Testing Your Question Answering Software via Asking Recursively

2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE)

ACM SIGSOFT
Distinguished
Paper Award

Presenter: Zhu Jie

2022.3.17

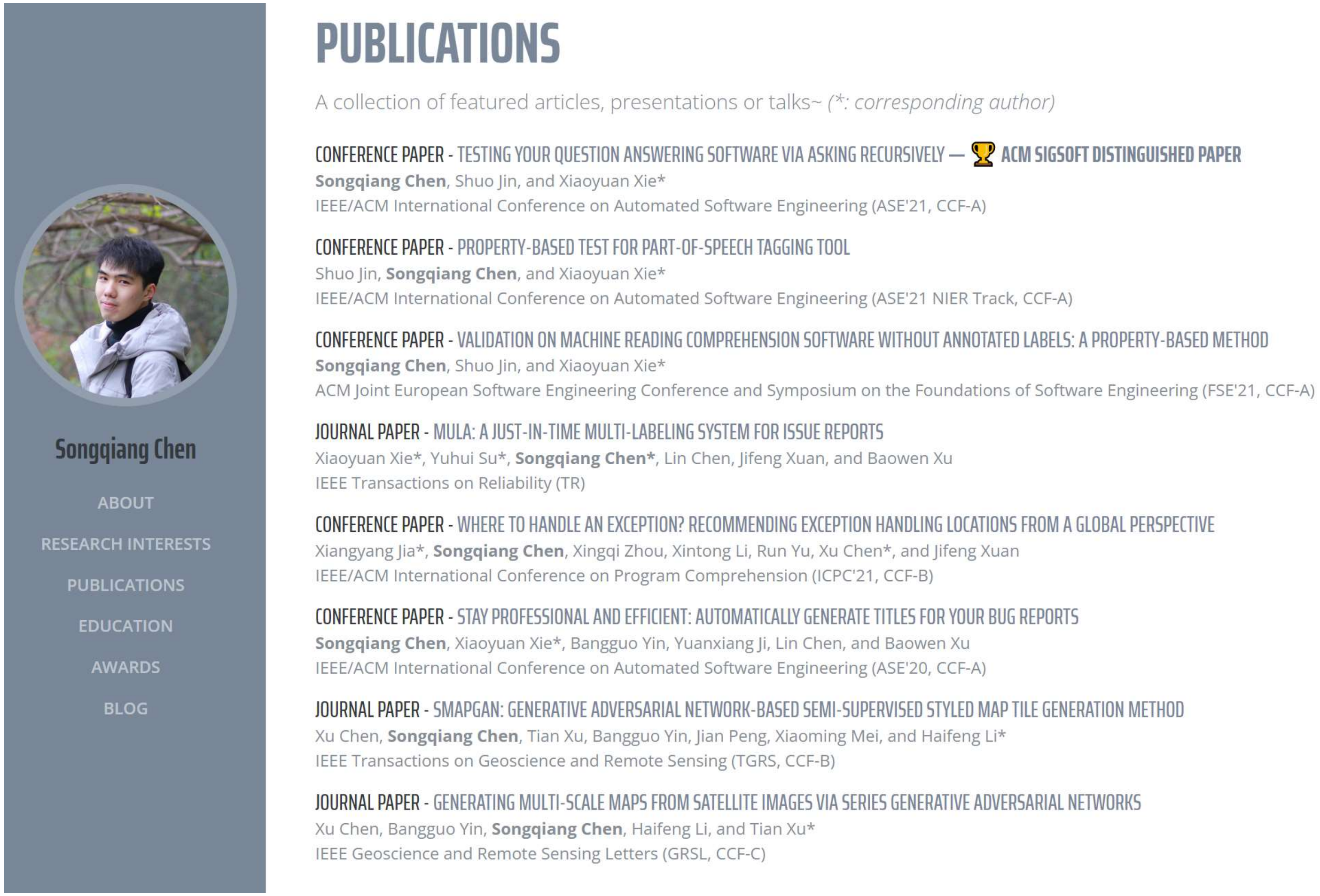
Authors

Songqiang Chen

Master Student (2020 - Now)

Wuhan University

- *Intelligent Software Engineering*
- *Software Testing*
- *Natural Language Processing*



PUBLICATIONS

A collection of featured articles, presentations or talks~ *(*: corresponding author)*

CONFERENCE PAPER - TESTING YOUR QUESTION ANSWERING SOFTWARE VIA ASKING RECURSIVELY — 🏆 **ACM SIGSOFT DISTINGUISHED PAPER**
Songqiang Chen, Shuo Jin, and Xiaoyuan Xie*
IEEE/ACM International Conference on Automated Software Engineering (ASE'21, CCF-A)

CONFERENCE PAPER - PROPERTY-BASED TEST FOR PART-OF-SPEECH TAGGING TOOL
Shuo Jin, **Songqiang Chen**, and Xiaoyuan Xie*
IEEE/ACM International Conference on Automated Software Engineering (ASE'21 NIER Track, CCF-A)

CONFERENCE PAPER - VALIDATION ON MACHINE READING COMPREHENSION SOFTWARE WITHOUT ANNOTATED LABELS: A PROPERTY-BASED METHOD
Songqiang Chen, Shuo Jin, and Xiaoyuan Xie*
ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (FSE'21, CCF-A)

JOURNAL PAPER - MULA: A JUST-IN-TIME MULTI-LABELING SYSTEM FOR ISSUE REPORTS
Xiaoyuan Xie*, Yuhui Su*, **Songqiang Chen***, Lin Chen, Jifeng Xuan, and Baowen Xu
IEEE Transactions on Reliability (TR)

CONFERENCE PAPER - WHERE TO HANDLE AN EXCEPTION? RECOMMENDING EXCEPTION HANDLING LOCATIONS FROM A GLOBAL PERSPECTIVE
Xiangyang Jia*, **Songqiang Chen**, Xingqi Zhou, Xintong Li, Run Yu, Xu Chen*, and Jifeng Xuan
IEEE/ACM International Conference on Program Comprehension (ICPC'21, CCF-B)

CONFERENCE PAPER - STAY PROFESSIONAL AND EFFICIENT: AUTOMATICALLY GENERATE TITLES FOR YOUR BUG REPORTS
Songqiang Chen, Xiaoyuan Xie*, Bangguo Yin, Yuanxiang Ji, Lin Chen, and Baowen Xu
IEEE/ACM International Conference on Automated Software Engineering (ASE'20, CCF-A)

JOURNAL PAPER - SMAPGAN: GENERATIVE ADVERSARIAL NETWORK-BASED SEMI-SUPERVISED STYLED MAP TILE GENERATION METHOD
Xu Chen, **Songqiang Chen**, Tian Xu, Bangguo Yin, Jian Peng, Xiaoming Mei, and Haifeng Li*
IEEE Transactions on Geoscience and Remote Sensing (TGRS, CCF-B)

JOURNAL PAPER - GENERATING MULTI-SCALE MAPS FROM SATELLITE IMAGES VIA SERIES GENERATIVE ADVERSARIAL NETWORKS
Xu Chen, Bangguo Yin, **Songqiang Chen**, Haifeng Li, and Tian Xu*
IEEE Geoscience and Remote Sensing Letters (GRSL, CCF-C)

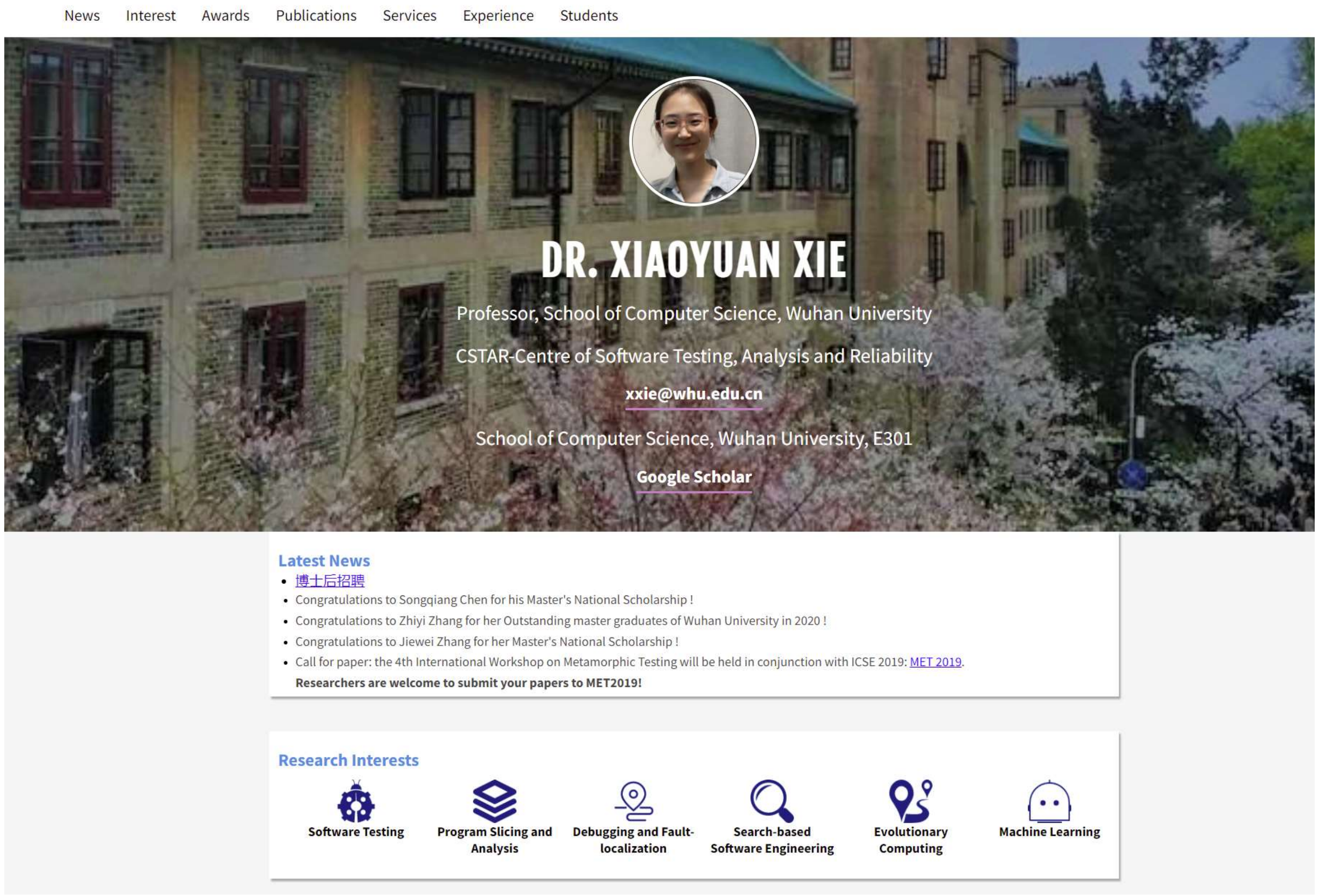
https://imcsq.github.io/

Authors

Xiaoyuan Xie

Professor
Leader of CSTAR Group
Wuhan University

- *Software Testing(Metamorphic testing and Mutation testing)*
- *Program Slicing and Analysis*
- *Debugging and Fault Localization*
- *Search-based Software Engineering*
- *Machine Learning*



News Interest Awards Publications Services Experience Students

DR. XIAOYUAN XIE

Professor, School of Computer Science, Wuhan University

CSTAR-Centre of Software Testing, Analysis and Reliability

xxie@whu.edu.cn

School of Computer Science, Wuhan University, E301

Google Scholar

Latest News

- 博士后招聘
- Congratulations to Songqiang Chen for his Master's National Scholarship !
- Congratulations to Zhiyi Zhang for her Outstanding master graduates of Wuhan University in 2020 !
- Congratulations to Jiewei Zhang for her Master's National Scholarship !
- Call for paper: the 4th International Workshop on Metamorphic Testing will be held in conjunction with ICSE 2019: MET 2019.
  Researchers are welcome to submit your papers to MET2019!

Research Interests

Software Testing
Program Slicing and Analysis
Debugging and Fault-localization
Search-based Software Engineering
Evolutionary Computing
Machine Learning

http://xiaoyuanxie.github.io/

Research Team

CSTAR

Center of Software Testing, Analysis, and Reliability

Computer School of Wuhan University

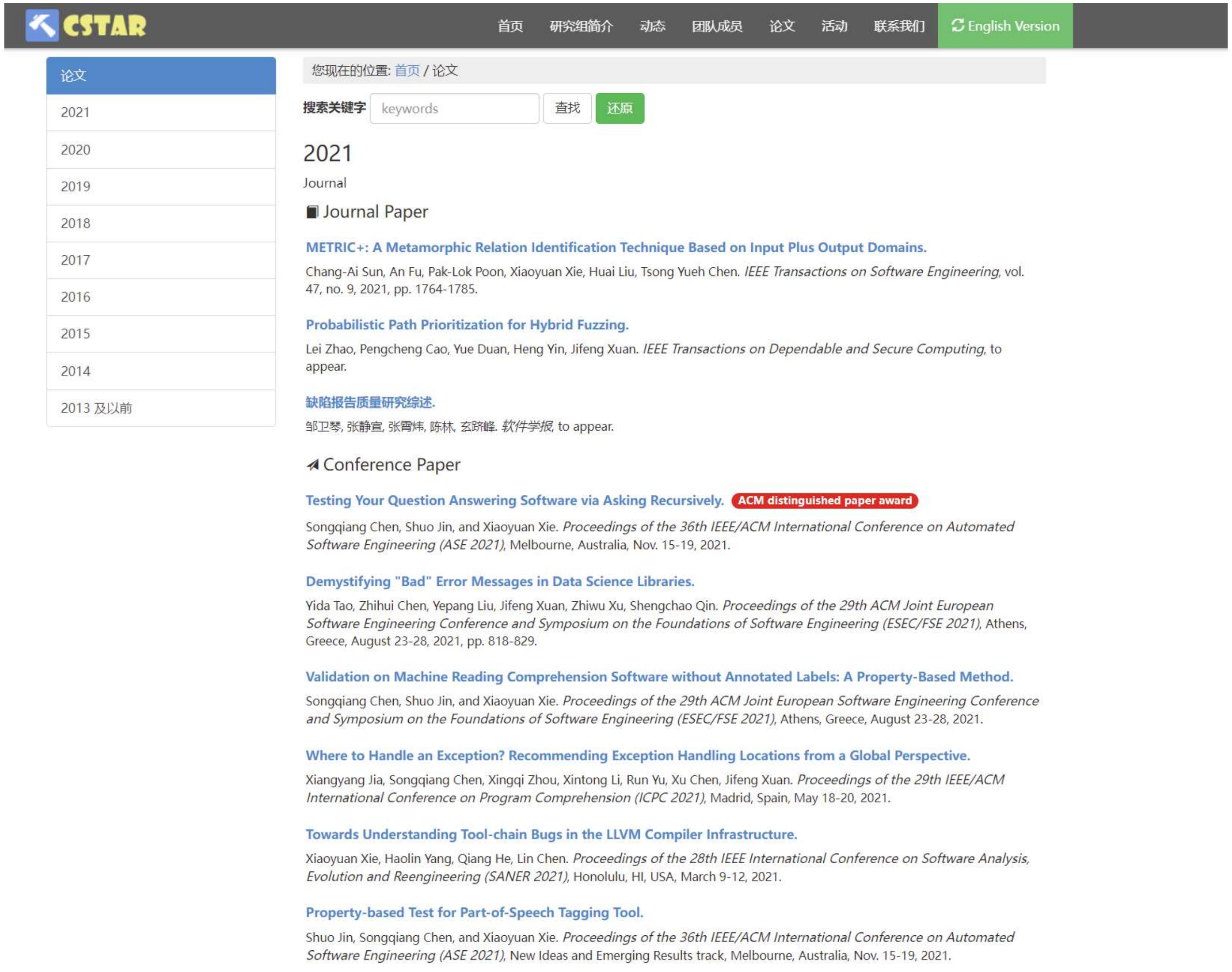
- *Software Testing*
- *Software Reliability and Quality*
- *Program Analysis*
- *Debugging and Fault Localization*
- *Machine Learning*

Jifeng Xuan

Xiaoyuan Xie

Songqiang Chen

http://cstar.whu.edu.cn/cn/index.html



2021

Journal

Journal Paper

METRIC+: A Metamorphic Relation Identification Technique Based on Input Plus Output Domains.

Chang-Ai Sun, An Fu, Pak-Lok Poon, Xiaoyuan Xie, Huai Liu, Tsong Yueh Chen. *IEEE Transactions on Software Engineering*, vol. 47, no. 9, 2021, pp. 1764-1785.

Probabilistic Path Prioritization for Hybrid Fuzzing.

Lei Zhao, Pengcheng Cao, Yue Duan, Heng Yin, Jifeng Xuan. *IEEE Transactions on Dependable and Secure Computing*, to appear.

缺陷报告质量研究综述.

邹卫琴, 张静宣, 张霄炜, 陈林, 玄跻峰. *软件学报*, to appear.

Conference Paper

Testing Your Question Answering Software via Asking Recursively. ACM distinguished paper award

Songqiang Chen, Shuo Jin, and Xiaoyuan Xie. *Proceedings of the 36th IEEE/ACM International Conference on Automated Software Engineering (ASE 2021)*, Melbourne, Australia, Nov. 15-19, 2021.

Demystifying "Bad" Error Messages in Data Science Libraries.

Yida Tao, Zhihui Chen, Yepang Liu, Jifeng Xuan, Zhiwu Xu, Shengchao Qin. *Proceedings of the 29th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE 2021)*, Athens, Greece, August 23-28, 2021, pp. 818-829.

Validation on Machine Reading Comprehension Software without Annotated Labels: A Property-Based Method.

Songqiang Chen, Shuo Jin, and Xiaoyuan Xie. *Proceedings of the 29th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE 2021)*, Athens, Greece, August 23-28, 2021.

Where to Handle an Exception? Recommending Exception Handling Locations from a Global Perspective.

Xiangyang Jia, Songqiang Chen, Xingqi Zhou, Xintong Li, Run Yu, Xu Chen, Jifeng Xuan. *Proceedings of the 29th IEEE/ACM International Conference on Program Comprehension (ICPC 2021)*, Madrid, Spain, May 18-20, 2021.

Towards Understanding Tool-chain Bugs in the LLVM Compiler Infrastructure.

Xiaoyuan Xie, Haolin Yang, Qiang He, Lin Chen. *Proceedings of the 28th IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER 2021)*, Honolulu, HI, USA, March 9-12, 2021.

Property-based Test for Part-of-Speech Tagging Tool.

Shuo Jin, Songqiang Chen, and Xiaoyuan Xie. *Proceedings of the 36th IEEE/ACM International Conference on Automated Software Engineering (ASE 2021)*, New Ideas and Emerging Results track, Melbourne, Australia, Nov. 15-19, 2021.

Why Choose This Paper?

- ACM SIGSOFT Distinguished Paper
- A simple and novel idea
- SE for ML paper

ISE: “ML for SE” or “SE for ML”

ML for SE

1. Code Intelligence
(**code generation**/completion/edit/repair/representation/search/clone/reuse/type prediction/smells/verification/debug/api...)

2. Documentation
(**code comment**/review/var naming/app user review/bug report/commit message/release notes/github/stackoverflow/developer talks&vlogs…)

3. Quality and Maintenance
(**testing**/log/**AIOps**/performance/bug detection/code changes…)

https://ml4code.github.io/tags.html

ISE: “ML for SE” or “SE for ML”

SE for ML

1. Model Evaluation and Testing
(such as testing QA/autopilot/machine translation software)

2. Model Interpretation and Explainability

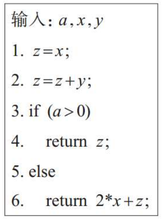
3. Auto ML

Background

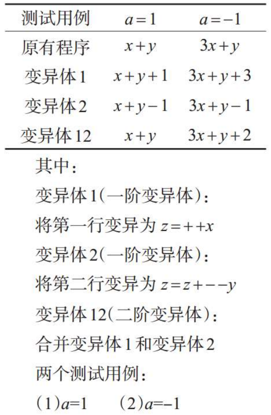
Background: Mutation/Metamorphic/Fuzz Testing

What is Mutation Testing(变异测试)?

- Motivation: Who guards the guardian? Who tests the tests?
- Target: Improve the test cases of program
- Idea: If we mutate the original program, the tests **still pass**. There will be **two possible reasons**: the original program is logically equal to the mutated program / Or the tests could not detect all possible cases of the program.



```
输入：a, x, y
1.  z=x;
2.  z=z+y;
3. if (a>0)
4.     return z;
5. else
6.     return 2*x+z;
```



测试用例	$a=1$	$a=-1$
原有程序	$x+y$	$3x+y$
变异体1	$x+y+1$	$3x+y+3$
变异体2	$x+y-1$	$3x+y-1$
变异体12	$x+y$	$3x+y+2$

其中：

变异体1(一阶变异体)：

将第一行变异为 $z=++x$

变异体2(一阶变异体)：

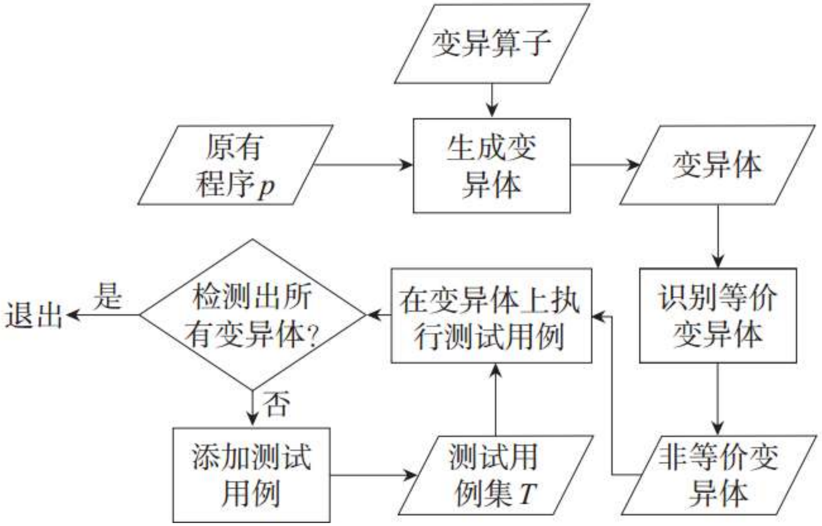
将第二行变异为 $z=z+--y$

变异体12(二阶变异体)：

合并变异体1和变异体2

两个测试用例：

(1)a=1　　(2)a=−1



[1] 软件测试：一个软件工艺师的方法（原书第4版）

[2] https://www.cnblogs.com/TongWee/p/4505289.html

Background: Mutation/Metamorphic/Fuzz Testing

What is Metamorphic Testing(蜕变测试)?

