Towards Grounding Commonsense Concepts in Commonsense Knowledge Bases

June 2020

1 Abstract

One promising solution to the current commonsense challenges is incorporating external commonsense knowledge in the form of knowledge bases. Several large-scale commonsense knowledge bases have been developed, such as ConceptNet [5], ASER [8], and ATOMIC [6]. However, limited success has been observed when trying to utilize them as additional knowledge source to solve commonsense tasks. Several factors may contribute to this issue. The most noted one is the coverage problem, in that it is hard to expect a knowledge base to contain all the knowledge needed to answer a broad range of commonsense knowledge. In this work, however, we investigate a rather less studied problem, which we phrase as "grounding commonsense concepts in commonsense knowledge bases". Commonsense concepts can be expressed in various surface forms. For example, the concept of "X defeats Y" can be expressed as "X wins over Y" or "X is the winner". It is hard to use simple methods to locate the concept of "X defeats Y" when it is expressed in a complex natural language sentence. This differs traditionally studied problems such as information retrieval (IR) or entity linking, where the focus is finding entity-centric terms in knowledge sources, such as "President Obama", and heuristic-based methods such as string match would yield decent performance.

In this work, we investigate the commonsense concept grounding problem. We take ASER and ATOMIC as example, and through human evaluation, confirm the severity of this issue. We also propose potential solutions to the problem and show promising results on solving a commonsense task SocialIQA [7] using ATOMIC.

2 Human Evaluation

2.1 ASER

ASER is a large collection of relations between events and is promised to provide good coverage of everyday events. It also provides a parsing-based way

to locate the relevant events. We investigate if this approach can successfully locate relevant knowledge given a problem in Winograd. Unsurprisingly, we find it challenging to locate useful and relevant events in ASER using the current pipeline. In the following, we conduct error analysis to illustrate reasons contributing the issue in detail.

- Parser failure. In ASER, each event consists of a center verb and its several arguments. We used a reasonably good parser [1] to conduct dependency parse on the sentences. However, we find that the parser makes a fair amount of mistakes, oftentimes confusing arguments of verbs. This impacts the quality of the events we extract from raw sentences and subsequently the events we are able to match in the ASER.

- Limited event patterns. When creating ASER, authors extract events from raw text following fixed syntactic patterns. This approach limits the kind of events we could extract from text. In addition, certain details (e.g. adjectives) are ignored when we keep only the skeleton of the events while in certain examples, the details are critical for solving the question.

- Simplified linking process. After extracting events (verbs and arguments), ASER defaults to strict string matching between verbs and fuzzy string matching between arguments to locate relevant concepts. This simple linking process only gives us lexically related events. For more than 50% of the events we extract from Winograd, we are unable to locate even one related event in ASER.

- Event quality. Being automatically collected, the quality of the events stored in ASER are lacking. Based on the subjective judgement of the author, more than 30% of the located events are of low quality and thus do not provide much useful information.

2.2 ATOMIC

ATOMIC is a recent large-scale knowledge base focusing on inferential relations between social events and emotions. We take twenty questions out of the Winograd Challenge and try to search in ATOMIC for knowledge potentially useful for solving the question. During searching, we can modify the surface form of the query without changing its semantic meaning. We find that 12 out of the 20 questions could benefit from knowledge from ATOMIC. However, the surface form problem we find still exists. In Table 1, we showcase three examples with varying difficulties of locating the relevant knowledge. In the top example, surface clues including word match would lead us to the correct knowledge. In the second example, it becomes much harder to arrive at the correct knowledge based on heuristics. In the final example, there is almost no surface string match between the query and the knowledge, thus a heuristic-based information retrieval (IR) system is almost bound to fail on it.

Winograd questions	ATOMIC Knowledge
The trophy doesn't fit into the brown suitcase because **it** is too large.	PersonX wouldn't fit. PersonX is seen as large.
Joan made sure to thank Susan for all the help **she** had received.	PersonX offers help. As a Result, others want to thank PersonX for it.
Frank felt crushed when his longtime rival Bill revealed that **he** was the winner of the competition.	PersonX defeats PersonY's purpose. As a Result, others feel humiliated/defeated/sad.

