case we think it is useful (iii) to use crowd-sourcing and (iv) to adapt the game techniques used in GWAPs.

4 How we know we got it?

(i) to verify whether intuitions of inference of ordinary speakers correspond to the intuitions of the authors of the dataset; and (ii) to evaluate a degree of variation of inference intuitions for individual examples which would allow us to examine how humans reason and therefore further study examples of inference that are normally considered as problematic. (iii) evaluating with more examples, more resources with different backgrounds and with different languages.