- Motivation: Oracle Problem (How could we test a software when we don’t know expected output)
- Example: we want to test function which computes sin(x)
- Metamorphic testing: we could calculate the result of a random num such as 1.3, then we check if f(1.3) == f(π-1.3), f(1.3) == -f(2π-1.3)
- Advantages: we could test a program we know nothing about
- Key procedures: Build Metamorphic Relation(such as sin(x)==sin(π-x))
- Applications: Machine Learning Software, Complex Software

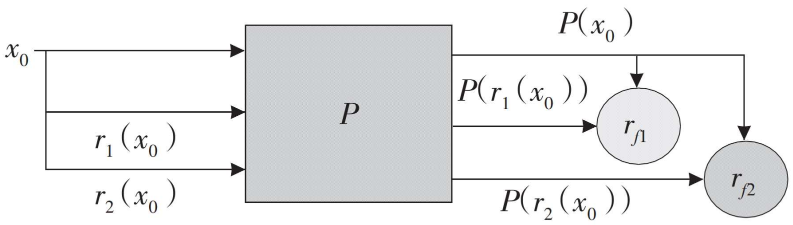


Fig.1 Illustration of metamorphic testing

[1] 钟文康, 葛季栋, 陈翔, 李传艺, 唐泽, 骆斌. 面向神经机器翻译系统的多粒度蜕变测试. 软件学报, 2021, 32(4): 1051-1066. http://www.jos.org.cn/1000-9825/6221.htm
[2] 董国伟, 徐宝文, 陈林, 聂长海, 王璐璐. 蜕变测试技术综述[J]. 计算机科学与探索, 2009, 3(2): 130-143.

Background: Metamorphic Testing

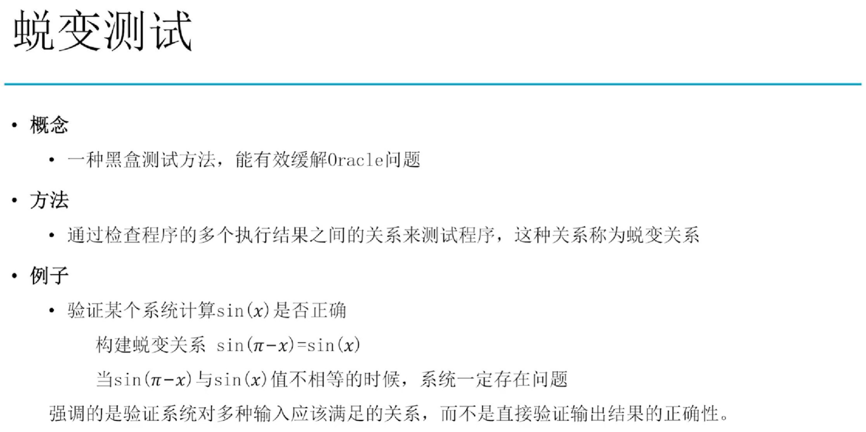
中国科学院软件研究所学术年会'2021
暨计算机科学国家重点实验室开放周

基于蜕变测试的文本定位系统稳定性测试

Stability Evaluation for Text Localization Systems via Metamorphic Testing

晏荣杰　王思琪　闫艺宣　高红雨　严　俊

ISCAS 中国科学院软件研究所 Institute of Software Chinese Academy of Sciences
北京工业大学 BEIJINGUNIVERSITYOFTECHNOLOGY



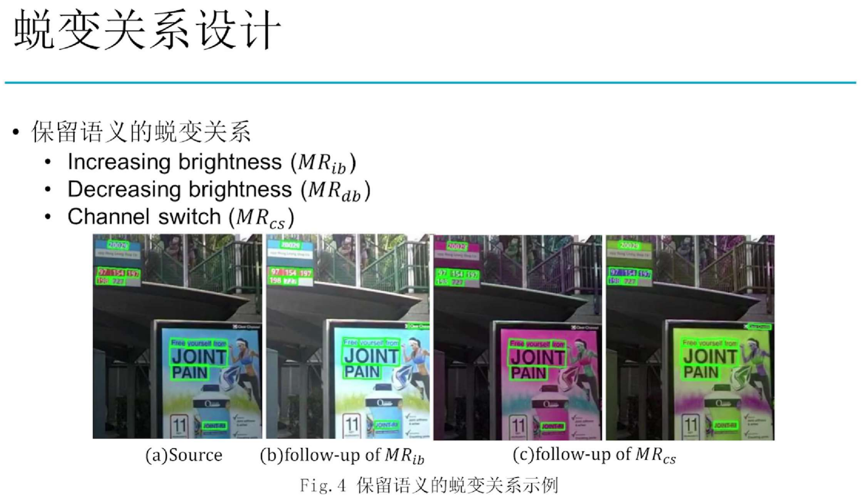
蜕变测试

- 概念
  - 一种黑盒测试方法，能有效缓解Oracle问题
- 方法
  - 通过检查程序的多个执行结果之间的关系来测试程序，这种关系称为蜕变关系
- 例子
  - 验证某个系统计算sin(x)是否正确

    构建蜕变关系 sin($\pi-x$)=sin(x)

    当sin($\pi-x$)与sin(x)值不相等的时候，系统一定存在问题

  强调的是验证系统对多种输入应该满足的关系，而不是直接验证输出结果的正确性。

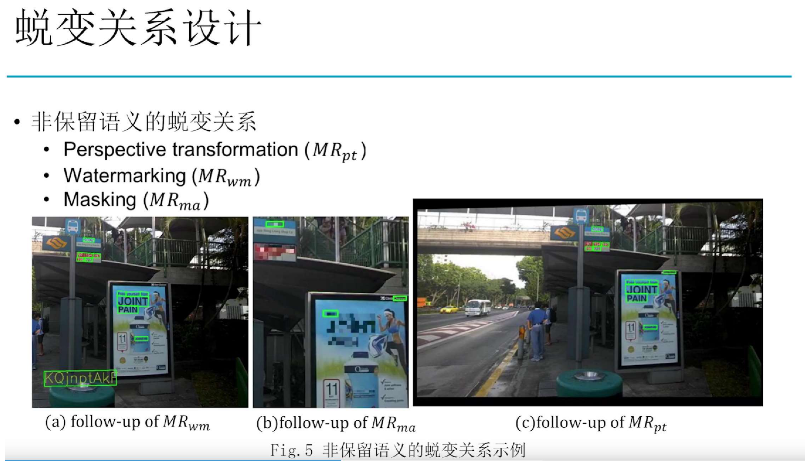


蜕变关系设计

- 保留语义的蜕变关系
  - Increasing brightness (MR_{ib})
  - Decreasing brightness (MR_{db})
  - Channel switch (MR_{cs})

(a)Source　(b)follow-up of MR_{ib}　(c)follow-up of MR_{cs}

Fig. 4 保留语义的蜕变关系示例



蜕变关系设计

- 非保留语义的蜕变关系
  - Perspective transformation (MR_{pt})
  - Watermarking (MR_{wm})
  - Masking (MR_{ma})

(a) follow-up of MR_{wm}　(b)follow-up of MR_{ma}　(c)follow-up of MR_{pt}

Fig. 5 非保留语义的蜕变关系示例

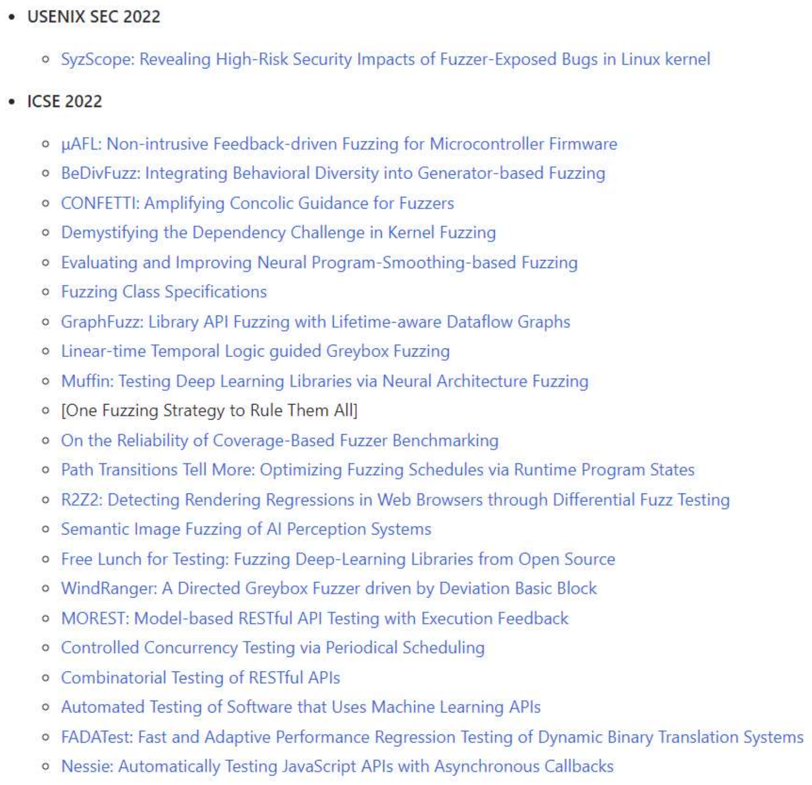
[1]基于蜕变测试的文本定位系统稳定性测试 https://www.bilibili.com/video/BV1b64y1Y7UY?spm_id_from=333.337.search-card.all.click

[2] http://www.is.cas.cn/ztzl2016/2021xsnh/2021hbzs/

Background: Mutation/Metamorphic/Fuzz Testing

What is Fuzz Testing(模糊测试)?

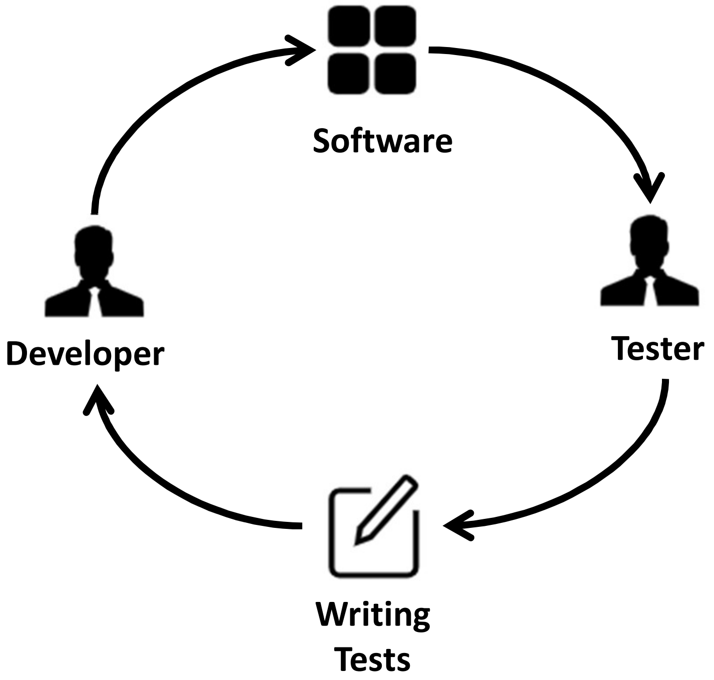
- Definition: Fuzzing or fuzz testing is an automated software testing technique that involves providing invalid, unexpected, or random data as inputs to a computer program.
- History: 1988, Prof. Barton Miller was testing the reliability of UNIX command line programs. But due to the heavy rain, there were some unexpected wrong inputs sent to the program, which caused the program to terminate.
- Motivation: We cannot list every possible case or predict every exception
- Application: Security & Vulnerability, Software Testing



- USENIX SEC 2022
  - SyzScope: Revealing High-Risk Security Impacts of Fuzzer-Exposed Bugs in Linux kernel
- ICSE 2022
  - µAFL: Non-intrusive Feedback-driven Fuzzing for Microcontroller Firmware
  - BeDivFuzz: Integrating Behavioral Diversity into Generator-based Fuzzing
  - CONFETTI: Amplifying Concolic Guidance for Fuzzers
  - Demystifying the Dependency Challenge in Kernel Fuzzing
  - Evaluating and Improving Neural Program-Smoothing-based Fuzzing
  - Fuzzing Class Specifications
  - GraphFuzz: Library API Fuzzing with Lifetime-aware Dataflow Graphs
  - Linear-time Temporal Logic guided Greybox Fuzzing
  - Muffin: Testing Deep Learning Libraries via Neural Architecture Fuzzing
  - [One Fuzzing Strategy to Rule Them All]
  - On the Reliability of Coverage-Based Fuzzer Benchmarking
  - Path Transitions Tell More: Optimizing Fuzzing Schedules via Runtime Program States
  - R2Z2: Detecting Rendering Regressions in Web Browsers through Differential Fuzz Testing
  - Semantic Image Fuzzing of AI Perception Systems
  - Free Lunch for Testing: Fuzzing Deep-Learning Libraries from Open Source
  - WindRanger: A Directed Greybox Fuzzer driven by Deviation Basic Block
  - MOREST: Model-based RESTful API Testing with Execution Feedback
  - Controlled Concurrency Testing via Periodical Scheduling
  - Combinatorial Testing of RESTful APIs
  - Automated Testing of Software that Uses Machine Learning APIs
  - FADATest: Fast and Adaptive Performance Regression Testing of Dynamic Binary Translation Systems
  - Nessie: Automatically Testing JavaScript APIs with Asynchronous Callbacks

[1] https://zhuanlan.zhihu.com/p/43432370
[2] https://github.com/wcventure/FuzzingPaper

Problem: Could all software be “tested”?



Front-End Software

- GUI Automatic Testing is not fully developed
- Take much manual effort to check every page
- Difficult to test on every possible environment (device)

Untestable

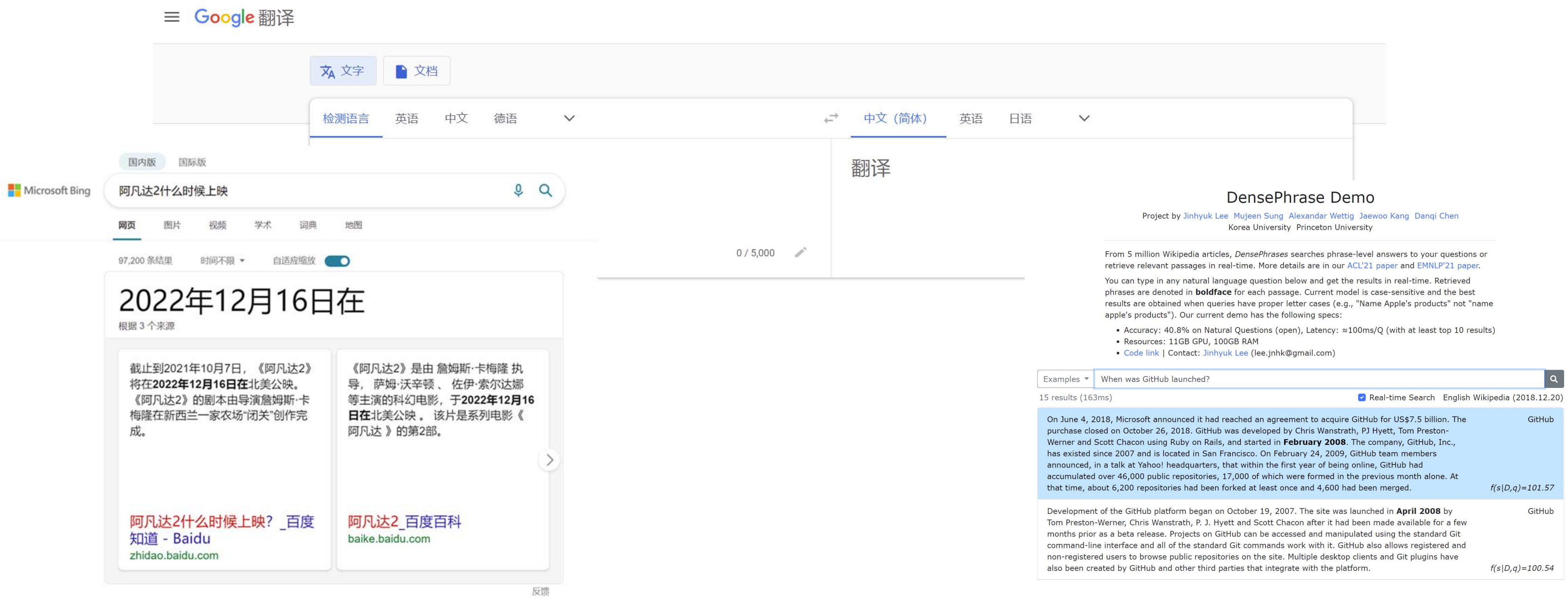
AI Software

- Difficult to define equivalent classes
- Impossible to list every case
- Lack of interpretability and explainability

https://scholar.google.com.hk/scholar?hl=zh-CN&as_sdt=0%2C5&q=untestable+software&btnG=

Background: Testing Untestable

How to test an AI software (intelligent software)?



阿凡达2什么时候上映

2022年12月16日在

根据 3 个来源

截止到2021年10月7日，《阿凡达2》将在**2022年12月16日在**北美公映。《阿凡达2》的剧本由导演詹姆斯·卡梅隆在新西兰一家农场"闭关"创作完成。

阿凡达2什么时候上映？_百度知道 - Baidu
zhidao.baidu.com

《阿凡达2》是由 詹姆斯·卡梅隆 执导， 萨姆·沃辛顿 、 佐伊·索尔达娜 等主演的科幻电影，于**2022年12月16日在**北美公映 。 该片是系列电影《 阿凡达 》的第2部。

阿凡达2_百度百科
baike.baidu.com

DensePhrase Demo

Project by Jinhyuk Lee Mujeen Sung Alexandar Wettig Jaewoo Kang Danqi Chen
Korea University Princeton University

From 5 million Wikipedia articles, *DensePhrases* searches phrase-level answers to your questions or retrieve relevant passages in real-time. More details are in our ACL'21 paper and EMNLP'21 paper.

You can type in any natural language question below and get the results in real-time. Retrieved phrases are denoted in **boldface** for each passage. Current model is case-sensitive and the best results are obtained when queries have proper letter cases (e.g., "Name Apple's products" not "name apple's products"). Our current demo has the following specs:

- Accuracy: 40.8% on Natural Questions (open), Latency: ≈100ms/Q (with at least top 10 results)
- Resources: 11GB GPU, 100GB RAM
- Code link | Contact: Jinhyuk Lee (lee.jnhk@gmail.com)

Examples ▾ When was GitHub launched?

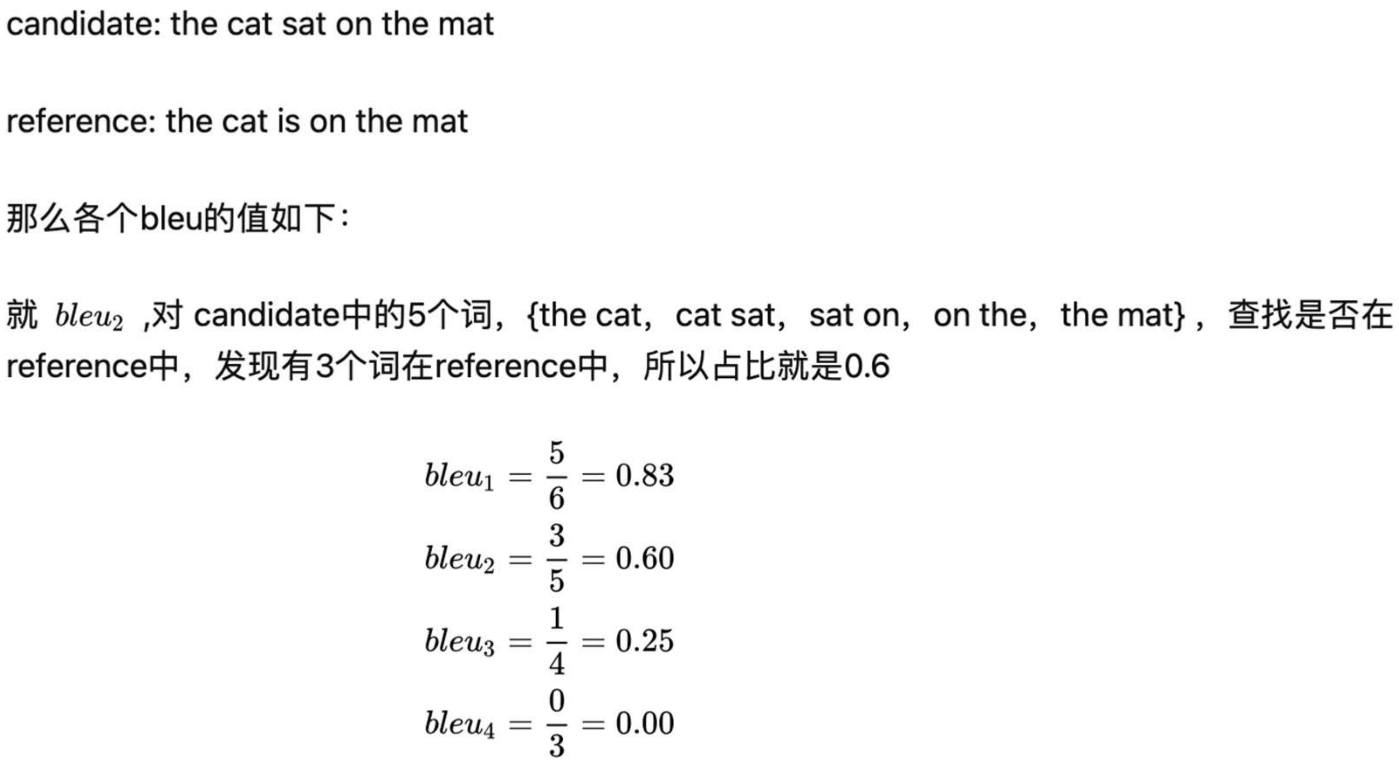
15 results (163ms) ☑ Real-time Search English Wikipedia (2018.12.20)

On June 4, 2018, Microsoft announced it had reached an agreement to acquire GitHub for US$7.5 billion. The purchase closed on October 26, 2018. GitHub was developed by Chris Wanstrath, PJ Hyett, Tom Preston-Werner and Scott Chacon using Ruby on Rails, and started in **February 2008**. The company, GitHub, Inc., has existed since 2007 and is located in San Francisco. On February 24, 2009, GitHub team members announced, in a talk at Yahoo! headquarters, that within the first year of being online, GitHub had accumulated over 46,000 public repositories, 17,000 of which were formed in the previous month alone. At that time, about 6,200 repositories had been forked at least once and 4,600 had been merged. GitHub $f(s|D,q)=101.57$

Development of the GitHub platform began on October 19, 2007. The site was launched in **April 2008** by Tom Preston-Werner, Chris Wanstrath, P. J. Hyett and Scott Chacon after it had been made available for a few months prior as a beta release. Projects on GitHub can be accessed and manipulated using the standard Git command-line interface and all of the standard Git commands work with it. GitHub also allows registered and non-registered users to browse public repositories on the site. Multiple desktop clients and Git plugins have also been created by GitHub and other third parties that integrate with the platform. GitHub $f(s|D,q)=100.54$