Table 1: Examples from Winograd which could benefit from ATOMIC Knowledge

3 Preliminary Solution

Having identified the problem, we present a seemingly promising solution and present our preliminary experimental results. As revealed, locating relevant commonsense concepts in raw text goes well beyond surface form matching and current IR techniques are ill-equipped for this task. It requires deep contextual understanding of the concept. Taking inspiration from recent development in open-domain QA [4, 2, 3], we propose a trainable neural commonsense knowledge retriever to locate commonsense concepts in raw text. The neural retriever benefits from pre-trained contextual language models and is equipped to deal with the surface form variation of commonsense concepts. We then use the presumably more accurately retrieved commonsense knowledge to enhance a standard question answering system.

3.1 Method

Suppose we are given a commonsense question answer pair $< q, a >$, and a commonsense knowledge base C, which is a collection of commonsense knowledge pieces c. Our goal is to find the most relevant knowledge c^* which could help answering q. This setting resembles the setting of open-domain QA and we use a recently proposed method to solve this problem. The retriever is implemented as two separate encoders E_q and E_c. E_q encodes the query q into a dense vector q while E_c encodes the knowledge pieces c into a collection of vectors c. Then finding the most relevant knowledge is transformed into a nearest neighbor search on C given q. Then we feed the retrieved knowledge c' along with q into a question answering system. Notably, E_q and E_c are initialized from pre-trained language models, and thus possess the power to project semantically similar sentences into the same vector space.

During training, the retriever does not require annotations for which knowledge c is the desired knowledge for a specific query q. Rather, the model can automatically learn to discard irrelevant knowledge and locate useful knowledge

through pairs of $< q, a >$. We refer the readers to [3] for the detailed learning process and ways to combat the cold-start problem.

3.2 Preliminary Results

We conduct preliminary experiments on SocialIQA, a commonsense multi-choice task targeted at social events. We choose ATOMIC as the external knowledge source. SocialIQA is partially derived from ATOMIC, where the annotators were asked to created natural language question and answers give a piece of knowledge from ATOMIC. Thus the experiment is under limited settings, in that the knowledge from ATOMIC is known to be useful for the task. However, the questions and answers have undergone substantial rewriting and serve as a suitable testbed for testing if the proposed model can perform semantic retrieval.

We now introduce the specification of the proposed model. E_q and E_c are initialized from the checkpoint from [3], providing non-trivial retrieval results at the beginning. During training, due to the computational cost issue as noted in , we freeze E_q and fine-tune E_c and the question answering system. We also include a baseline which receives no additional knowledge.

Results are shown in Table 2. Our system outperforms the baseline by a small margin. We note the several strategies in [4, 2] might further help with the results.

Model	Baseline	Our
Performance	57.5	58.0

Table 2: Performance of a baseline system and a system that benefits from external knowledge.

4 Team Members

- Honghua Zhang joshuacnf@ucla.edu 004644097
- Shanxiu He heshanxiu@g.ucla.edu 405173550
- Liunian Harold Li liunian.harold.li@cs.ucla.edu 005271406

References

[1] Matt Gardner et al. "Allennlp: A deep semantic natural language processing platform". In: *arXiv preprint arXiv:1803.07640* (2018).

[2] Kelvin Guu et al. "Realm: Retrieval-augmented language model pre-training". In: *arXiv preprint arXiv:2002.08909* (2020).

[3] Vladimir Karpukhin et al. “Dense Passage Retrieval for Open-Domain Question Answering”. In: *arXiv preprint arXiv:2004.04906* (2020).

[4] Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. “Latent retrieval for weakly supervised open domain question answering”. In: *arXiv preprint arXiv:1906.00300* (2019).

[5] Hugo Liu and Push Singh. “ConceptNet—a practical commonsense reasoning tool-kit”. In: *BT technology journal* 22.4 (2004), pp. 211–226.

[6] Maarten Sap et al. “Atomic: An atlas of machine commonsense for if-then reasoning”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 2019, pp. 3027–3035.

[7] Maarten Sap et al. “Socialiqa: Commonsense reasoning about social interactions”. In: *arXiv preprint arXiv:1904.09728* (2019).

[8] Hongming Zhang et al. “ASER: A large-scale eventuality knowledge graph”. In: *Proceedings of The Web Conference 2020*. 2020, pp. 201–211.