TheWebConf 2020 notification for paper 216

# Intro部分可以强调构建数据集的方法对中英文都是通用的

Dear Mengxue Zhang,  
  
We regret to inform you that your submission noted below has not been accepted to The Web Conference 2020:  
  
216: CoCon: A Dataset of Commonsense Contradiction on Short Phrase Constructed from the Web  
  
We received many submissions (1129 full papers and 397 short papers) and could accept only 19% of the full papers and 25% of the short papers.  
  
The program committee worked diligently to review all submitted papers and highlight areas for improvement. All papers were reviewed by at least three program committee members. In most cases, there are more than three reviews and/or meta-reviews – plus discussions among reviewers. The rebuttals from the authors were also seriously considered. Additionally, a track-chairs meeting was held in Amsterdam on Jan 6-7, 2020. Track chairs participated both on-site and remotely to discuss papers and ensure intertrack consistency.  
  
The reviews are included below. We trust that they are helpful for improving your paper.  
  
We hope that you can attend the conference and participate in other ways, such as workshops.  
  
Best regards,  
  
Tie-Yan Liu and Maarten van Steen  
TheWebConf2020 Program Chairs  
  
==========  
  
SUBMISSION: 216  
TITLE: CoCon: A Dataset of Commonsense Contradiction on Short Phrase Constructed from the Web  
  
**-------------------------  METAREVIEW  ------------------------**  
The reviewers agree that the dataset is a useful resource for the community, but identified a range of shortcomings with the current version of the work, which would need to be considered before publication. In particular, the reviewers asked for clarifications on the scope of the work, including fundamental concepts like commonsense, commonsense contradiction and commonsense reasoning (R1, R3). It was unclear, as noted by R3 where the boundaries are between the task the authors state they tackle and language modelling evaluation. R2 also commented on the relevance of the problem.  
  
Further on, all reviewers asked for specific details on the experimental design. R2 voices a concern about the performance of the approach, while R1 makes several useful points on the quality of the gold standard and missing details on the annotation methodology.  
  
Unfortunately we were not provided with a rebuttal to address any of these issues.  
  
  
  
**----------------------- REVIEW 1 ---------------------**  
SUBMISSION: 216  
TITLE: CoCon: A Dataset of Commonsense Contradiction on Short Phrase Constructed from the Web  
AUTHORS: Mengxue Zhang, Xusheng Luo, Yonghua Yang, Keping Yang, Zhiyi Luo, Qi Jia and Kenny Zhu  
  
----------- Strengths -----------  
Interesting topic.  
----------- Weaknesses -----------  
Poorly conducted research. Questionable definition of "common sense".  
----------- Relevance -----------  
SCORE: 2 (Yes, but only to a small group of people)  
----------- Summary and review comments -----------  
The paper presents research aiming to tackle common sense reasoning, previously mostly tested on long context-rich phrases, at the level of short, context-poor, phrases using a sort of Knowledge Graph. This is indeed an important problem because of a tendency of user-generated content to get shorter and shorter. The authors present their efforts to create a common sense short phrase data set, with some nice ideas indeed, but within an overall poorly conducted research.  
  
I am really surprised by some affirmations presented in the introduction. For instance "The performance of machines even surpasses humans on some tasks such as machine reading comprehension, image classification, and question answering." I am not an expert in the filed of image classification but it is difficult to imagine a non-human generated gold standard for image classification, and therefore it is difficult to understand how can machines surpass humans on such a task. The reference [4] given to support this surprising claim does not at all include any comparison of human vs. machine performance on image classification tasks.  
  
Other, similarly questionable assertions e.g. "statistical methods such as word co-occurrence are not that effective when dealing with short texts." lack supporting references.  
  
The authors claim that the phrase "watch Harry Potter" is more ambiguous to machines than it is to humans. They assume that common sense implies "watch" to be the action of watching a movie. However, clearly Harry Potter theme watches exist on the market (e.g. <https://www.harrypottershop.com/collections/hp-shop-watches> ) and the search query in question is clearly equally ambiguous to humans and machines. The authors should really make a better effort at defining what is "commonsense" for them and provide better examples where such common sense can be observed making a true difference.  
  
Due to such imprecise affirmations, the paper is really far from ready for publishing at a conference of this level.  
  
In the related work the authors should also discuss the work presented in the Making Sense of Microposts workshop and challenge over several years, where questions related to multiple short-text processing tasks and the use of Knowledge Graphs have been studied.【<http://microposts2016.seas.upenn.edu/topics.html>】【这个workshop主要关注NER，以及2015年开始关注NER后的entity linking】  
  
In 3.2. "some noisy and meaningless queries are filtered" - the authors should provide details. What are "meaningless queries"? How is the filtering exactly performed?  
  
In 3.3. "we first randomly select 10,000 samples from preprocessed raw data and do the human annotation." How is the human annotation performed? How many humans? What did they do? How is the coherence of annotations achieved between annotators? 【只有一个人标注，所以仿佛算不了这些东西，在文中加了一个词“rough”】How many annotators worked on each sample? What exactly is a "negative sample" and a "positive sample" - how are they defined? Give examples. In 3.3.1. one gets an impression that "moisturizing sunscreen" is a negative sample and "cleansing sunscreen" a positive one. Why?  
  