Background: Model Evaluation Metrics

BLEU



candidate: the cat sat on the mat

reference: the cat is on the mat

那么各个bleu的值如下：

就 $bleu_2$,对 candidate中的5个词，{the cat，cat sat，sat on，on the，the mat}，查找是否在reference中，发现有3个词在reference中，所以占比就是0.6

$$bleu_1 = \frac{5}{6} = 0.83$$

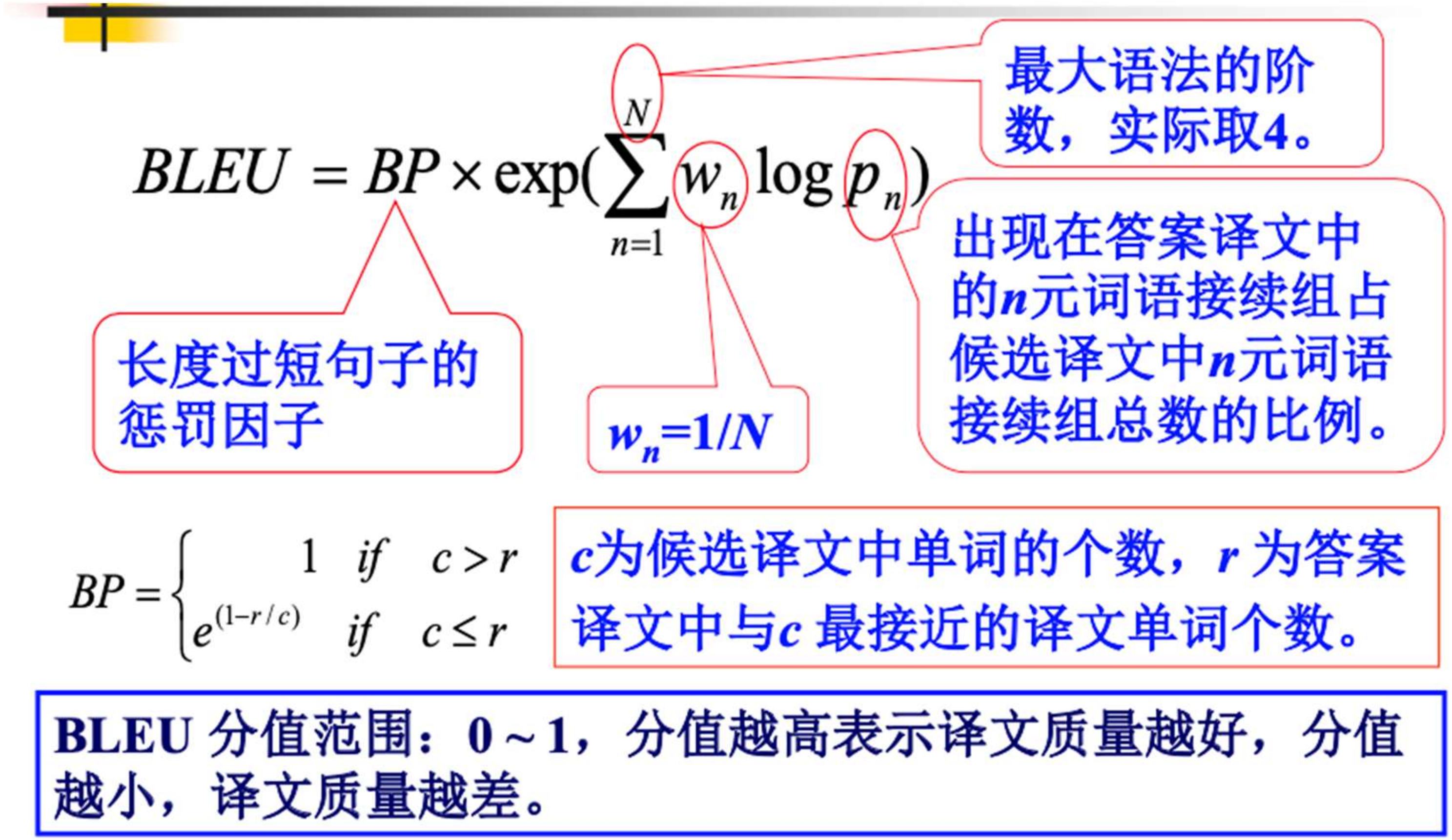
$$bleu_2 = \frac{3}{5} = 0.60$$

$$bleu_3 = \frac{1}{4} = 0.25$$

$$bleu_4 = \frac{0}{3} = 0.00$$

Background: Model Evaluation Metrics

BLEU



$$BLEU = BP \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

最大语法的阶数，实际取4。

长度过短句子的惩罚因子

w_n=1/N

出现在答案译文中的n元词语接续组占候选译文中n元词语接续组总数的比例。

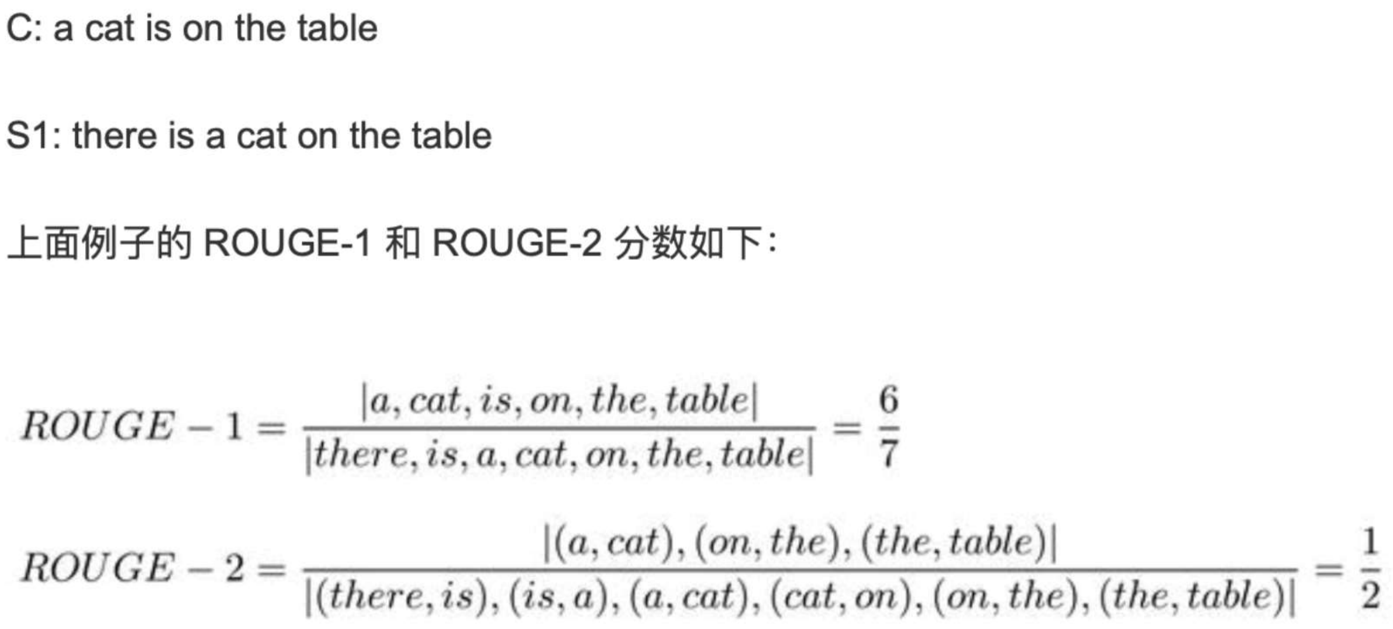
$$BP = \begin{cases} 1 & if \quad c > r \\ e^{(1-r/c)} & if \quad c \le r \end{cases}$$

c为候选译文中单词的个数，r 为答案译文中与c 最接近的译文单词个数。

BLEU 分值范围：0～1，分值越高表示译文质量越好，分值越小，译文质量越差。

Background: Model Evaluation Metrics

ROUGE



C: a cat is on the table

S1: there is a cat on the table

上面例子的 ROUGE-1 和 ROUGE-2 分数如下：

$$ROUGE-1 = \frac{|a, cat, is, on, the, table|}{|there, is, a, cat, on, the, table|} = \frac{6}{7}$$

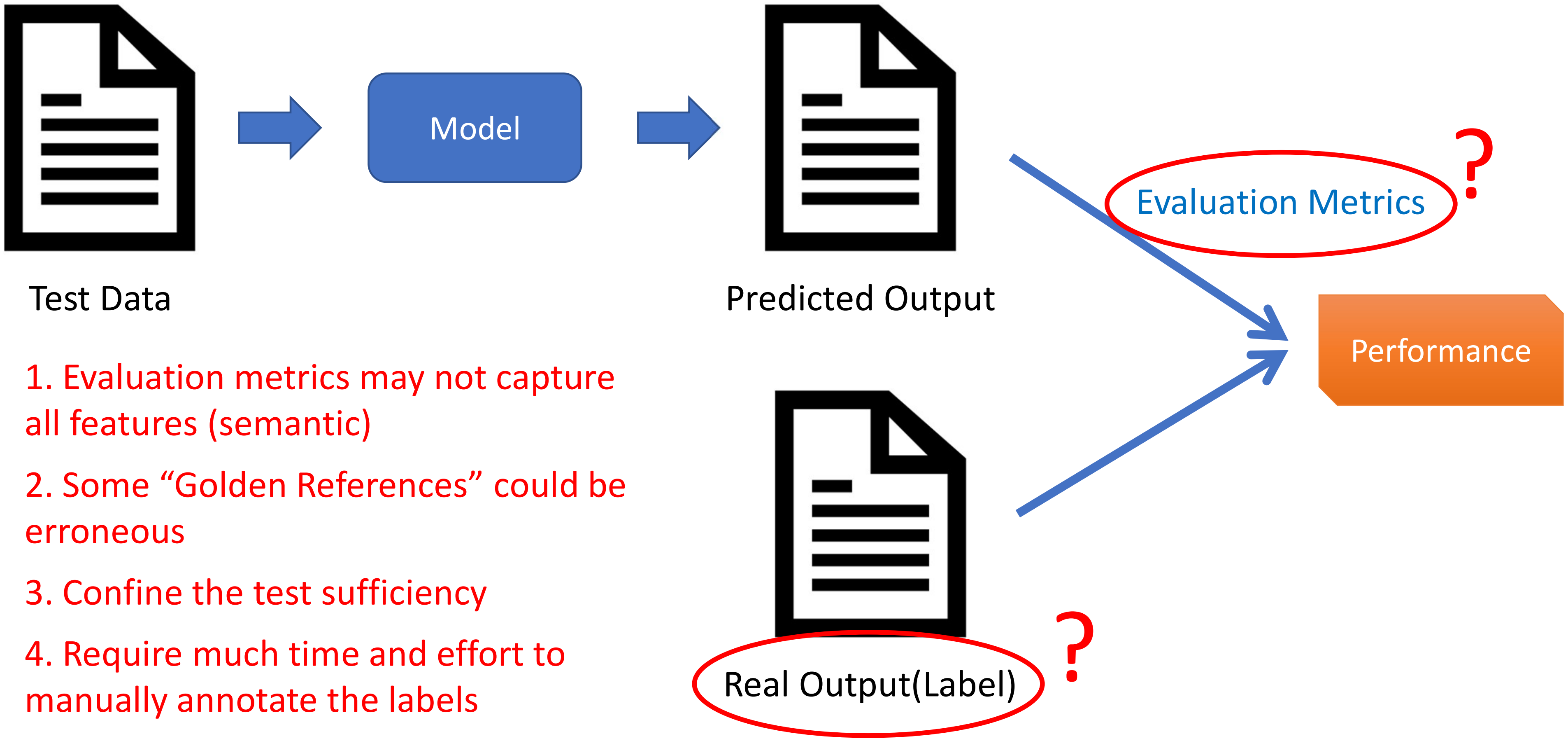
$$ROUGE-2 = \frac{|(a, cat), (on, the), (the, table)|}{|(there, is), (is, a), (a, cat), (cat, on), (on, the), (the, table)|} = \frac{1}{2}$$

Background: Model Evaluation Metrics

ROUGE

$$\text{ROUGE - N} = \frac{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_N \in S} Count_{match}(gram_N)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_N \in S} Count(gram_N)}$$

Background: Model Evaluation



1. Evaluation metrics may not capture all features (semantic)

2. Some “Golden References” could be erroneous

3. Confine the test sufficiency

4. Require much time and effort to manually annotate the labels

5. Just-in-time Testing (Real time / detect issues during usage)

Background: Testing Untestable

Bad Case Mining for Machine Translation using Back Translation

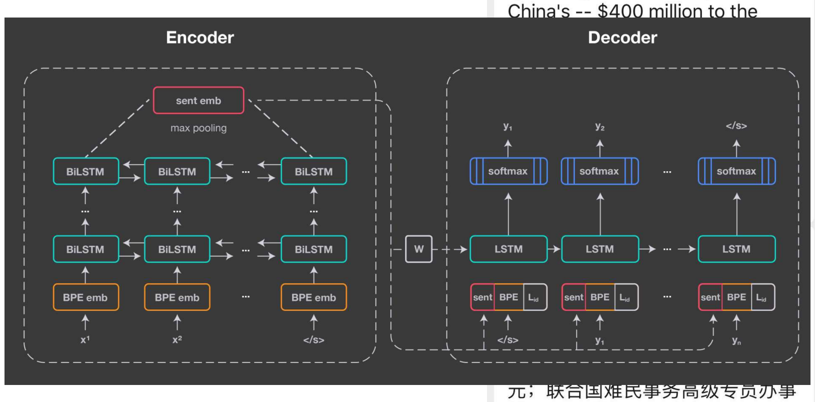
- Bad Case: Cases that could not receive expected result
- Back Translation: A common data augmentation method
- Translate source sentence to a sentence in another language, and then translate back into the original language
- Finally, compute the similarity between source sentence and target sentence (Representation of sentences: LASER)

Eg: "**周杰伦是一位华语乐坛的实力唱将，他的专辑卖遍了全球**" —>

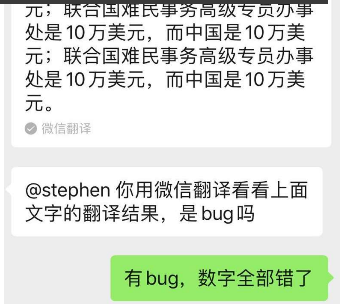
"**Jay Chou is a strength singer in the Chinese music scene, his albums are sold all over the world.** "—>

"**周杰伦是中国音乐界的优秀歌手，他的专辑畅销全世界。**"

The State Department followed up by issuing a fact sheet Tuesday, showing how U.S. contributions to global agencies far surpass China's -- $400 million to the



LASER

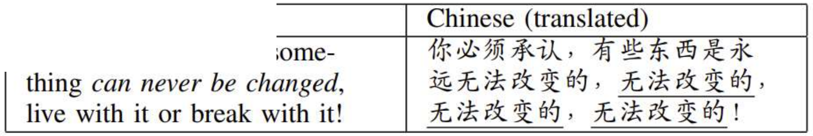


元；联合国难民事务高级专员办事处是10万美元，而中国是10万美元；联合国难民事务高级专员办事处是10万美元，而中国是10万美元。

微信翻译

@stephen 你用微信翻译看看上面文字的翻译结果，是bug吗

有bug，数字全部错了



	Chinese (translated)
ome- thing *can never be changed*, live with it or break with it!	你必须承认，有些东西是永远无法改变的，无法改变的，无法改变的，无法改变的！

Bad Case

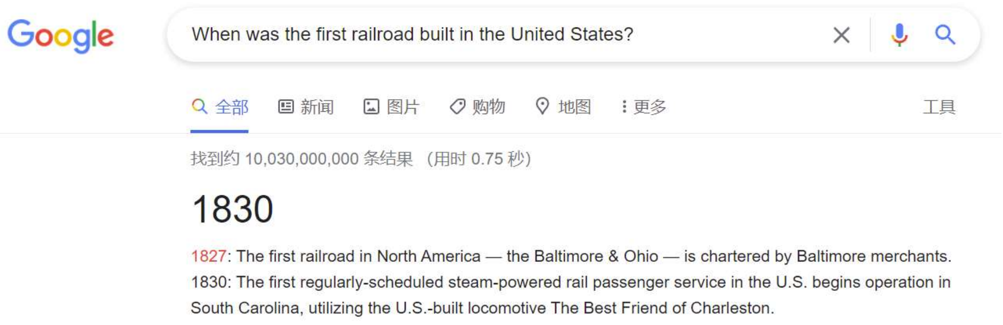
[1] Zheng W, Wang W, Liu D, et al. Testing untestable neural machine translation: An industrial case[C]//2019 IEEE/ACM 41st International Conference on Software Engineering: Companion Proceedings (ICSE-Companion). IEEE, 2019: 314-315.

[2] Wang W, Zheng W, Liu D, et al. Detecting failures of neural machine translation in the absence of reference translations[C]//2019 49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks–Industry Track. IEEE, 2019: 1-4.

Introduction

Is it possible to transfer this method to other AI software?

- Metamorphic Testing: Testing the Untestable
- Question Answering Software?

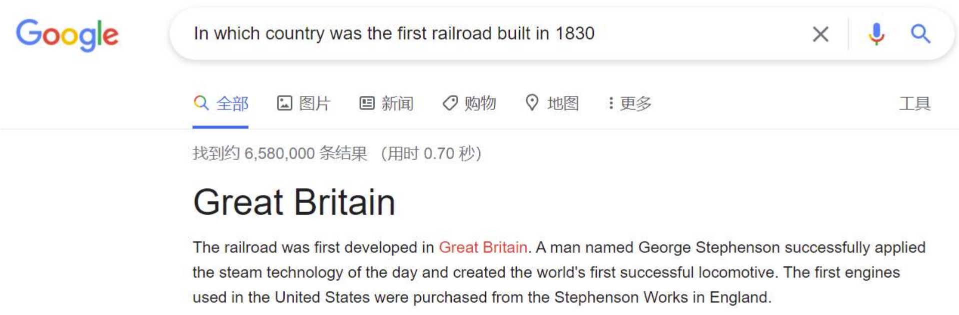


When was the first railroad built in the United States?

找到约 10,030,000,000 条结果 （用时 0.75 秒）

1830

1827: The first railroad in North America — the Baltimore & Ohio — is chartered by Baltimore merchants. 1830: The first regularly-scheduled steam-powered rail passenger service in the U.S. begins operation in South Carolina, utilizing the U.S.-built locomotive The Best Friend of Charleston.



In which country was the first railroad built in 1830

找到约 6,580,000 条结果 （用时 0.70 秒）

Great Britain

The railroad was first developed in Great Britain. A man named George Stephenson successfully applied the steam technology of the day and created the world's first successful locomotive. The first engines used in the United States were purchased from the Stephenson Works in England.

t?

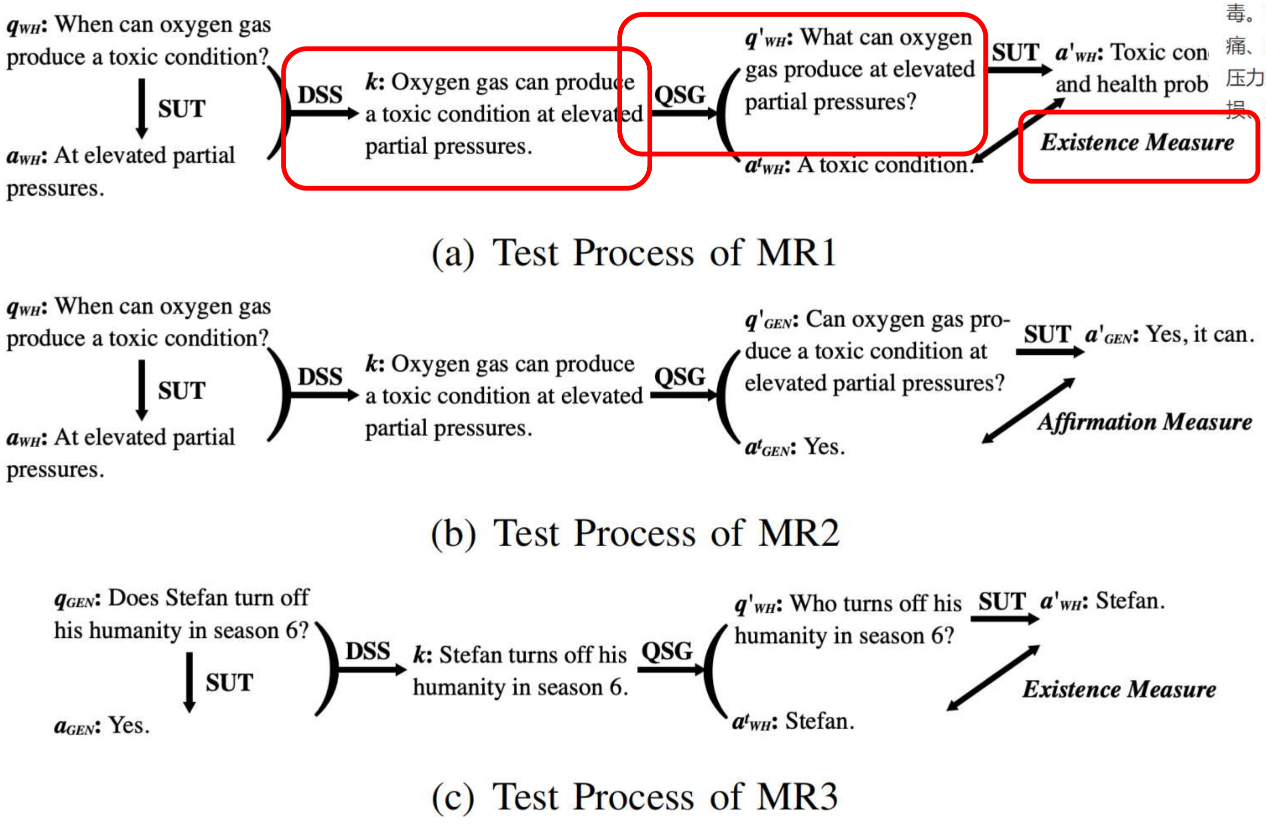
There is a violation between the two answers

（Actually, the first railroad is built in United States on 1827-02-28)

[1] Segura S, Towey D, Zhou Z Q, et al. Metamorphic testing: Testing the untestable[J]. IEEE Software, 2018, 37(3): 46-53.

