Commonsense Causal Reasoning Using Million of Personal Stories *Andrew S. Gordon, Cosmin Adrian Bejan, and Kenji Sagae*

Summary:

Commonsense reasoning aims to help computers understand and interact with people more naturally by finding ways to collect these assumptions and teach them to computers. The formal approaches lack to store commonsense knowledge with sufficient inferential soundness and development of effective data-driven methods for automated commonsense reasoning remains an open challenge. There are many research efforts presented by acquiring general world knowledge from text corpora based on parsing sentences and mapping syntactic forms into logic forms then inferring them with simple propositional facts from these forms through abstraction. This kind of approach is proven useful for some tasks but it is still not clear that how this knowledge can be applied toward benchmark problems in commonsense reasoning. In this paper, authors describe approaches that utilize unstructured web text on a massive scale as commonsense knowledge. This work specifically looks at the unique properties of the genre of the personal story, for which millions of examples are readily available in weblogs. Authors believe that personal stories from weblogs are ideally suited as a source of commonsense causal information, in that causality is a central component of coherent narrative and weblogs are focused on everyday situations. This paper is mainly concerned with the following question:

- Are personal stories from weblogs a better source of causal information than text corpora?

- Can we overcome the challenges of open domain commonsense causal reasoning by exploiting the unique properties of personal stories?

The main obstacle on the way to the solution is the scalability of this approach. To tackle these questions authors conducted four experiments (1) choice of plausible alternatives (2) Personal stories as a knowledge base (3) Reasoning with discourse relations (4) Reasoning with sentence proximity from these authors concluded that the personal stories that people write in their weblogs are a good source of commonsense causal information. The strong performance of PMI-based techniques with a moderate word window size (W=25) suggests that this causal information exists largely within the scope of adjacent clauses and sentences, which is consistent with analyses by discourse psychologists on other narrative genres. To overcome the obstacles of scalability author conducted these experiments with millions of stories in which authors developed a pipeline for automatically identifying personal stories in weblogs. This involved supervised machine learning techniques used to identify personal stories in a stream of English language weblog posts. The results showed that one million stories are probably enough for the best results. This paper shows that using PMI statistics from a corpus of personal stories is an appropriate step in this direction.