In 3.4. what is the ratio of kept samples with regards to the discarded ones? It would be important to also present the average inter-rater agreement between 5 raters in order to establish that the task was properly defined and that high level of consensus is actually emerging. The fact that only samples reaching high-agreement are kept is not sufficient【这仿佛需要当时的标注统计结果啊……如果没有的话仿佛也无能为力了】 - if the raters diverge, it would discredit the whole process of gold standard creation. The authors should present more convincing statistics related to the annotators work.  
----------- Overall score -----------  
SCORE: -2 (reject)  
----------- Reviewer's confidence -----------  
SCORE: 3 ((medium))  
  
  
  
**----------------------- REVIEW 2 ---------------------**  
SUBMISSION: 216  
TITLE: CoCon: A Dataset of Commonsense Contradiction on Short Phrase Constructed from the Web  
AUTHORS: Mengxue Zhang, Xusheng Luo, Yonghua Yang, Keping Yang, Zhiyi Luo, Qi Jia and Kenny Zhu  
  
----------- Strengths -----------  
S1. The paper is well written and organized and the proposed solution is clearly described  
  
S2. Construction of new dataset (CoCon) of commonsense short phrases in Chinese for evaluating the capability of commonsense reasoning for machines on short phrases.  
  
S3. Description of the methodology for dataset construction, from data collection to validation.  
----------- Weaknesses -----------  
W1. Applicability. It is not clear if the problem of commonsense reasoning with short phrases is a problem that has relevant application implications (in the Related Work section, the authors state that only a minimal percentage of queries on a search engine is associated with commonsense modifiers). Furthermore, the proposed dataset is for the Chinese language. Is it easy/feasibile to extend the dataset to other languages? Is the methodology of constructing the dataset easily reproducible?【适用性问题】【在搜索引擎上！commonsense modifier只占很小的比例，啊是说搜狗那个？？去掉去掉】【中英文的问题】  
  
W2. Results. The authors demonstrate that the use of the CoCon dataset to fine-tune standard models for NLP allows an increase in performance that brings existing models to over 70% accuracy on common sense reasoning tasks. However the results are far from human performances (95%). Authors should add further comments on how these results can be improved and what CoCon may lack to set up a complete and useful training set to improve performance.  
  
W3. Experimental section. It would be interesting that the authors include in their paper a more detailed analysis of BERT and ERNIE errors in the resolution of the common sense contradiction task. Which one of the False Positive or the False Negative numbers is greater? In particular, authors could analyze the set of False Positives more accurately to understand if some pairs of texts present in CoCon introduce noise in the training process.  
----------- Relevance -----------  
SCORE: 3 (Yes, to a large group of people)  
----------- Summary and review comments -----------  
The paper addresses the commonsense reasoning problem and it provides a dataset (CoCon) of short commonsense phrases in Chinese collected from search queries over a popular e-commerce platform for learning and reasoning purposes. The data collection process  and the dataset validation are described. Experiments are conducted using two state-of-the-art publicly available models for Chinese. Benchmark results shows that there is large margin for machines to achieve capability for judging commonsense contradiction in short texts and that external knowledge needs to be integrated into models.  
----------- Overall score -----------  
SCORE: 1 (weak accept)  
----------- Reviewer's confidence -----------  
SCORE: 3 ((medium))  
  
  
  
**----------------------- REVIEW 3 ---------------------**  
SUBMISSION: 216  
TITLE: CoCon: A Dataset of Commonsense Contradiction on Short Phrase Constructed from the Web  
AUTHORS: Mengxue Zhang, Xusheng Luo, Yonghua Yang, Keping Yang, Zhiyi Luo, Qi Jia and Kenny Zhu  
  
----------- Strengths -----------  
- The focus of this work is on creating a dataset for evaluating models on commonsense reasoning  
- Dataset is created from search queries, addressing a gap in the literature  
- Authors show a significant gap between the human performance and baselines  
----------- Weaknesses -----------  
- It is not entirely clear whether the introduced task is more relevant to commonsense reasoning or evaluating language modeling 【是说投稿方向不对？】【】  
- While the evaluation uses state-of-the-art models, the baselines are not necessarily effective because of the lack of context.  
----------- Relevance -----------  
SCORE: 3 (Yes, to a large group of people)  
----------- Summary and review comments -----------  
The paper introduces a dataset and a task for predicting which of two short phrases is more reasonable, with the aim to evaluate different models at their ability to detect common sense contradictions. While there are other datasets that aim to assess the ability of models to incorporate some aspect of common sense reasoning, the proposed one is constructed from user queries. This a novel aspect, however, it is not clear to what extend this dataset validates common sense reasoning or how more likely one short phrase is compared to another. To address this issue, it would be required to apply to this dataset models that do make use of common sense reasoning and compare them against models that do not make use of such knowledge.  
  
More specific comments follow below:  
In the introduction, the claim that "humans can easily understand the query intent ..." would require a citation. It seems that while a person could perceive the query based on what is more likely, the user could have a different intent, which could be inferred from clicked documents. If such information is not taken into account, then, we can only reasonably predict what is the most likely interpretation of the query. However, other intents or interpretations can exist that do not go against common sense. 【感觉是不是还是得在intro的第一段好好定义一下common sense】  
  
In Section 5.2, it is not clear what is meant by "Those words that have obvious statistical feature..."  
  
  
The paper has some typos:  
- In Sec. 2: "... made up of instances one premise... ", needs "of" before "one"  
- In Sec. 3.3: "... far away being ..." => "... far away from being ..."  
- In Sec. 3.4: "... by 5 person, ..." => "... by 5 persons, ..."  
- In Sec. 5.5, "However, we remain..." => "However, we keep...""  
----------- Overall score -----------  
SCORE: -2 (reject)  
----------- Reviewer's confidence -----------  
SCORE: 4 ((high))