Implementation

Overview of the procedures


q_{WH}: When can oxygen gas produce a toxic condition?
SUT
a_{WH}: At elevated partial pressures.
DSS
k: Oxygen gas can produce a toxic condition at elevated partial pressures.
QSG
q'_{WH}: What can oxygen gas produce at elevated partial pressures?
SUT
a'_{WH}: Toxic con
and health prob
a^t_{WH}: A toxic condition.
Existence Measure

(a) Test Process of MR1

q_{WH}: When can oxygen gas produce a toxic condition?
SUT
a_{WH}: At elevated partial pressures.
DSS
k: Oxygen gas can produce a toxic condition at elevated partial pressures.
QSG
q'_{GEN}: Can oxygen gas produce a toxic condition at elevated partial pressures?
SUT
a'_{GEN}: Yes, it can.
a^t_{GEN}: Yes.
Affirmation Measure

(b) Test Process of MR2

q_{GEN}: Does Stefan turn off his humanity in season 6?
SUT
a_{GEN}: Yes.
DSS
k: Stefan turns off his humanity in season 6.
QSG
q'_{WH}: Who turns off his humanity in season 6?
SUT
a'_{WH}: Stefan.
a^t_{WH}: Stefan.
Existence Measure

(c) Test Process of MR3

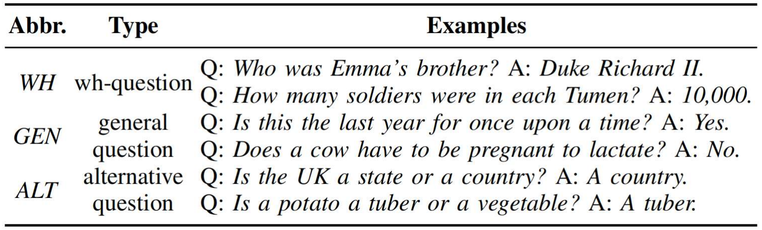
Fig. 1. Proposed Recursive Metamorphic Relations

你好！因为当氧气浓度超过70%的时候，高纯度氧气会对人体产生危害，也就是所谓的“氧气”中毒。吸入在标准大气压下80%或更高纯度的氧气2-3个小时，会引起鼻子不通、咳嗽、咽喉疼痛、胸痛、呼吸困难等症状，若是在更高的压力下吸入氧气，会在更短的时间内出现以上症状。若是吸入有压力的纯氧，可能损害肺功能和中枢神经系统，从而导致弦晕、感觉迟钝、刺痛感觉、视觉及听觉受损、肌肉抽动、意识减退和痉

DSS: Declarative Sentence Synthesis
QSG: Question Sentence Generation
k: Knowledge(Declarative Sentence)

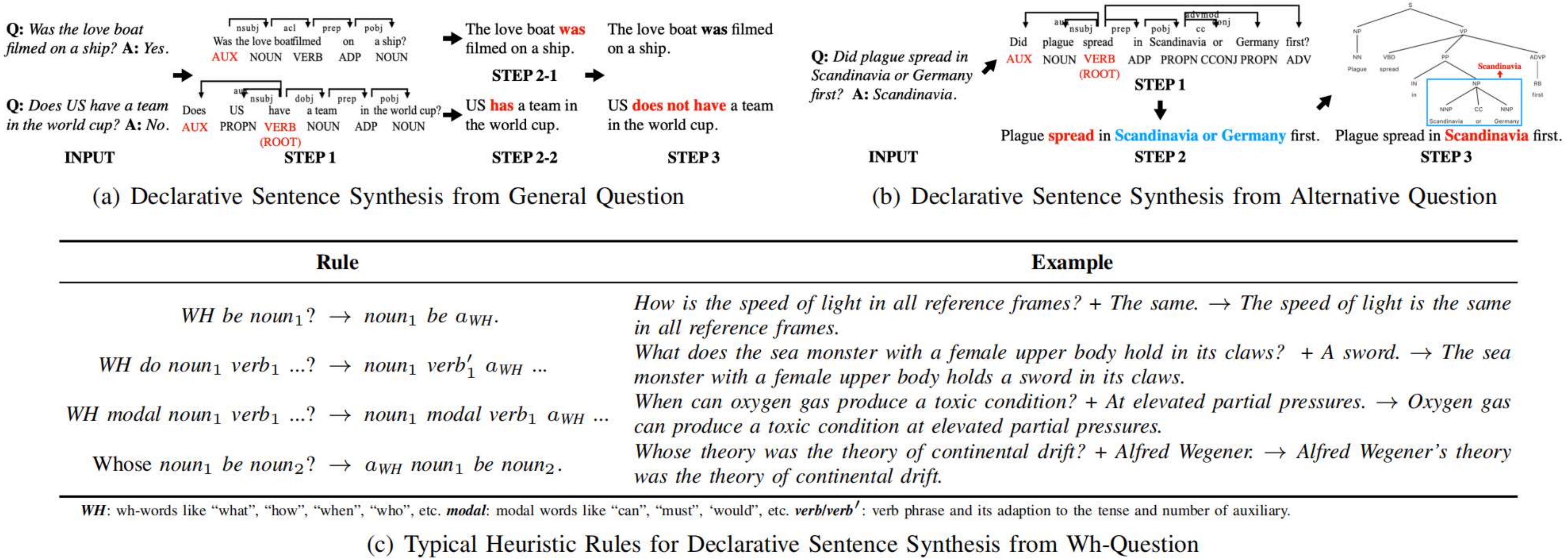
Question Definition:

WH: wh-question(what when who how many)
GEN: general question
ALT: alternative question



Abbr.	Type	Examples
WH	wh-question	Q: *Who was Emma's brother?* A: *Duke Richard II.* Q: *How many soldiers were in each Tumen?* A: *10,000.*
GEN	general question	Q: *Is this the last year for once upon a time?* A: *Yes.* Q: *Does a cow have to be pregnant to lactate?* A: *No.*
ALT	alternative question	Q: *Is the UK a state or a country?* A: *A country.* Q: *Is a potato a tuber or a vegetable?* A: *A tuber.*

Declaration Sentence Synthesis



Q: *Was the love boat filmed on a ship?* **A:** *Yes*.

The love boat **was** filmed on a ship.

The love boat **was** filmed on a ship.

Q: *Does US have a team in the world cup?* **A:** *No*.

US **has** a team in the world cup.

US **does not have** a team in the world cup.

(a) Declarative Sentence Synthesis from General Question

Q: *Did plague spread in Scandinavia or Germany first?* **A:** *Scandinavia*.

Plague **spread** in **Scandinavia or Germany** first.

Plague spread in **Scandinavia** first.

(b) Declarative Sentence Synthesis from Alternative Question

Rule	Example
WH be noun$_1$? → *noun*$_1$ *be* a_{WH}.	*How is the speed of light in all reference frames? + The same. → The speed of light is the same in all reference frames.*
WH do noun$_1$ *verb*$_1$ ...? → *noun*$_1$ *verb*$'_1$ a_{WH} ...	*What does the sea monster with a female upper body hold in its claws? + A sword. → The sea monster with a female upper body holds a sword in its claws.*
WH modal noun$_1$ *verb*$_1$ ...? → *noun*$_1$ *modal verb*$_1$ a_{WH} ...	*When can oxygen gas produce a toxic condition? + At elevated partial pressures. → Oxygen gas can produce a toxic condition at elevated partial pressures.*
Whose *noun*$_1$ *be noun*$_2$? → a_{WH} *noun*$_1$ *be noun*$_2$.	*Whose theory was the theory of continental drift? + Alfred Wegener. → Alfred Wegener's theory was the theory of continental drift.*

WH: wh-words like "what", "how", "when", "who", etc. ***modal***: modal words like "can", "must", 'would", etc. ***verb/verb'***: verb phrase and its adaption to the tense and number of auxiliary.

(c) Typical Heuristic Rules for Declarative Sentence Synthesis from Wh-Question

Fig. 2. Process and Example of Declarative Sentence Synthesis

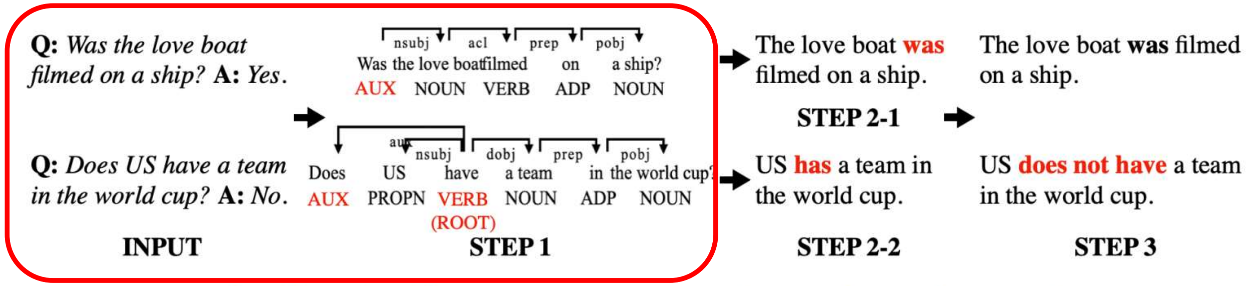
[1] Spacy简介: https://www.jianshu.com/p/e6b3565e159d
[2] Pattern简介: https://python.freelycode.com/contribution/detail/1609

Declaration Sentence Synthesis

Step1: POS tagging(词性标注) and Dependency Parsing (using Spacy)

Step2: Adjust the place of AUX
(AUX could be “be/can/may” or “do/did”)
(might need to transform the tense and number)

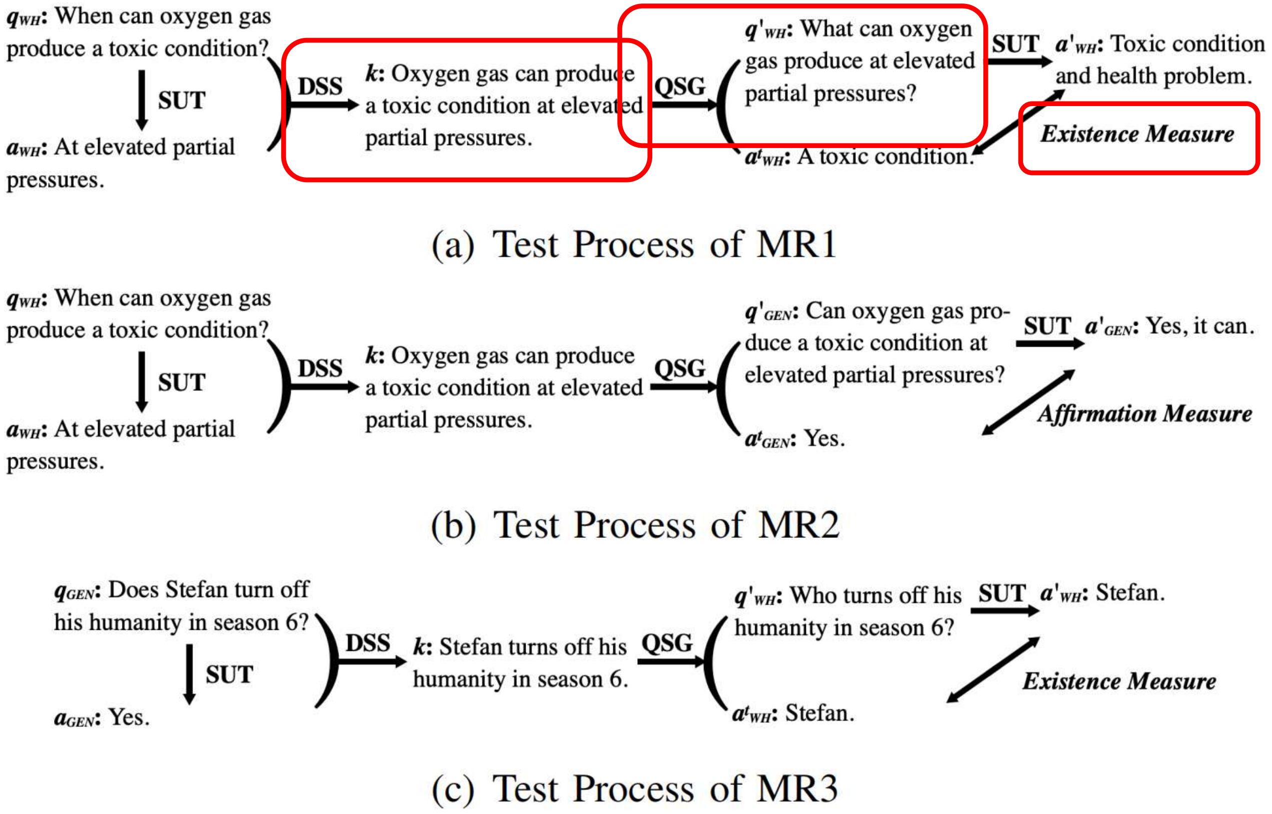
Step3: Negate k if answer is “No”


Q: Was the love boat filmed on a ship? A: Yes.
nsubj
acl
prep
pobj
Was the love boat filmed on a ship?
AUX NOUN VERB ADP NOUN
The love boat was filmed on a ship.
The love boat was filmed on a ship.
STEP 2-1
Q: Does US have a team in the world cup? A: No.
aux
nsubj
dobj
prep
pobj
Does US have a team in the world cup?
AUX PROPN VERB NOUN ADP NOUN
(ROOT)
US has a team in the world cup.
US does not have a team in the world cup.
INPUT
STEP 1
STEP 2-2
STEP 3

(a) Declarative Sentence Synthesis from General Question

[1] Spacy简介: https://www.jianshu.com/p/e6b3565e159d
[2] Pattern简介: https://python.freelycode.com/contribution/detail/1609

After DSS …



q_{WH}: When can oxygen gas produce a toxic condition?

SUT

a_{WH}: At elevated partial pressures.

DSS

k: Oxygen gas can produce a toxic condition at elevated partial pressures.

QSG

q'_{WH}: What can oxygen gas produce at elevated partial pressures?

SUT

a'_{WH}: Toxic condition and health problem.

a^t_{WH}: A toxic condition.

Existence Measure

(a) Test Process of MR1

q_{WH}: When can oxygen gas produce a toxic condition?

SUT

a_{WH}: At elevated partial pressures.

DSS

k: Oxygen gas can produce a toxic condition at elevated partial pressures.

QSG

q'_{GEN}: Can oxygen gas produce a toxic condition at elevated partial pressures?

SUT

a'_{GEN}: Yes, it can.

a^t_{GEN}: Yes.

Affirmation Measure

(b) Test Process of MR2

q_{GEN}: Does Stefan turn off his humanity in season 6?

SUT

a_{GEN}: Yes.

DSS

k: Stefan turns off his humanity in season 6.

QSG

q'_{WH}: Who turns off his humanity in season 6?

SUT

a'_{WH}: Stefan.

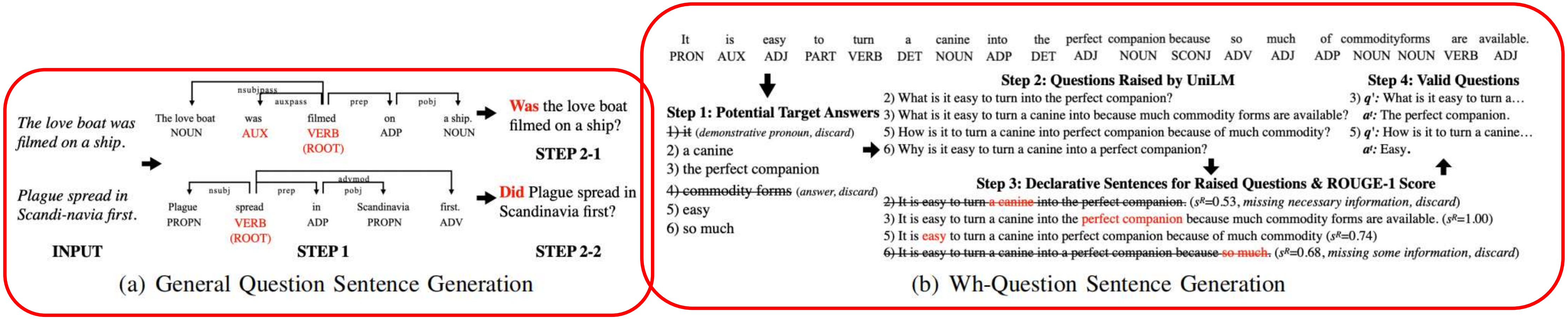
a^t_{WH}: Stefan.

Existence Measure

(c) Test Process of MR3

Fig. 1. Proposed Recursive Metamorphic Relations

Question Sentence Generation



The love boat was filmed on a ship.

Was the love boat filmed on a ship?

Plague spread in Scandi-navia first.

Did Plague spread in Scandinavia first?

(a) General Question Sentence Generation

Step 1: Potential Target Answers

~~1) it~~ *(demonstrative pronoun, discard)*
2) a canine
3) the perfect companion
~~4) commodity forms~~ *(answer, discard)*
5) easy
6) so much

Step 2: Questions Raised by UniLM

2) What is it easy to turn into the perfect companion?
3) What is it easy to turn a canine into because much commodity forms are available?
5) How is it to turn a canine into perfect companion because of much commodity?
6) Why is it easy to turn a canine into a perfect companion?

Step 3: Declarative Sentences for Raised Questions & ROUGE-1 Score

~~2) It is easy to turn a canine into the perfect companion.~~ (s^R=0.53, *missing necessary information, discard*)
3) It is easy to turn a canine into the perfect companion because much commodity forms are available. (s^R=1.00)
5) It is easy to turn a canine into perfect companion because of much commodity (s^R=0.74)
~~6) It is easy to turn a canine into a perfect companion because so much.~~ (s^R=0.68, *missing some information, discard*)

Step 4: Valid Questions

3) ***q'***: What is it easy to turn a…
a^t: The perfect companion.
5) ***q'***: How is it to turn a canine…
a^t: Easy.

(b) Wh-Question Sentence Generation

Fig. 3. Process and Example of Follow-up Question Sentence Generation

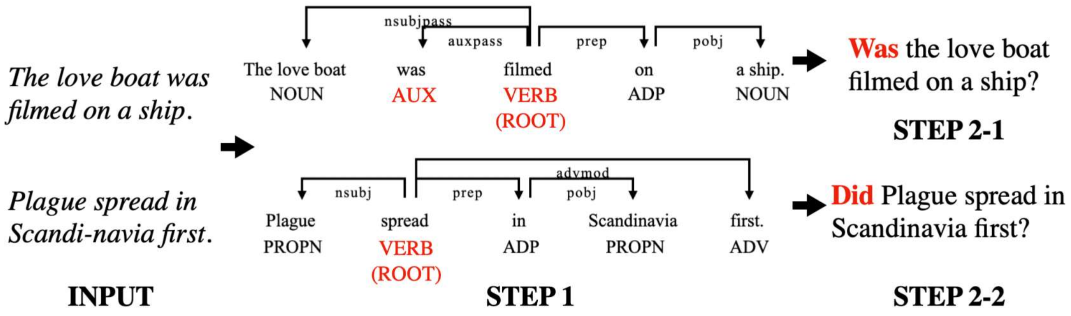
Question Sentence Generation (General question)

Step1: POS tagging(词性标注) and Dependency Parsing (using Spacy)

Step2: Check if AUX exists
(AUX could be “be/can/may” or “do/did”)

Step2-1: If AUX exists, move it to the beginning of the whole sentence.

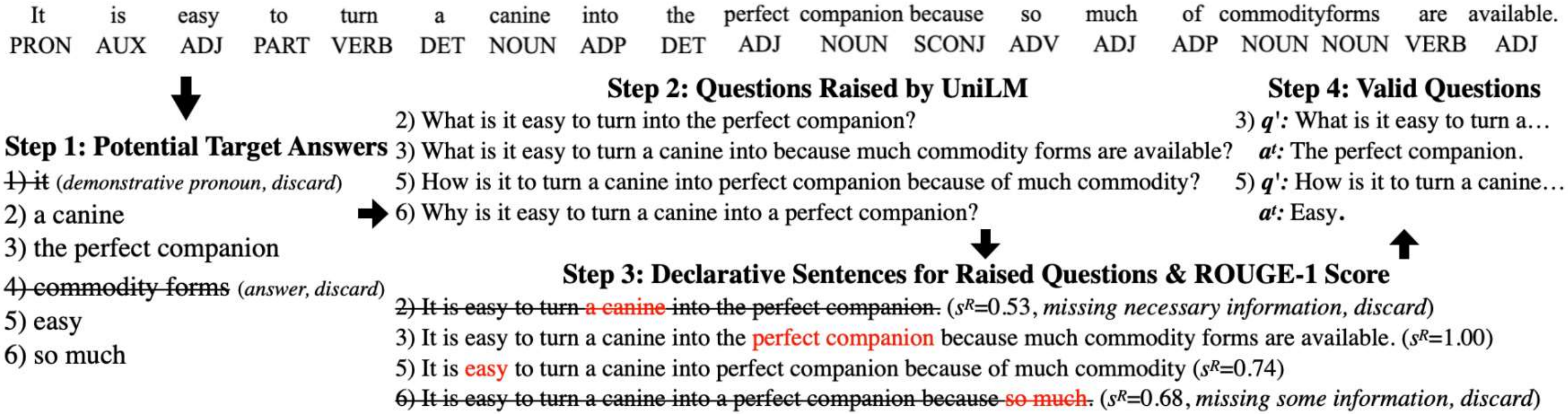
Step2-2: If AUX doesn’t exist, use Pattern Library to recognize the tense and number of VERB(ROOT). And Insert a do with suitable tense and number.



(a) General Question Sentence Generation

[1] Spacy简介: https://www.jianshu.com/p/e6b3565e159d
[2] Pattern简介: https://python.freelycode.com/contribution/detail/1609

Question Sentence Generation (Wh-question)



It is easy to turn a canine into the perfect companion because so much of commodityforms are available.
PRON AUX ADJ PART VERB DET NOUN ADP DET ADJ NOUN SCONJ ADV ADJ ADP NOUN NOUN VERB ADJ

Step 1: Potential Target Answers

~~1) it~~ *(demonstrative pronoun, discard)*
2) a canine
3) the perfect companion
~~4) commodity forms~~ *(answer, discard)*
5) easy
6) so much

Step 2: Questions Raised by UniLM

2) What is it easy to turn into the perfect companion?
3) What is it easy to turn a canine into because much commodity forms are available?
5) How is it to turn a canine into perfect companion because of much commodity?
6) Why is it easy to turn a canine into a perfect companion?

Step 3: Declarative Sentences for Raised Questions & ROUGE-1 Score

~~2) It is easy to turn a canine into the perfect companion.~~ (s^R=0.53, *missing necessary information, discard*)
3) It is easy to turn a canine into the perfect companion because much commodity forms are available. (s^R=1.00)
5) It is easy to turn a canine into perfect companion because of much commodity (s^R=0.74)
~~6) It is easy to turn a canine into a perfect companion because so much.~~ (s^R=0.68, *missing some information, discard*)

Step 4: Valid Questions

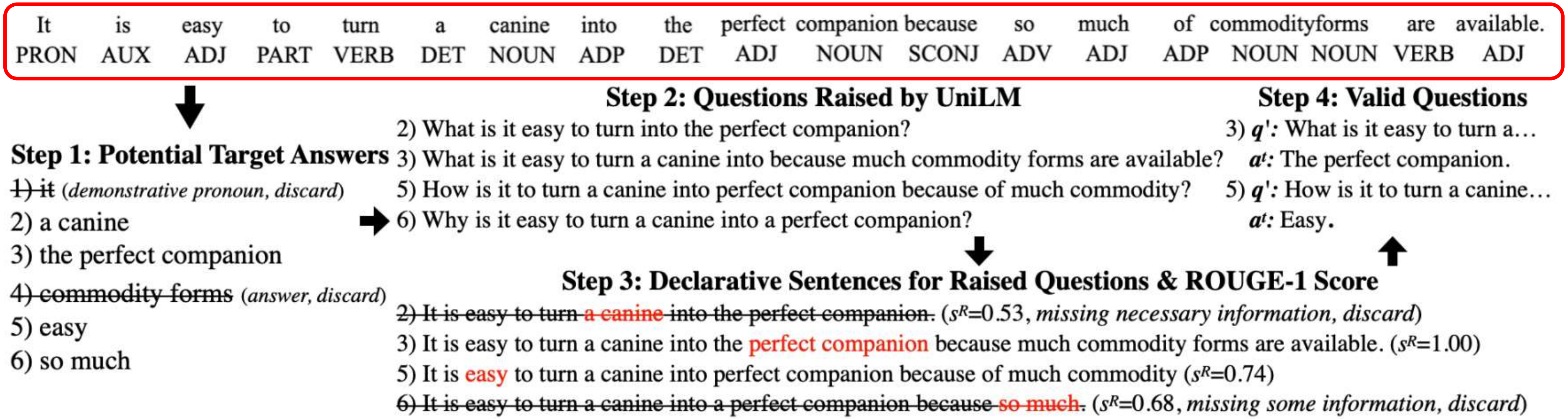
3) ***q*'*:*** What is it easy to turn a…
a*^t*: The perfect companion.
5) ***q*'*:*** How is it to turn a canine…
a*^t*: Easy.

(b) Wh-Question Sentence Generation

Step1: Choose proper target answers

Step2: Produce the corresponding questions

Question Sentence Generation (Wh-question)



It	is	easy	to	turn	a	canine	into	the	perfect	companion	because	so	much	of	commodity	forms	are	available.
PRON	AUX	ADJ	PART	VERB	DET	NOUN	ADP	DET	ADJ	NOUN	SCONJ	ADV	ADJ	ADP	NOUN	NOUN	VERB	ADJ

Step 1: Potential Target Answers

~~1) it~~ *(demonstrative pronoun, discard)*
2) a canine
3) the perfect companion
~~4) commodity forms~~ *(answer, discard)*
5) easy
6) so much

Step 2: Questions Raised by UniLM

2) What is it easy to turn into the perfect companion?
3) What is it easy to turn a canine into because much commodity forms are available?
5) How is it to turn a canine into perfect companion because of much commodity?
6) Why is it easy to turn a canine into a perfect companion?

Step 3: Declarative Sentences for Raised Questions & ROUGE-1 Score

~~2) It is easy to turn a canine into the perfect companion.~~ (s^R=0.53, *missing necessary information, discard*)
3) It is easy to turn a canine into the perfect companion because much commodity forms are available. (s^R=1.00)
5) It is easy to turn a canine into perfect companion because of much commodity (s^R=0.74)
~~6) It is easy to turn a canine into a perfect companion because so much.~~ (s^R=0.68, *missing some information, discard*)

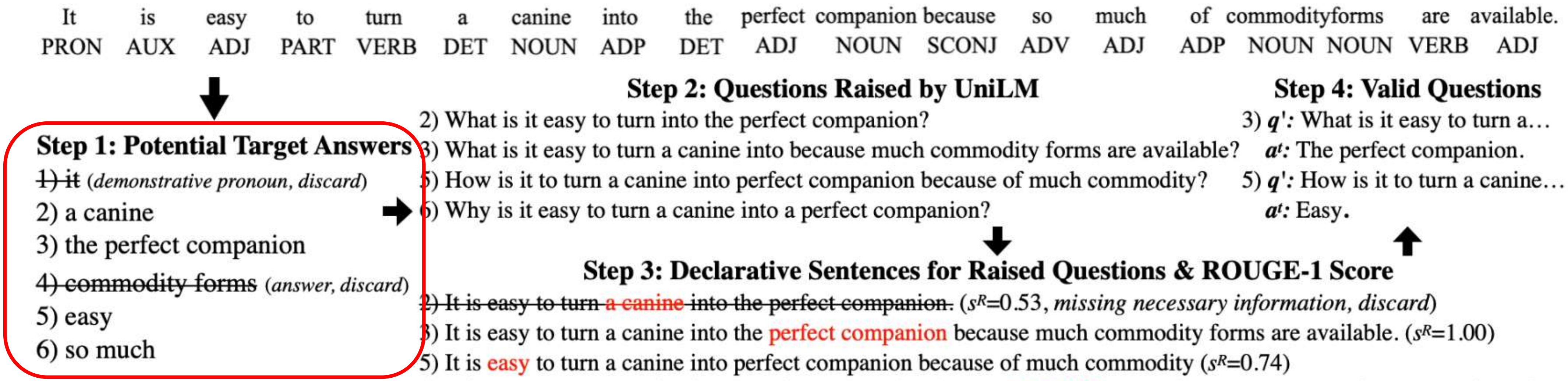
Step 4: Valid Questions

3) ***q'*:** What is it easy to turn a…
***a^t*:** The perfect companion.
5) ***q'*:** How is it to turn a canine…
***a^t*:** Easy.

(b) Wh-Question Sentence Generation

Step1:
POS tagging and dependency parsing

Question Sentence Generation (Wh-question)



It	is	easy	to	turn	a	canine	into	the	perfect	companion	because	so	much	of	commodity	forms	are	available.
PRON	AUX	ADJ	PART	VERB	DET	NOUN	ADP	DET	ADJ	NOUN	SCONJ	ADV	ADJ	ADP	NOUN	NOUN	VERB	ADJ

Step 1: Potential Target Answers

~~1) it~~ *(demonstrative pronoun, discard)*

2) a canine

3) the perfect companion

~~4) commodity forms~~ *(answer, discard)*

5) easy

6) so much

Step 2: Questions Raised by UniLM

2) What is it easy to turn into the perfect companion?

3) What is it easy to turn a canine into because much commodity forms are available?

5) How is it to turn a canine into perfect companion because of much commodity?

6) Why is it easy to turn a canine into a perfect companion?

Step 3: Declarative Sentences for Raised Questions & ROUGE-1 Score

~~2) It is easy to turn a canine into the perfect companion.~~ (s^R=0.53, *missing necessary information, discard*)

3) It is easy to turn a canine into the perfect companion because much commodity forms are available. (s^R=1.00)

5) It is easy to turn a canine into perfect companion because of much commodity (s^R=0.74)

~~6) It is easy to turn a canine into a perfect companion because so much.~~ (s^R=0.68, *missing some information, discard*)

Step 4: Valid Questions

3) q': What is it easy to turn a…

a^t: The perfect companion.

5) q': How is it to turn a canine…

a^t: Easy.

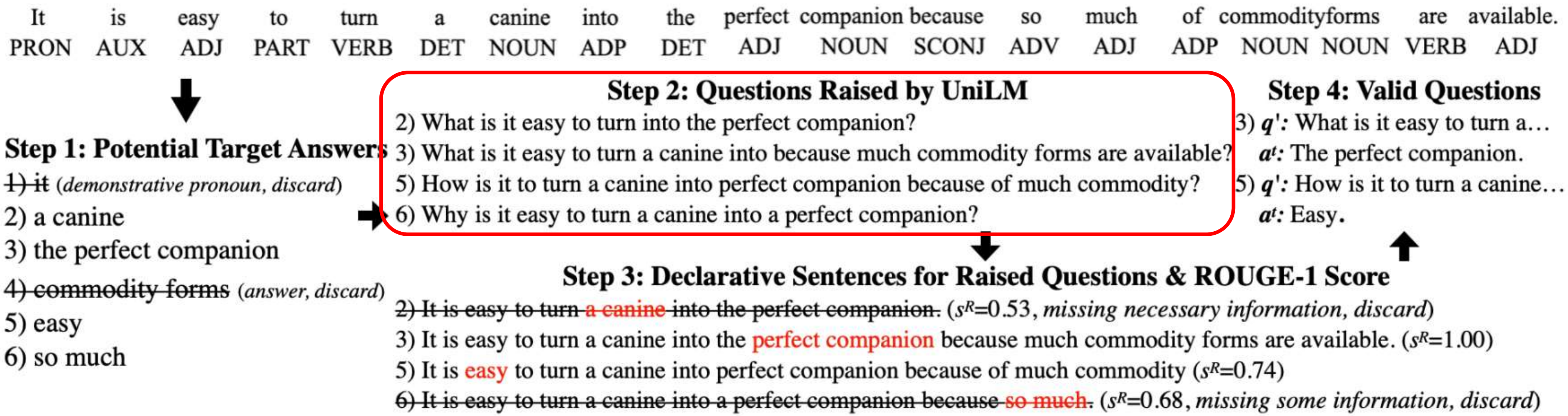
(b) Wh-Question Sentence Generation

Step1:

Extract noun phrases and adjective phrases

(some unsuitable answers, such as phrases, with demonstrative pronouns and the original answer, are excluded from potential target answers)

Question Sentence Generation (Wh-question)



It	is	easy	to	turn	a	canine	into	the	perfect	companion	because	so	much	of	commodity	forms	are	available.
PRON	AUX	ADJ	PART	VERB	DET	NOUN	ADP	DET	ADJ	NOUN	SCONJ	ADV	ADJ	ADP	NOUN	NOUN	VERB	ADJ

Step 1: Potential Target Answers

~~1) it~~ (*demonstrative pronoun, discard*)
2) a canine
3) the perfect companion
~~4) commodity forms~~ (*answer, discard*)
5) easy
6) so much

Step 2: Questions Raised by UniLM

2) What is it easy to turn into the perfect companion?
3) What is it easy to turn a canine into because much commodity forms are available?
5) How is it to turn a canine into perfect companion because of much commodity?
6) Why is it easy to turn a canine into a perfect companion?

Step 3: Declarative Sentences for Raised Questions & ROUGE-1 Score

~~2) It is easy to turn a canine into the perfect companion.~~ (s^R=0.53, *missing necessary information, discard*)
3) It is easy to turn a canine into the perfect companion because much commodity forms are available. (s^R=1.00)
5) It is easy to turn a canine into perfect companion because of much commodity (s^R=0.74)
~~6) It is easy to turn a canine into a perfect companion because so much.~~ (s^R=0.68, *missing some information, discard*)

Step 4: Valid Questions

3) ***q*'*:*** What is it easy to turn a…
a*ᵗ*: The perfect companion.
5) ***q*'*:*** How is it to turn a canine…
a*ᵗ*: Easy.

(b) Wh-Question Sentence Generation

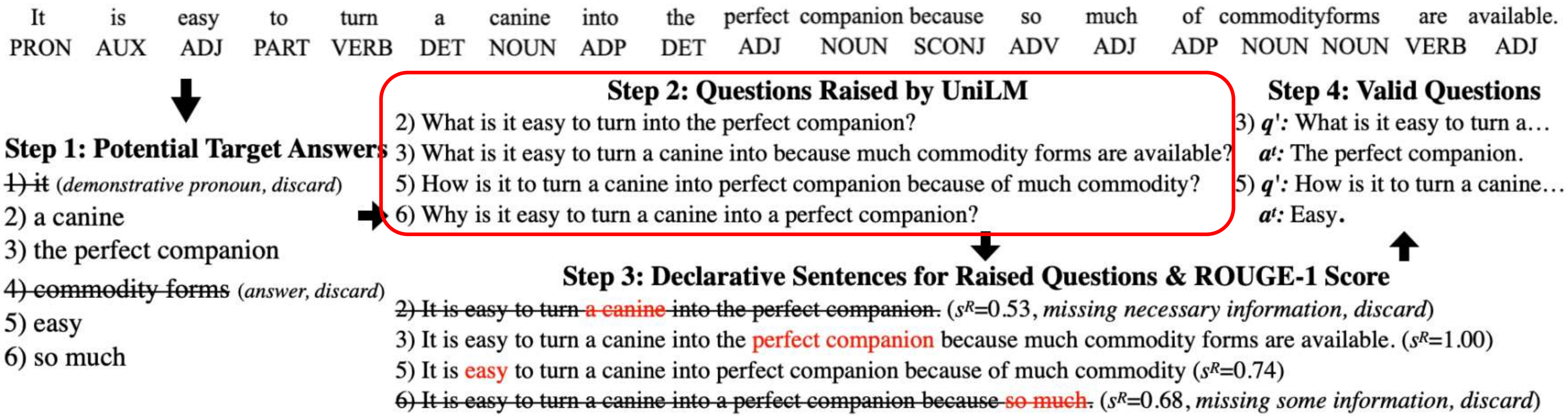
Step2:

Use UniLM to raise a reasonable question for each target answer

(UniLM is a pretrained language model which is good at generative QA/summarization/**question generation**)

[1] UniLM简介: https://cloud.tencent.com/developer/article/1573393

Question Sentence Generation (Wh-question)



It	is	easy	to	turn	a	canine	into	the	perfect	companion	because	so	much	of	commodity	forms	are	available.
PRON	AUX	ADJ	PART	VERB	DET	NOUN	ADP	DET	ADJ	NOUN	SCONJ	ADV	ADJ	ADP	NOUN	NOUN	VERB	ADJ

Step 1: Potential Target Answers

~~1) it~~ (*demonstrative pronoun, discard*)
2) a canine
3) the perfect companion
~~4) commodity forms~~ (*answer, discard*)
5) easy
6) so much

Step 2: Questions Raised by UniLM

2) What is it easy to turn into the perfect companion?
3) What is it easy to turn a canine into because much commodity forms are available?
5) How is it to turn a canine into perfect companion because of much commodity?
6) Why is it easy to turn a canine into a perfect companion?

Step 3: Declarative Sentences for Raised Questions & ROUGE-1 Score

~~2) It is easy to turn a canine into the perfect companion.~~ (s^R=0.53, *missing necessary information, discard*)
3) It is easy to turn a canine into the perfect companion because much commodity forms are available. (s^R=1.00)
5) It is easy to turn a canine into perfect companion because of much commodity (s^R=0.74)
~~6) It is easy to turn a canine into a perfect companion because so much.~~ (s^R=0.68, *missing some information, discard*)

Step 4: Valid Questions

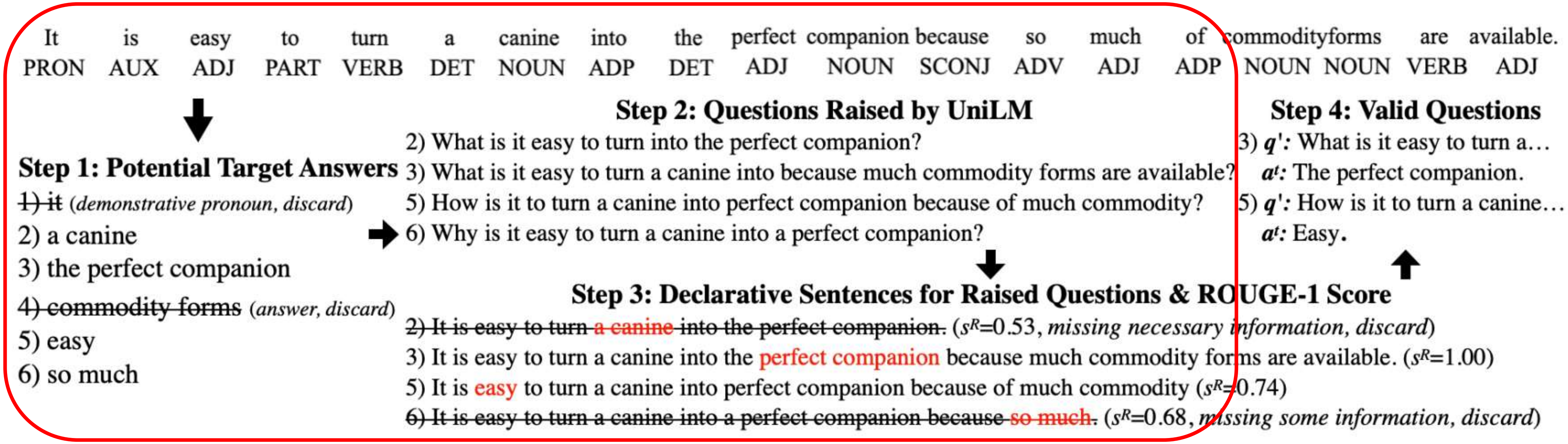
3) ***q*'**: What is it easy to turn a…
a^t: The perfect companion.
5) ***q*'**: How is it to turn a canine…
a^t: Easy.

(b) Wh-Question Sentence Generation

Problem:
Not every question generated by UniLM is reasonable
(Unreasonable question may lead to potential false positive issues)

[1] UniLM简介: https://cloud.tencent.com/developer/article/1573393

Question Sentence Generation (Wh-question)



It	is	easy	to	turn	a	canine	into	the	perfect	companion	because	so	much	of	commodity	forms	are	available.
PRON	AUX	ADJ	PART	VERB	DET	NOUN	ADP	DET	ADJ	NOUN	SCONJ	ADV	ADJ	ADP	NOUN	NOUN	VERB	ADJ

Step 1: Potential Target Answers

~~1) it~~ *(demonstrative pronoun, discard)*
2) a canine
3) the perfect companion
~~4) commodity forms~~ *(answer, discard)*
5) easy
6) so much

Step 2: Questions Raised by UniLM

2) What is it easy to turn into the perfect companion?
3) What is it easy to turn a canine into because much commodity forms are available?
5) How is it to turn a canine into perfect companion because of much commodity?
6) Why is it easy to turn a canine into a perfect companion?

Step 3: Declarative Sentences for Raised Questions & ROUGE-1 Score

~~2) It is easy to turn a canine into the perfect companion.~~ (s^R=0.53, *missing necessary information, discard*)
3) It is easy to turn a canine into the perfect companion because much commodity forms are available. (s^R=1.00)
5) It is easy to turn a canine into perfect companion because of much commodity (s^R=0.74)
~~6) It is easy to turn a canine into a perfect companion because so much.~~ (s^R=0.68, *missing some information, discard*)

Step 4: Valid Questions

3) ***q*'**: What is it easy to turn a…
***a*^{t}**: The perfect companion.
5) ***q*'**: How is it to turn a canine…
***a*^{t}**: Easy.

(b) Wh-Question Sentence Generation

Solution:

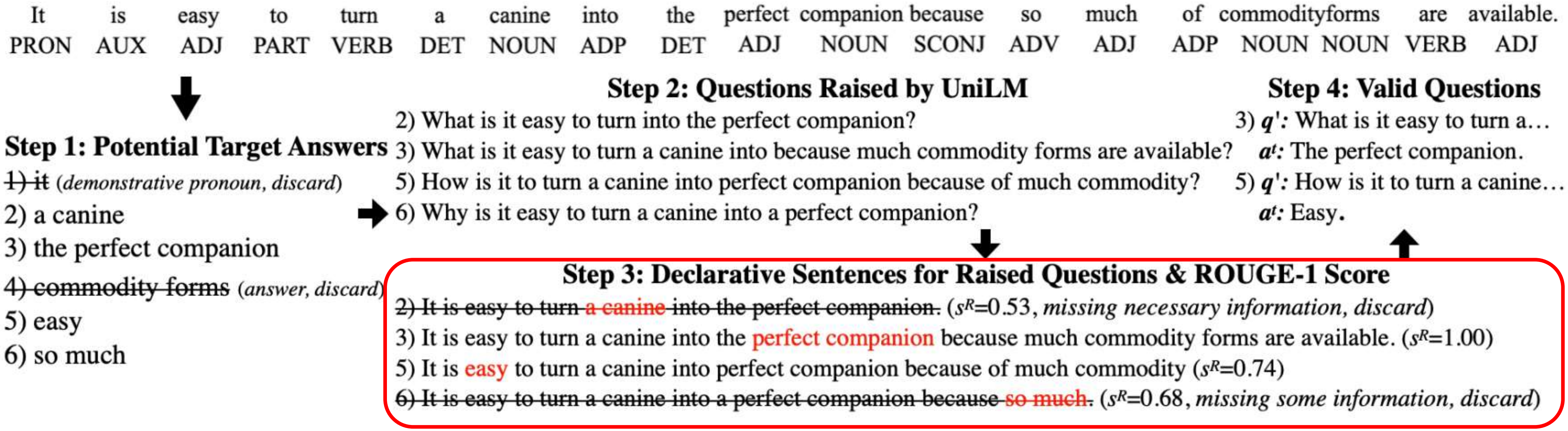
Use target answer and question raised by UniLM to produce new declarative sentence(knowledge)

Then compute the similarity between the generated knowledge and original knowledge

(somewhat similar to the idea of back-translation we mentioned before)

[1] UniLM简介: https://cloud.tencent.com/developer/article/1573393

Question Sentence Generation (Wh-question)



It	is	easy	to	turn	a	canine	into	the	perfect	companion	because	so	much	of	commodity	forms	are	available.
PRON	AUX	ADJ	PART	VERB	DET	NOUN	ADP	DET	ADJ	NOUN	SCONJ	ADV	ADJ	ADP	NOUN	NOUN	VERB	ADJ

Step 1: Potential Target Answers

~~1) it~~ (*demonstrative pronoun, discard*)

2) a canine

3) the perfect companion

~~4) commodity forms~~ (*answer, discard*)

5) easy

6) so much

Step 2: Questions Raised by UniLM

2) What is it easy to turn into the perfect companion?

3) What is it easy to turn a canine into because much commodity forms are available?

5) How is it to turn a canine into perfect companion because of much commodity?

6) Why is it easy to turn a canine into a perfect companion?

Step 3: Declarative Sentences for Raised Questions & ROUGE-1 Score

~~2) It is easy to turn a canine into the perfect companion.~~ (s^R=0.53, *missing necessary information, discard*)

3) It is easy to turn a canine into the perfect companion because much commodity forms are available. (s^R=1.00)

5) It is easy to turn a canine into perfect companion because of much commodity (s^R=0.74)

~~6) It is easy to turn a canine into a perfect companion because so much.~~ (s^R=0.68, *missing some information, discard*)

Step 4: Valid Questions

3) q': What is it easy to turn a…

a^t: The perfect companion.

5) q': How is it to turn a canine…

a^t: Easy.

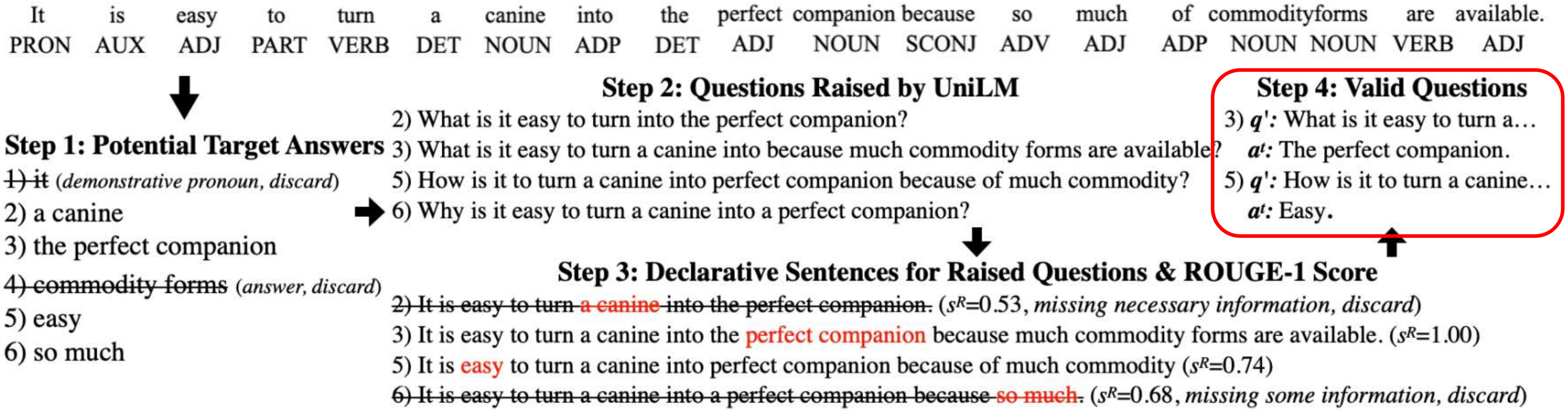
(b) Wh-Question Sentence Generation

Step3:

Filter out questions whose corresponding similarity is greater than 0.7 (in Rouge-1)

(somewhat similar to the idea of back-translation we mentioned before)

Question Sentence Generation (Wh-question)



It	is	easy	to	turn	a	canine	into	the	perfect	companion	because	so	much	of	commodity	forms	are	available.
PRON	AUX	ADJ	PART	VERB	DET	NOUN	ADP	DET	ADJ	NOUN	SCONJ	ADV	ADJ	ADP	NOUN	NOUN	VERB	ADJ

Step 1: Potential Target Answers

~~1) it~~ *(demonstrative pronoun, discard)*
2) a canine
3) the perfect companion
~~4) commodity forms~~ *(answer, discard)*
5) easy
6) so much

Step 2: Questions Raised by UniLM

2) What is it easy to turn into the perfect companion?
3) What is it easy to turn a canine into because much commodity forms are available?
5) How is it to turn a canine into perfect companion because of much commodity?
6) Why is it easy to turn a canine into a perfect companion?

Step 3: Declarative Sentences for Raised Questions & ROUGE-1 Score

~~2) It is easy to turn a canine into the perfect companion.~~ (s^R=0.53, *missing necessary information, discard*)
3) It is easy to turn a canine into the perfect companion because much commodity forms are available. (s^R=1.00)
5) It is easy to turn a canine into perfect companion because of much commodity (s^R=0.74)
~~6) It is easy to turn a canine into a perfect companion because so much.~~ (s^R=0.68, *missing some information, discard*)

Step 4: Valid Questions

3) ***q*'**: What is it easy to turn a…
a^t: The perfect companion.
5) ***q*'**: How is it to turn a canine…
a^t: Easy.

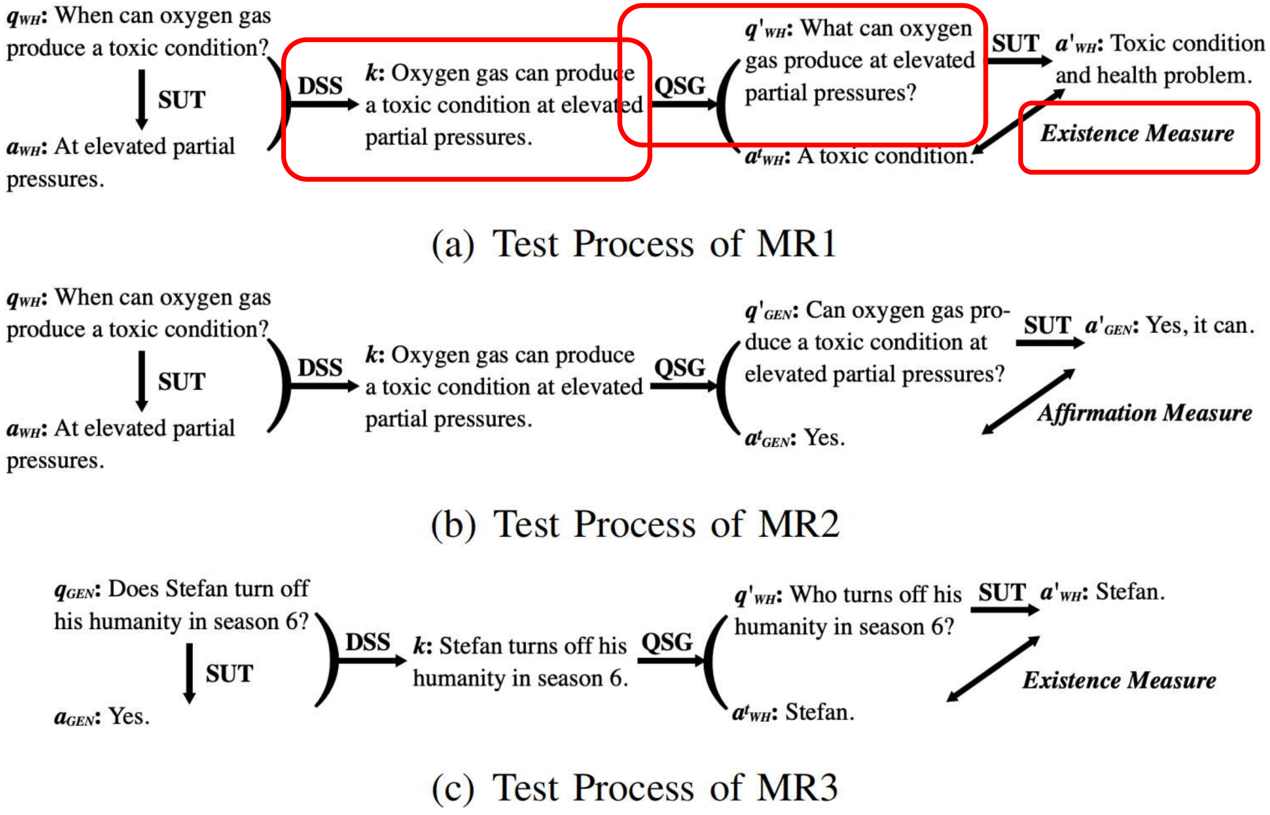
(b) Wh-Question Sentence Generation

Step4: Get Valid Questions

After 3 steps before, we finally generate the wh-questions from the declarative sentence(knowledge)

(and also the corresponding answer directly from potential target answer in step1)

After DSS and QSG …



q_{WH}: When can oxygen gas produce a toxic condition?

SUT

a_{WH}: At elevated partial pressures.

DSS

k: Oxygen gas can produce a toxic condition at elevated partial pressures.

QSG

q'_{WH}: What can oxygen gas produce at elevated partial pressures?

SUT

a'_{WH}: Toxic condition and health problem.

a^t_{WH}: A toxic condition.

Existence Measure

(a) Test Process of MR1

q_{WH}: When can oxygen gas produce a toxic condition?

SUT

a_{WH}: At elevated partial pressures.

DSS

k: Oxygen gas can produce a toxic condition at elevated partial pressures.

QSG

q'_{GEN}: Can oxygen gas produce a toxic condition at elevated partial pressures?

SUT

a'_{GEN}: Yes, it can.

a^t_{GEN}: Yes.

Affirmation Measure

(b) Test Process of MR2

q_{GEN}: Does Stefan turn off his humanity in season 6?

SUT

a_{GEN}: Yes.

DSS

k: Stefan turns off his humanity in season 6.

QSG

q'_{WH}: Who turns off his humanity in season 6?

SUT

a'_{WH}: Stefan.

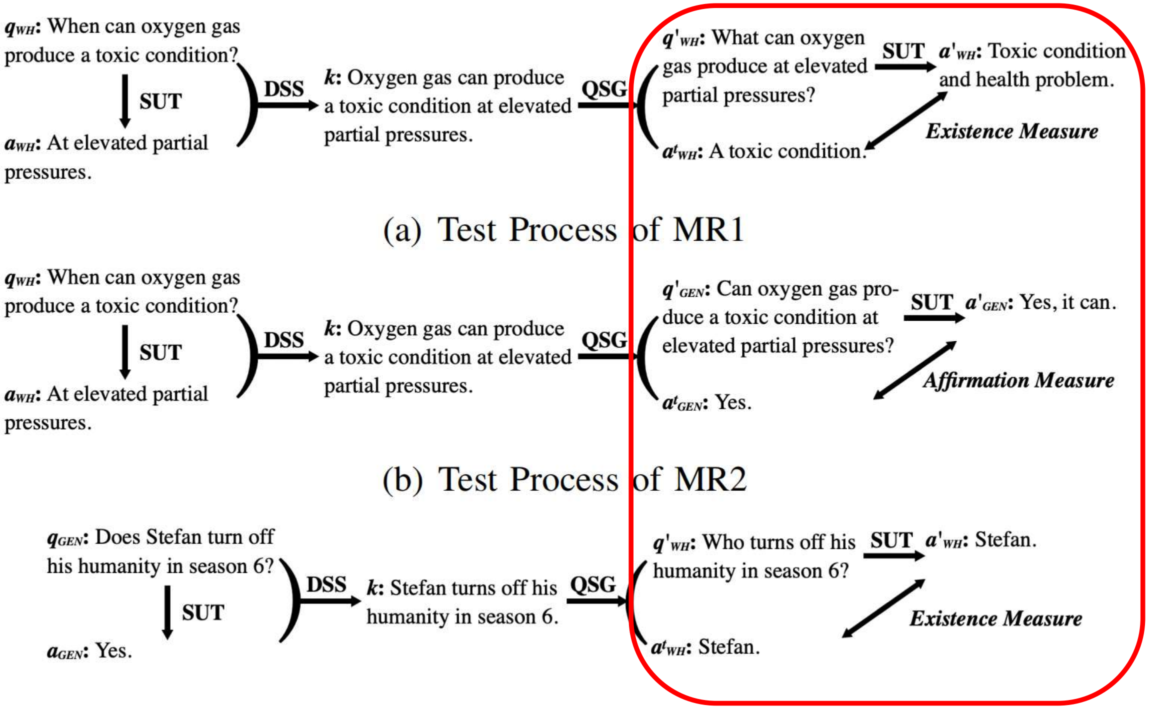
a^t_{WH}: Stefan.

Existence Measure

(c) Test Process of MR3

Fig. 1. Proposed Recursive Metamorphic Relations

Violation Measurement: Existence Measure



q_{WH}: When can oxygen gas produce a toxic condition?

SUT

a_{WH}: At elevated partial pressures.

DSS

k: Oxygen gas can produce a toxic condition at elevated partial pressures.

QSG

q'_{WH}: What can oxygen gas produce at elevated partial pressures?

SUT

a'_{WH}: Toxic condition and health problem.

Existence Measure

a^t_{WH}: A toxic condition.

(a) Test Process of MR1

q_{WH}: When can oxygen gas produce a toxic condition?

SUT

a_{WH}: At elevated partial pressures.

DSS

k: Oxygen gas can produce a toxic condition at elevated partial pressures.

QSG

q'_{GEN}: Can oxygen gas produce a toxic condition at elevated partial pressures?

SUT

a'_{GEN}: Yes, it can.

Affirmation Measure

a^t_{GEN}: Yes.

(b) Test Process of MR2

q_{GEN}: Does Stefan turn off his humanity in season 6?

SUT

a_{GEN}: Yes.

DSS

k: Stefan turns off his humanity in season 6.

QSG

q'_{WH}: Who turns off his humanity in season 6?

SUT

a'_{WH}: Stefan.

Existence Measure

a^t_{WH}: Stefan.

Step1: Discard stop words

Step2: Compute cosine similarity between each word in α'_{WH} and α^t_{WH}

Step3: Average all the word-wise maximum similarity into an overall score to indicate the existence

Step4: Report violation if score < 0.6

Violation Measurement: Existence Measure

If we want to conduct existence measurement on α_{WH}^{t}="the president of Egypt" and α'_{WH}="egyptian president"

(Whether α'_{WH} contains α_{WH}^{t})

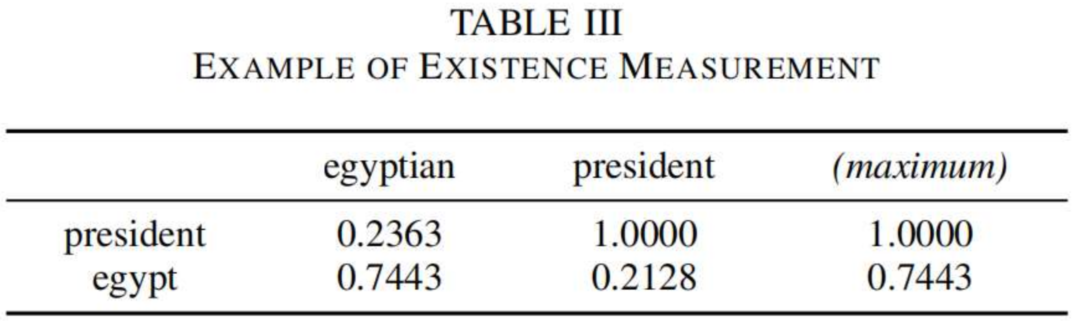


TABLE III
EXAMPLE OF EXISTENCE MEASUREMENT

	egyptian	president	*(maximum)*
president	0.2363	1.0000	1.0000
egypt	0.7443	0.2128	0.7443

Score = (1.0000+0.7443)/2=0.8722 > 0.6

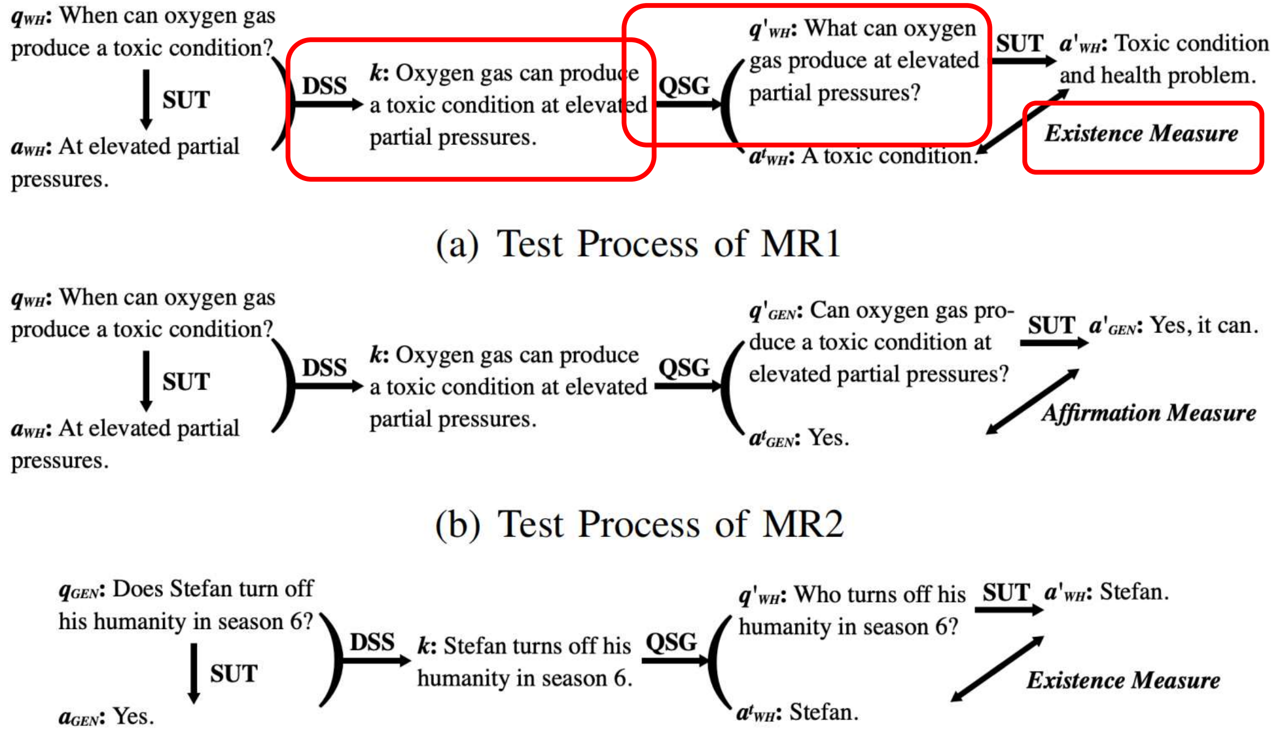
Step1: Discard stop words

Step2: Compute cosine similarity between each word in α'_{WH} and α_{WH}^{t}

Step3: Average all the word-wise maximum similarity into an overall score to indicate the existence

Step4: Report violation if score < 0.6

Implementation


q_{WH}: When can oxygen gas produce a toxic condition?
SUT
a_{WH}: At elevated partial pressures.
DSS
k: Oxygen gas can produce a toxic condition at elevated partial pressures.
QSG
q'_{WH}: What can oxygen gas produce at elevated partial pressures?
SUT
a'_{WH}: Toxic condition and health problem.
Existence Measure
a^{t}_{WH}: A toxic condition.
q_{WH}: When can oxygen gas produce a toxic condition?
SUT
a_{WH}: At elevated partial pressures.
DSS
k: Oxygen gas can produce a toxic condition at elevated partial pressures.
QSG
q'_{GEN}: Can oxygen gas produce a toxic condition at elevated partial pressures?
SUT
a'_{GEN}: Yes, it can.
Affirmation Measure
a^{t}_{GEN}: Yes.
q_{GEN}: Does Stefan turn off his humanity in season 6?
SUT
a_{GEN}: Yes.
DSS
k: Stefan turns off his humanity in season 6.
QSG
q'_{WH}: Who turns off his humanity in season 6?
SUT
a'_{WH}: Stefan.
Existence Measure
a^{t}_{WH}: Stefan.

(a) Test Process of MR1

(b) Test Process of MR2

(c) Test Process of MR3

Fig. 1. Proposed Recursive Metamorphic Relations

Module Definition:

SUT: software under test(QA software)

DSS: Declarative Sentence Synthesis

QSG: Question Sentence Generation

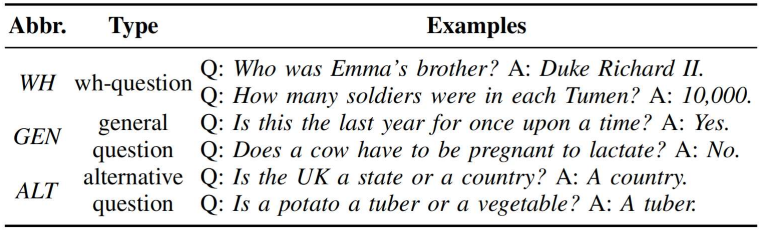
k: Knowledge(Declarative Sentence)

Question Definition:

WH: wh-question(what when who how many)

GEN: general question

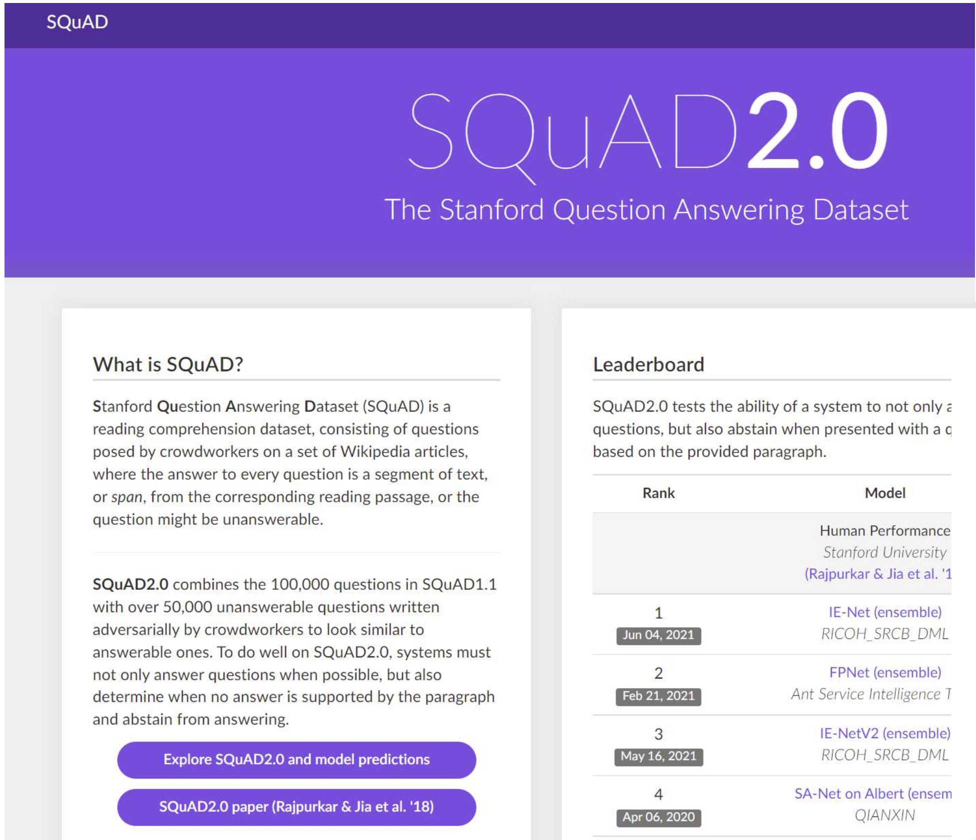
ALT: alternative question



Abbr.	Type	Examples
WH	wh-question	Q: *Who was Emma's brother?* A: *Duke Richard II.* Q: *How many soldiers were in each Tumen?* A: *10,000.*
GEN	general question	Q: *Is this the last year for once upon a time?* A: *Yes.* Q: *Does a cow have to be pregnant to lactate?* A: *No.*
ALT	alternative question	Q: *Is the UK a state or a country?* A: *A country.* Q: *Is a potato a tuber or a vegetable?* A: *A tuber.*

Experiment

Experiment: SQuAD Dataset



What is SQuAD?

Stanford **Q**uestion **A**nswering **D**ataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or *span*, from the corresponding reading passage, or the question might be unanswerable.

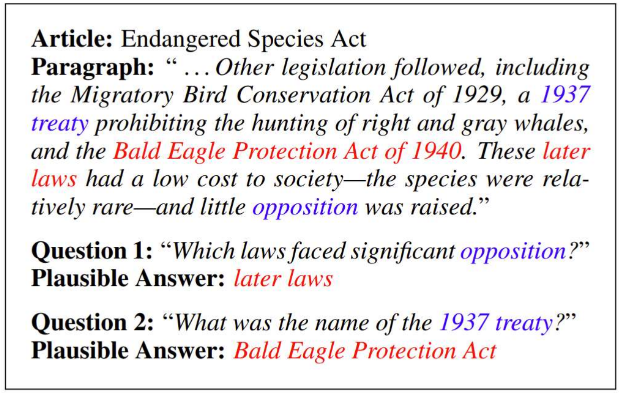
SQuAD2.0 combines the 100,000 questions in SQuAD1.1 with over 50,000 unanswerable questions written adversarially by crowdworkers to look similar to answerable ones. To do well on SQuAD2.0, systems must not only answer questions when possible, but also determine when no answer is supported by the paragraph and abstain from answering.

Explore SQuAD2.0 and model predictions
SQuAD2.0 paper (Rajpurkar & Jia et al. '18)

Leaderboard

SQuAD2.0 tests the ability of a system to not only a
questions, but also abstain when presented with a q
based on the provided paragraph.

Rank	Model
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '1
1 Jun 04, 2021	IE-Net (ensemble) *RICOH_SRCB_DML*
2 Feb 21, 2021	FPNet (ensemble) *Ant Service Intelligence T*
3 May 16, 2021	IE-NetV2 (ensemble) *RICOH_SRCB_DML*
4 Apr 06, 2020	SA-Net on Albert (ensem *QIANXIN*

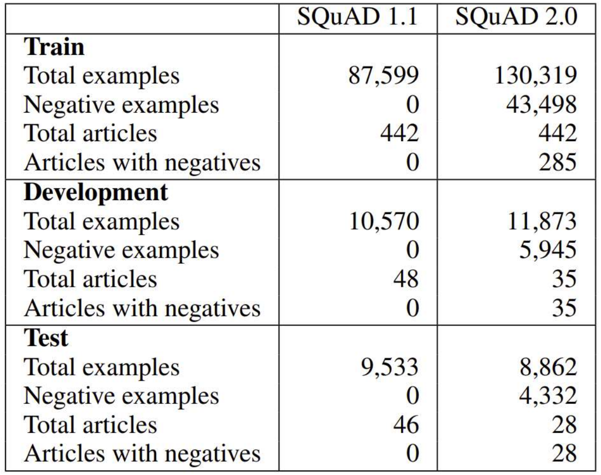


Article: Endangered Species Act
Paragraph: " *. . . Other legislation followed, including the Migratory Bird Conservation Act of 1929, a 1937 treaty prohibiting the hunting of right and gray whales, and the Bald Eagle Protection Act of 1940. These later laws had a low cost to society—the species were relatively rare—and little opposition was raised.*"

Question 1: "*Which laws faced significant opposition?*"
Plausible Answer: *later laws*

Question 2: "*What was the name of the 1937 treaty?*"
Plausible Answer: *Bald Eagle Protection Act*

Figure 1: Two unanswerable questions written by crowdworkers, along with plausible (but incorrect) answers. Relevant keywords are shown in blue.



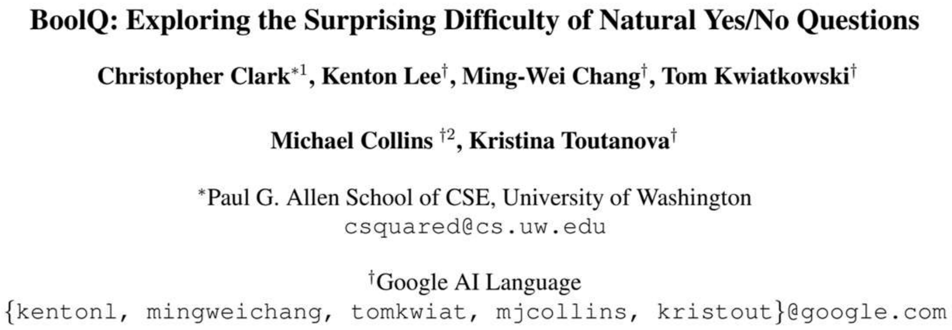
	SQuAD 1.1	SQuAD 2.0
Train		
Total examples	87,599	130,319
Negative examples	0	43,498
Total articles	442	442
Articles with negatives	0	285
Development		
Total examples	10,570	11,873
Negative examples	0	5,945
Total articles	48	35
Articles with negatives	0	35
Test		
Total examples	9,533	8,862
Negative examples	0	4,332
Total articles	46	28
Articles with negatives	0	28

Table 2: Dataset statistics of SQuAD 2.0, compared to the previous SQuAD 1.1.

[1] https://rajpurkar.github.io/SQuAD-explorer/
[2] https://zhuanlan.zhihu.com/p/137828922

Experiment: BoolQ Dataset

BoolQ is a dataset totally composed of general questions obtained from Google Search queries and paired with passages from Wikipedia that are considered sufficient to deduce the answer. The answer is expected to be either "Yes" or "No" (or sentences with similar meanings [13]). It has 9.4k training samples and 3.3k test samples.



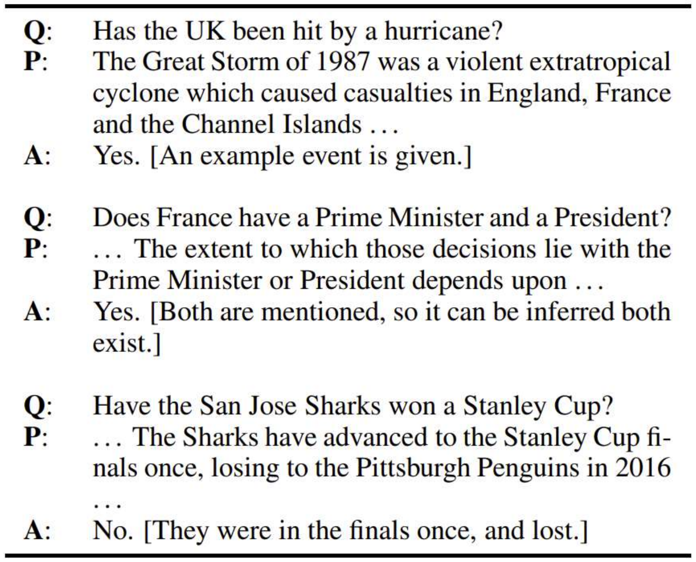
BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions

Christopher Clark[*1], **Kenton Lee**[†], **Ming-Wei Chang**[†], **Tom Kwiatkowski**[†]

Michael Collins [†2], **Kristina Toutanova**[†]

[*]Paul G. Allen School of CSE, University of Washington
csquared@cs.uw.edu

[†]Google AI Language
{kentonl, mingweichang, tomkwiat, mjcollins, kristout}@google.com



Q: Has the UK been hit by a hurricane?
P: The Great Storm of 1987 was a violent extratropical cyclone which caused casualties in England, France and the Channel Islands …
A: Yes. [An example event is given.]

Q: Does France have a Prime Minister and a President?
P: … The extent to which those decisions lie with the Prime Minister or President depends upon …
A: Yes. [Both are mentioned, so it can be inferred both exist.]

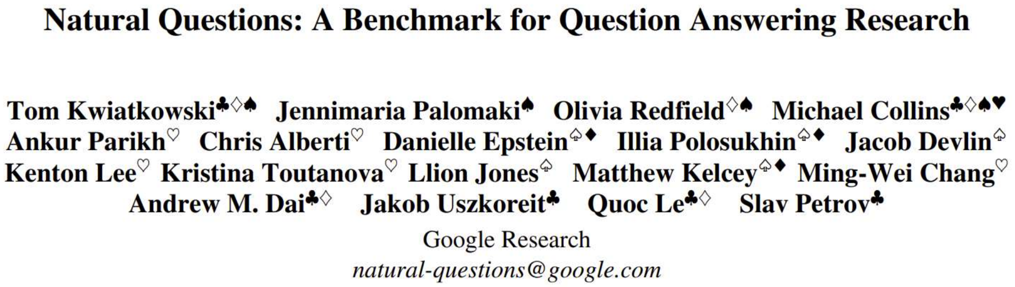
Q: Have the San Jose Sharks won a Stanley Cup?
P: … The Sharks have advanced to the Stanley Cup finals once, losing to the Pittsburgh Penguins in 2016 …
A: No. [They were in the finals once, and lost.]

Figure 1: Example yes/no questions from the BoolQ dataset. Each example consists of a question (**Q**), an excerpt from a passage (**P**), and an answer (**A**) with an explanation added for clarity.

[1] https://github.com/google-research-datasets/boolean-questions
[2] https://paperswithcode.com/dataset/boolq

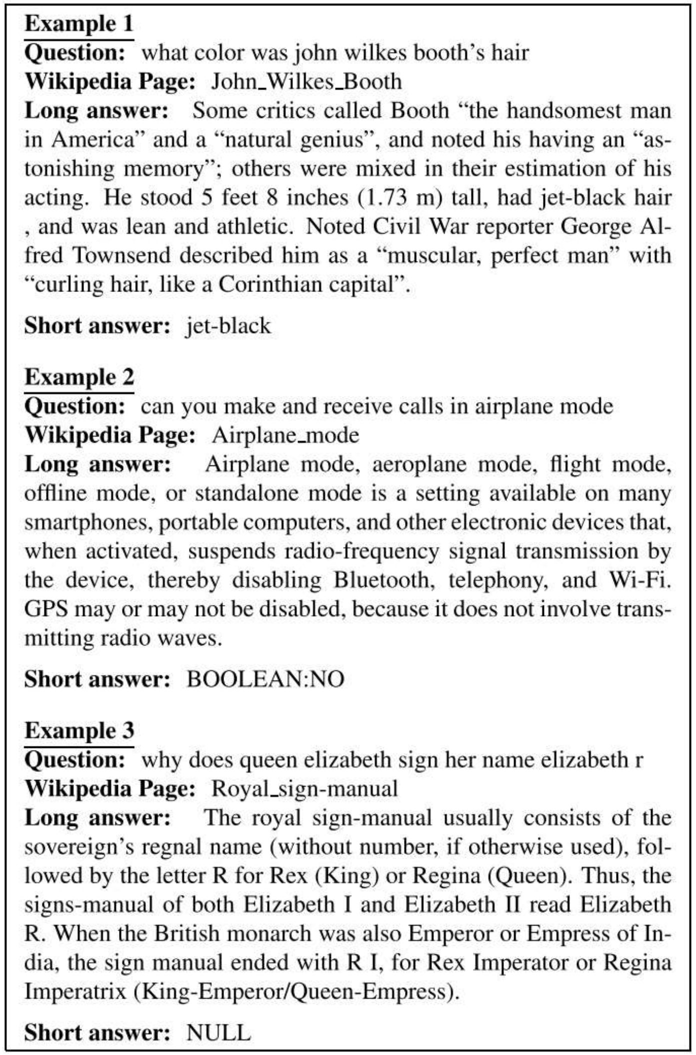
Experiment: NatQA Dataset

NatQA is one abstractive QA dataset, which means it requires the model to return answers that are not mere substrings of the reference passage. We use the version provided by UnifiedQA where each question is appended with a reference passage. It includes 98k wh-questions and 299 general and alternative questions[5], which are then divided into 97k training samples and 11k test samples.



Natural Questions: A Benchmark for Question Answering Research

Tom Kwiatkowski[♣◇♠] **Jennimaria Palomaki**[♠] **Olivia Redfield**[◇♠] **Michael Collins**[♣◇♠♥]
Ankur Parikh[♡] **Chris Alberti**[♡] **Danielle Epstein**[♤♦] **Illia Polosukhin**[♤♦] **Jacob Devlin**[♤]
Kenton Lee[♡] **Kristina Toutanova**[♡] **Llion Jones**[♤] **Matthew Kelcey**[♤♦] **Ming-Wei Chang**[♡]
Andrew M. Dai[♣◇] **Jakob Uszkoreit**[♣] **Quoc Le**[♣◇] **Slav Petrov**[♣]
Google Research
natural-questions@google.com



Example 1
Question: what color was john wilkes booth's hair
Wikipedia Page: John_Wilkes_Booth
Long answer: Some critics called Booth "the handsomest man in America" and a "natural genius", and noted his having an "astonishing memory"; others were mixed in their estimation of his acting. He stood 5 feet 8 inches (1.73 m) tall, had jet-black hair , and was lean and athletic. Noted Civil War reporter George Alfred Townsend described him as a "muscular, perfect man" with "curling hair, like a Corinthian capital".
Short answer: jet-black

Example 2
Question: can you make and receive calls in airplane mode
Wikipedia Page: Airplane_mode
Long answer: Airplane mode, aeroplane mode, flight mode, offline mode, or standalone mode is a setting available on many smartphones, portable computers, and other electronic devices that, when activated, suspends radio-frequency signal transmission by the device, thereby disabling Bluetooth, telephony, and Wi-Fi. GPS may or may not be disabled, because it does not involve transmitting radio waves.
Short answer: BOOLEAN:NO

Example 3
Question: why does queen elizabeth sign her name elizabeth r
Wikipedia Page: Royal_sign-manual
Long answer: The royal sign-manual usually consists of the sovereign's regnal name (without number, if otherwise used), followed by the letter R for Rex (King) or Regina (Queen). Thus, the signs-manual of both Elizabeth I and Elizabeth II read Elizabeth R. When the British monarch was also Emperor or Empress of India, the sign manual ended with R I, for Rex Imperator or Regina Imperatrix (King-Emperor/Queen-Empress).
Short answer: NULL

Figure 1: Example annotations from the corpus.

[1] https://aclanthology.org/Q19-1026.pdf
[2] https://arxiv.org/pdf/2005.00700v3.pdf

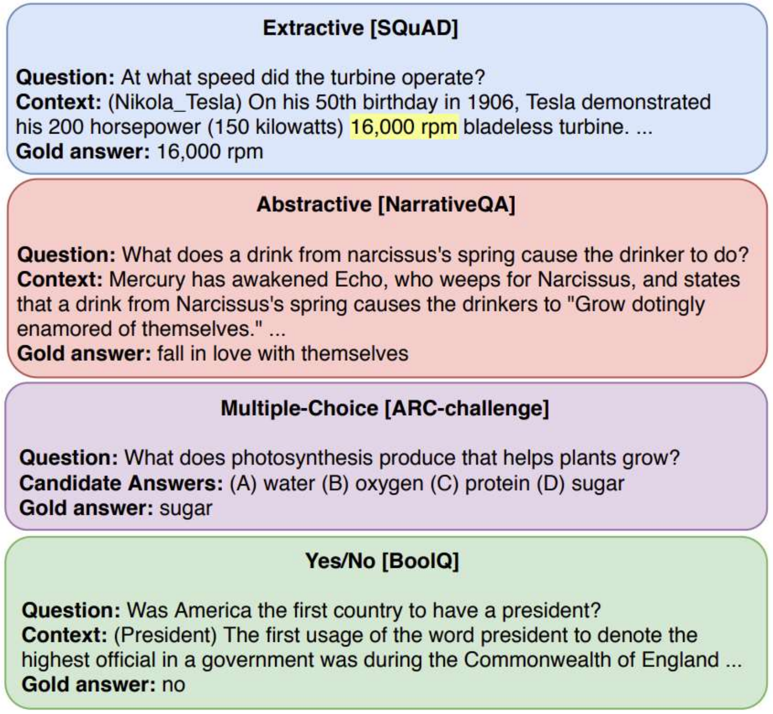
Experiment: Unified QA

Unified QA: Why do we need different QA model?

- Motivation: There are different models for different types of questions. But the ability of inference should be unified.
- Idea: Make a unified QA pretrained model / **Unifying QA solutions**
- Implementation: based on T5
- How to use: Fine-tune the pretrained model into specialized models for better performance on the specific QA tasks

How did they fine-tune the QA model?

- Dataset to fine-tuned: SQuAD2 BoolQ NatQA (236422 samples)
- Pretrained-Model: UnifiedQA (T5)
- Evaluation(When to stop): Exact Match(EM) Score (per 5000 steps)
- Device: RTX3090(24GB memory)



Extractive [SQuAD]

Question: At what speed did the turbine operate?
Context: (Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) 16,000 rpm bladeless turbine. ...
Gold answer: 16,000 rpm

Abstractive [NarrativeQA]

Question: What does a drink from narcissus's spring cause the drinker to do?
Context: Mercury has awakened Echo, who weeps for Narcissus, and states that a drink from Narcissus's spring causes the drinkers to "Grow dotingly enamored of themselves." ...
Gold answer: fall in love with themselves

Multiple-Choice [ARC-challenge]

Question: What does photosynthesis produce that helps plants grow?
Candidate Answers: (A) water (B) oxygen (C) protein (D) sugar
Gold answer: sugar

Yes/No [BoolQ]

Question: Was America the first country to have a president?
Context: (President) The first usage of the word president to denote the highest official in a government was during the Commonwealth of England ...
Gold answer: no

Figure 1: Four formats (color-coded throughout the paper) commonly used for posing questions and answering them: Extractive (EX), Abstractive (AB), Multiple-Choice (MC), and Yes/No (YN). Sample dataset names are shown in square brackets. We study generalization and transfer across these formats.

[1] https://github.com/allenai/unifiedqa
[2] UnifiedQA: https://arxiv.org/pdf/2005.00700.pdf

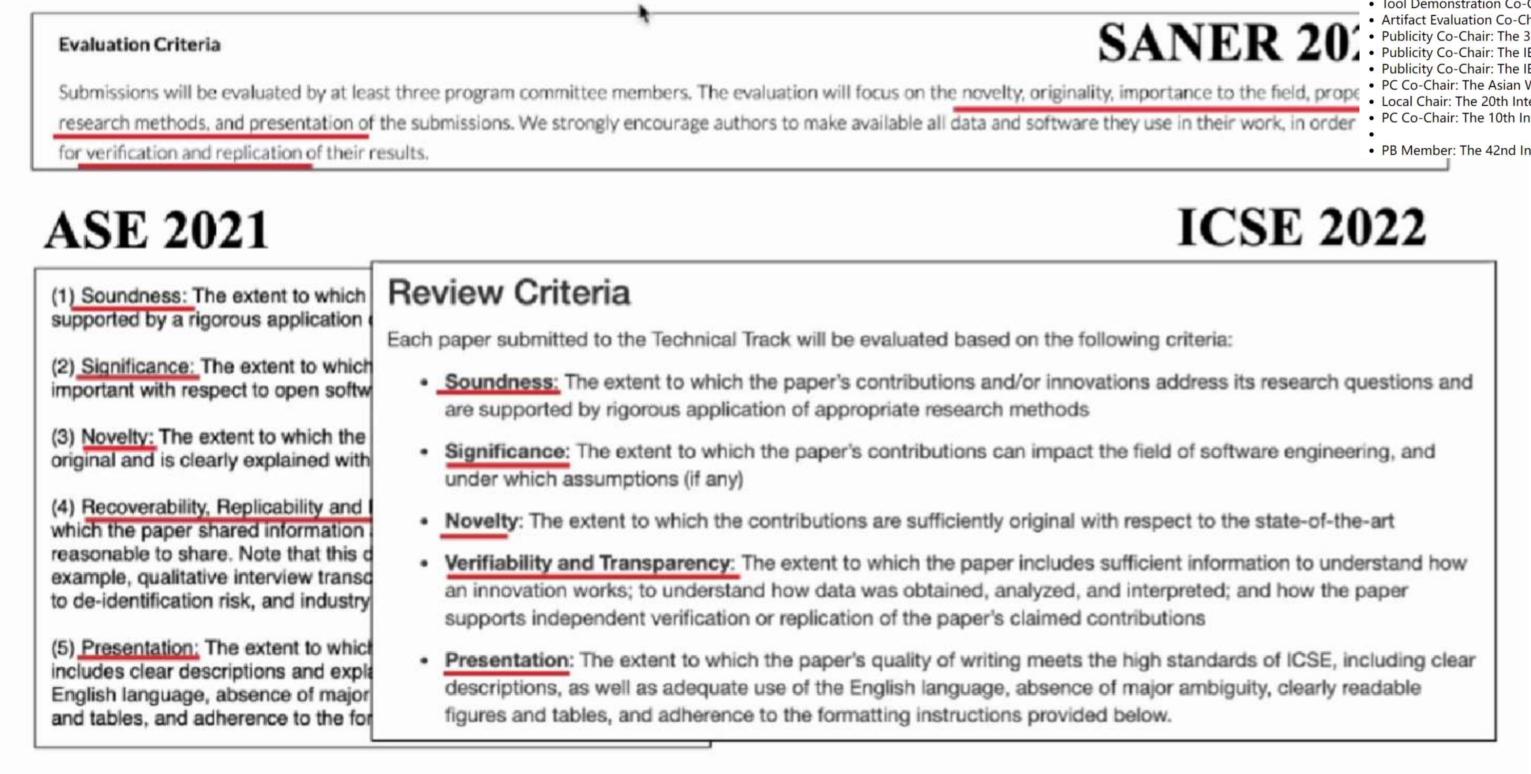
Evaluation

Evaluation

Academic Services

- Program Co-Chair: The 29th IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER 2022)
- Program Co-Chair: The 36th IEEE/ACM International Conference on Automated Software Engineering (ASE 2021)
- General Co-Chair: The 22nd International Systems and Software Product Line Conference (SPLC 2018)
- Steering Committee Member: The International Systems and Software Product Line Conference (SPLC)
- Editorial Board: Software Testing, Verification and Validation (STVR) (since 2019)
- Editorial Board: IEEE Transactions on Software Engineering (TSE) (since 2019)
- Editorial Board: Empirical Software Engineering (EMSE) (since 2020)
- New Ideas and Emerging Results Co-Chair: The 45th International Conference on Software Engineering (ICSE 2023)
- Tool Demonstration Co-Chair: The 44th International Conference on Software Engineering (ICSE 2022)
- Tool Demonstration Co-Chair: The 35th IEEE/ACM International Conference on Automated Software Engineering (ASE 2020)
- Artifact Evaluation Co-Chair: The ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA 2019)
- Publicity Co-Chair: The 36th International Conference on Software Maintenance and Evolution (ICSME 2020)
- Publicity Co-Chair: The IEEE International Conference on Software Testing, Verification, and Validation (ICST 2019)
- Publicity Co-Chair: The IEEE International Working Conference on Source Code Analysis and Manipulation (SCAM 2017)
- PC Co-Chair: The Asian Workshop for ICST 2017
- Local Chair: The 20th International Systems and Software Product Line Conference (SPLC 2016)
- PC Co-Chair: The 10th International Workshop on Automation of Software Testing (AST 2015)
- PB Member: The 42nd International Conference on Software Engineering (ICSE 2020)

评审标准



SANER 202

Evaluation Criteria

Submissions will be evaluated by at least three program committee members. The evaluation will focus on the novelty, originality, importance to the field, prope
research methods, and presentation of the submissions. We strongly encourage authors to make available all data and software they use in their work, in order
for verification and replication of their results.

ASE 2021

(1) Soundness: The extent to which
supported by a rigorous application

(2) Significance: The extent to which
important with respect to open softw

(3) Novelty: The extent to which the
original and is clearly explained with

(4) Recoverability, Replicability and
which the paper shared information
reasonable to share. Note that this d
example, qualitative interview transc
to de-identification risk, and industry

(5) Presentation: The extent to whic
includes clear descriptions and expl
English language, absence of major
and tables, and adherence to the for

ICSE 2022

Review Criteria

Each paper submitted to the Technical Track will be evaluated based on the following criteria:

- **Soundness:** The extent to which the paper's contributions and/or innovations address its research questions and are supported by rigorous application of appropriate research methods
- **Significance:** The extent to which the paper's contributions can impact the field of software engineering, and under which assumptions (if any)
- **Novelty**: The extent to which the contributions are sufficiently original with respect to the state-of-the-art
- **Verifiability and Transparency:** The extent to which the paper includes sufficient information to understand how an innovation works; to understand how data was obtained, analyzed, and interpreted; and how the paper supports independent verification or replication of the paper's claimed contributions
- **Presentation:** The extent to which the paper's quality of writing meets the high standards of ICSE, including clear descriptions, as well as adequate use of the English language, absence of major ambiguity, clearly readable figures and tables, and adherence to the formatting instructions provided below.

Evaluation

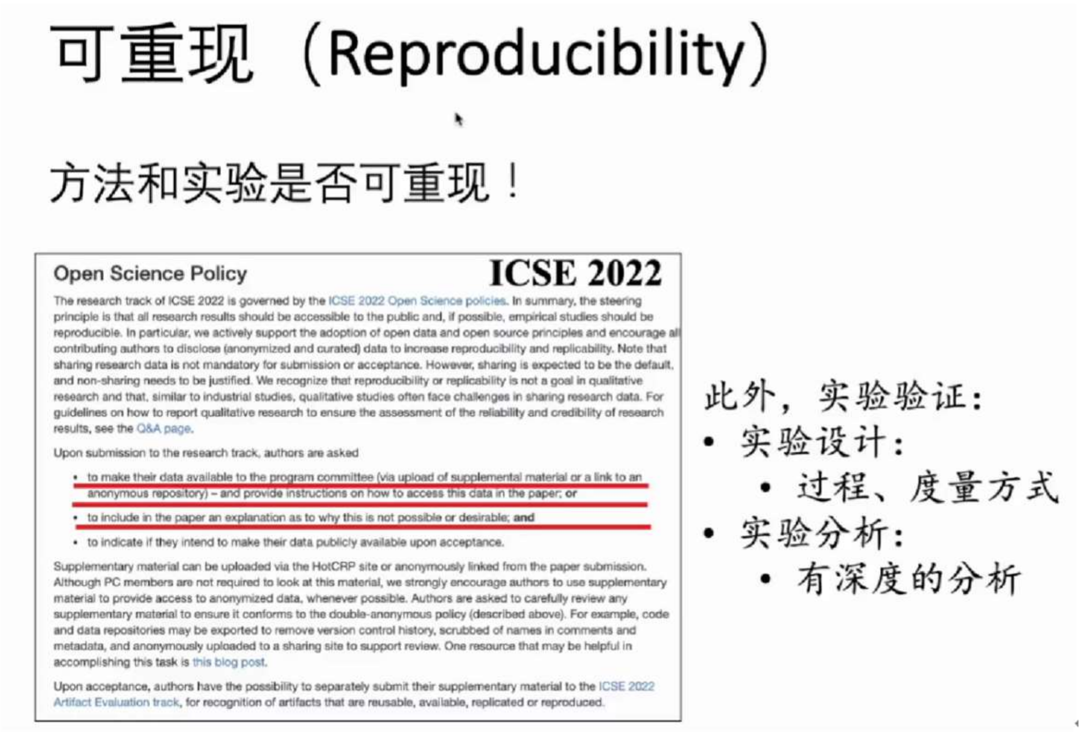
1 Soundness

2 Significance

3 Novelty

4 Reproducibility

5 Presentation



可重现（Reproducibility）

方法和实验是否可重现！

Open Science Policy **ICSE 2022**

The research track of ICSE 2022 is governed by the ICSE 2022 Open Science policies. In summary, the steering principle is that all research results should be accessible to the public and, if possible, empirical studies should be reproducible. In particular, we actively support the adoption of open data and open source principles and encourage all contributing authors to disclose (anonymized and curated) data to increase reproducibility and replicability. Note that sharing research data is not mandatory for submission or acceptance. However, sharing is expected to be the default, and non-sharing needs to be justified. We recognize that reproducibility or replicability is not a goal in qualitative research and that, similar to industrial studies, qualitative studies often face challenges in sharing research data. For guidelines on how to report qualitative research to ensure the assessment of the reliability and credibility of research results, see the Q&A page.

Upon submission to the research track, authors are asked

- to make their data available to the program committee (via upload of supplemental material or a link to an anonymous repository) – and provide instructions on how to access this data in the paper; **or**
- to include in the paper an explanation as to why this is not possible or desirable; **and**
- to indicate if they intend to make their data publicly available upon acceptance.

Supplementary material can be uploaded via the HotCRP site or anonymously linked from the paper submission. Although PC members are not required to look at this material, we strongly encourage authors to use supplementary material to provide access to anonymized data, whenever possible. Authors are asked to carefully review any supplementary material to ensure it conforms to the double-anonymous policy (described above). For example, code and data repositories may be exported to remove version control history, scrubbed of names in comments and metadata, and anonymously uploaded to a sharing site to support review. One resource that may be helpful in accomplishing this task is this blog post.

Upon acceptance, authors have the possibility to separately submit their supplementary material to the ICSE 2022 Artifact Evaluation track, for recognition of artifacts that are reusable, available, replicated or reproduced.

此外，实验验证：

- 实验设计：
  - 过程、度量方式
- 实验分析：
  - 有深度的分析

其实是 evaluation，必须应该写清楚结果和评估对比。但是你在对比其他技术的时候，你也必须说清楚你的方法为什么会 work？审稿人不只是想看到你的方法是work的，他更想看到你对问题和方法的理解的深刻性。比如你的方法有不work的情况吗？在哪些情况上 work？可以推广到其他情况吗？千万不要让审稿人觉得你的方法不可重现，因为这个很容易说服其他审稿人也这么认为，然后就被拒

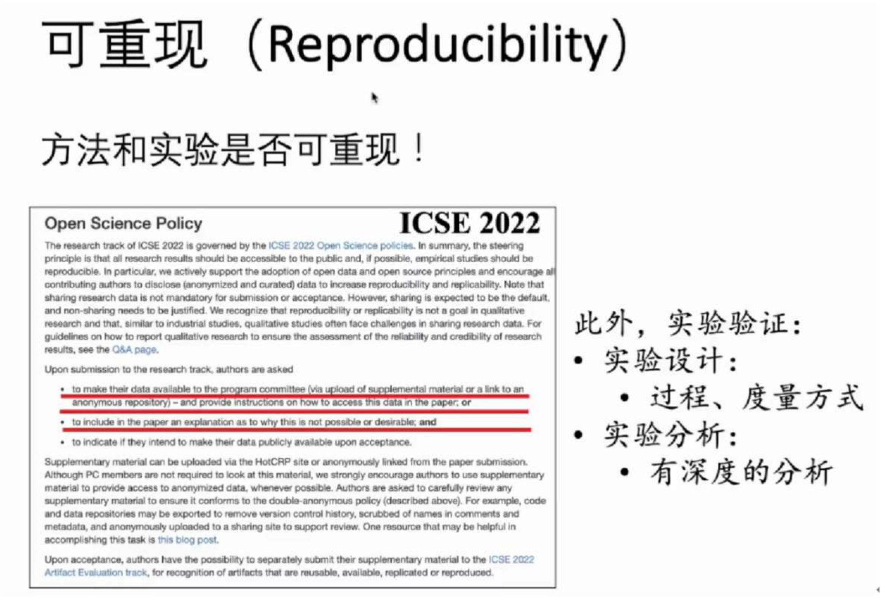
Evaluation: Research Questions

RQ1: The overall effectiveness of *QAASKER*

RQ2: Validity of the revealed violations

RQ3: Types of the revealed true violations

RQ4: Helpfulness to fix the revealed answering issues



可重现（Reproducibility）

方法和实验是否可重现！

Open Science Policy

ICSE 2022

The research track of ICSE 2022 is governed by the ICSE 2022 Open Science policies. In summary, the steering principle is that all research results should be accessible to the public and, if possible, empirical studies should be reproducible. In particular, we actively support the adoption of open data and open source principles and encourage all contributing authors to disclose (anonymized and curated) data to increase reproducibility and replicability. Note that sharing research data is not mandatory for submission or acceptance. However, sharing is expected to be the default, and non-sharing needs to be justified. We recognize that reproducibility or replicability is not a goal in qualitative research and that, similar to industrial studies, qualitative studies often face challenges in sharing research data. For guidelines on how to report qualitative research to ensure the assessment of the reliability and credibility of research results, see the Q&A page.

Upon submission to the research track, authors are asked

- to make their data available to the program committee (via upload of supplemental material or a link to an anonymous repository) – and provide instructions on how to access this data in the paper; **or**
- to include in the paper an explanation as to why this is not possible or desirable; **and**
- to indicate if they intend to make their data publicly available upon acceptance.

Supplementary material can be uploaded via the HotCRP site or anonymously linked from the paper submission. Although PC members are not required to look at this material, we strongly encourage authors to use supplementary material to provide access to anonymized data, whenever possible. Authors are asked to carefully review any supplementary material to ensure it conforms to the double-anonymous policy (described above). For example, code and data repositories may be exported to remove version control history, scrubbed of names in comments and metadata, and anonymously uploaded to a sharing site to support review. One resource that may be helpful in accomplishing this task is this blog post.

Upon acceptance, authors have the possibility to separately submit their supplementary material to the ICSE 2022 Artifact Evaluation track, for recognition of artifacts that are reusable, available, replicated or reproduced.

此外，实验验证：
- 实验设计：
  - 过程、度量方式
- 实验分析：
  - 有深度的分析

其实是 evaluation，必须应该写清楚结果和评估对比。但是你在对比其他技术的时候，你也必须说清楚你的方法为什么会 work？审稿人不只是想看到你的方法是work的，他更想看到你对问题和方法的理解的深刻性。比如你的方法有不work的情况吗？在哪些情况上 work？可以推广到其他情况吗？千万不要让审稿人觉得你的方法不可重现，因为这个很容易说服其他审稿人也这么认为，然后就被拒

RQ1: The overall effectiveness of QAASKER

What is the definition of effectiveness?

- Demonstrates the effectiveness of QAASKER to **reveal the answering issues without the need for the ground truth labels**

Why could MR2 and MR3 find more violation than MR1?

- MR1: WH question=>WH question
- MR2: WH question=>General question
- MR3: General question=>WH question
- **Reason(conjecture): UnifiedQA Overfit the training sample?** Could only pass the test cases whose question is of the frequent types among the training samples from their corresponding dataset
- Indicate the **potential insufficient generalization of UnifiedQA** to figure out the questions of distinct types across datasets

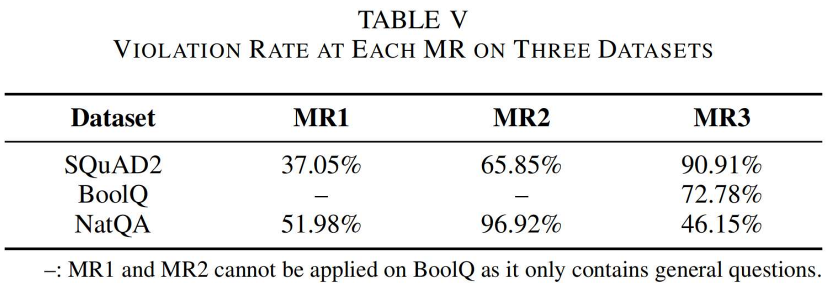
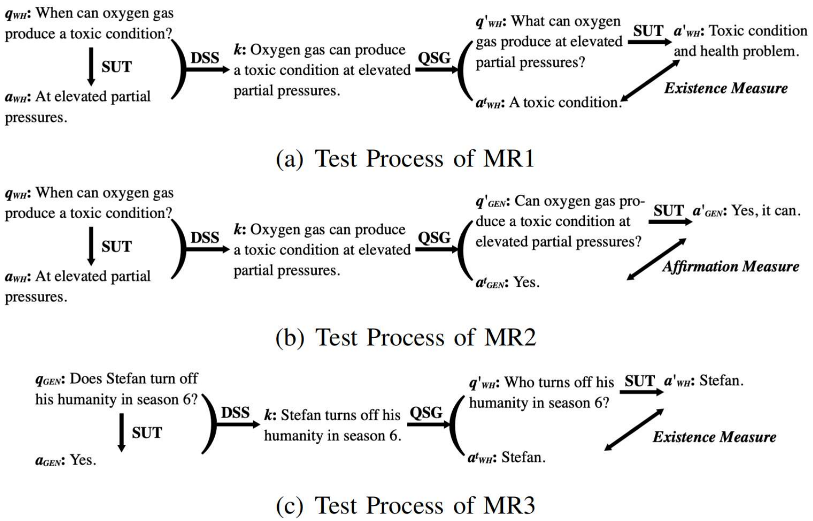


TABLE V
VIOLATION RATE AT EACH MR ON THREE DATASETS

Dataset	MR1	MR2	MR3
SQuAD2	37.05%	65.85%	90.91%
BoolQ	–	–	72.78%
NatQA	51.98%	96.92%	46.15%

–: MR1 and MR2 cannot be applied on BoolQ as it only contains general questions.


q_{WH}: When can oxygen gas produce a toxic condition?
SUT
a_{WH}: At elevated partial pressures.
DSS
k: Oxygen gas can produce a toxic condition at elevated partial pressures.
QSG
q'_{WH}: What can oxygen gas produce at elevated partial pressures?
SUT
a'_{WH}: Toxic condition and health problem.
Existence Measure
a^t_{WH}: A toxic condition.
q_{WH}: When can oxygen gas produce a toxic condition?
SUT
a_{WH}: At elevated partial pressures.
DSS
k: Oxygen gas can produce a toxic condition at elevated partial pressures.
QSG
q'_{GEN}: Can oxygen gas produce a toxic condition at elevated partial pressures?
SUT
a'_{GEN}: Yes, it can.
Affirmation Measure
a^t_{GEN}: Yes.
q_{GEN}: Does Stefan turn off his humanity in season 6?
SUT
a_{GEN}: Yes.
DSS
k: Stefan turns off his humanity in season 6.
QSG
q'_{WH}: Who turns off his humanity in season 6?
SUT
a'_{WH}: Stefan.
Existence Measure
a^t_{WH}: Stefan.

(a) Test Process of MR1

(b) Test Process of MR2

(c) Test Process of MR3

Fig. 1. Proposed Recursive Metamorphic Relations

RQ2: Validity of the Revealed Violations

What is next after we obtain lots of violations?

- We need to evaluate the validity of the revealed violations
- Valid Violation(factuality): at least one incorrect answer from the source and follow-up question

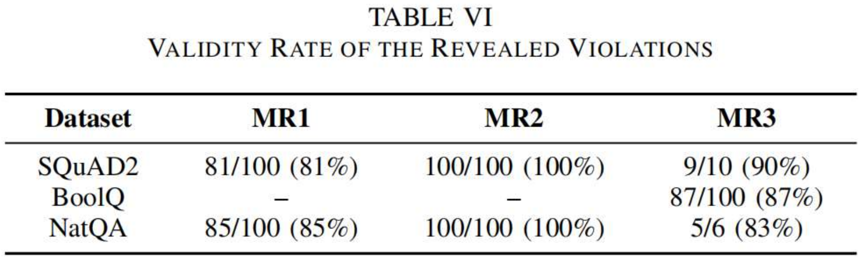


TABLE VI
VALIDITY RATE OF THE REVEALED VIOLATIONS

Dataset	MR1	MR2	MR3
SQuAD2	81/100 (81%)	100/100 (100%)	9/10 (90%)
BoolQ	–	–	87/100 (87%)
NatQA	85/100 (85%)	100/100 (100%)	5/6 (83%)

What may cause the invalid violation (false positive)?

- Generating Questions and Measurement of semantic similarity is challenging

How to detect invalid violation (false positive)?

- Perform the inspection manually and independently (2 people)
- Check the validity of all violations (at most 100)

Meaningful and convincing

RQ3: Types of the Revealed True Violations

What is next after we inspect invalid violations?

- We could further study the details of valid violations
- What we could do and what we couldn’t do

What kinds of violations QAAsKeR could detect?

- <NoAnswer> for answerable questions
- Format mismatch between the answer and the question
- Irrelevant content of the answer
- Grammatical error
- Missing information in the answer

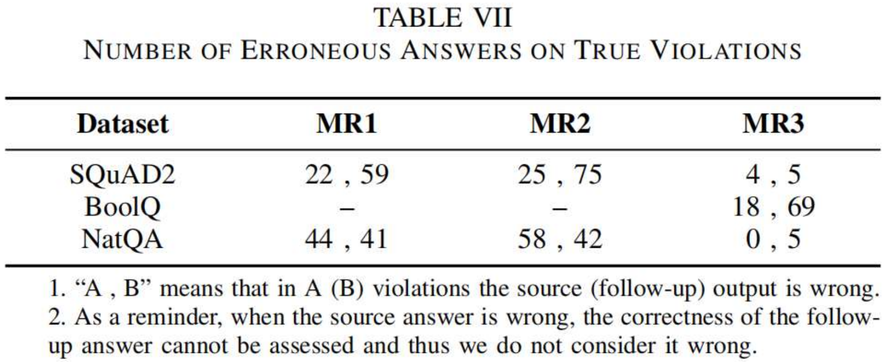


TABLE VII
NUMBER OF ERRONEOUS ANSWERS ON TRUE VIOLATIONS

Dataset	MR1	MR2	MR3
SQuAD2	22 , 59	25 , 75	4 , 5
BoolQ	–	–	18 , 69
NatQA	44 , 41	58 , 42	0 , 5

1. “A , B” means that in A (B) violations the source (follow-up) output is wrong.
2. As a reminder, when the source answer is wrong, the correctness of the follow-up answer cannot be assessed and thus we do not consider it wrong.

RQ3: Types of the Revealed True Violations

What kinds of violations QAAsKeR could detect?

- <NoAnswer> for answerable questions
- Format mismatch between the answer and the question
- Irrelevant content of the answer
- Grammatical error
- Missing information in the answer

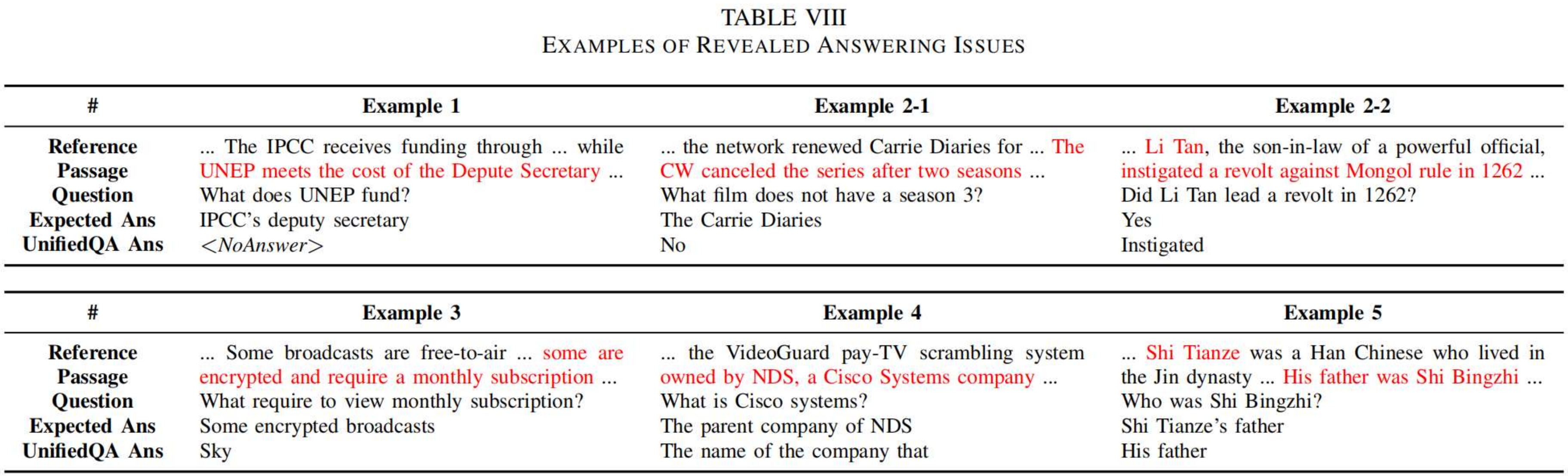


TABLE VIII
EXAMPLES OF REVEALED ANSWERING ISSUES

#	Example 1	Example 2-1	Example 2-2
Reference Passage	... The IPCC receives funding through ... while UNEP meets the cost of the Depute Secretary ...	... the network renewed Carrie Diaries for ... The CW canceled the series after two seasons ...	... Li Tan, the son-in-law of a powerful official, instigated a revolt against Mongol rule in 1262 ...
Question	What does UNEP fund?	What film does not have a season 3?	Did Li Tan lead a revolt in 1262?
Expected Ans	IPCC's deputy secretary	The Carrie Diaries	Yes
UnifiedQA Ans	*<NoAnswer>*	No	Instigated

#	Example 3	Example 4	Example 5
Reference Passage	... Some broadcasts are free-to-air ... some are encrypted and require a monthly subscription ...	... the VideoGuard pay-TV scrambling system owned by NDS, a Cisco Systems company ...	... Shi Tianze was a Han Chinese who lived in the Jin dynasty ... His father was Shi Bingzhi ...
Question	What require to view monthly subscription?	What is Cisco systems?	Who was Shi Bingzhi?
Expected Ans	Some encrypted broadcasts	The parent company of NDS	Shi Tianze's father
UnifiedQA Ans	Sky	The name of the company that	His father

RQ4: Helpfulness to Fix the Revealed Answering Issues

How could the revealed violations help fix the QA software?

- Violation rate about all MRs decreases a lot (**By retrain a new model using training data augmented**)
- Reference-based test metric stays stable (original:0.5574 –> retrained:0.5483)

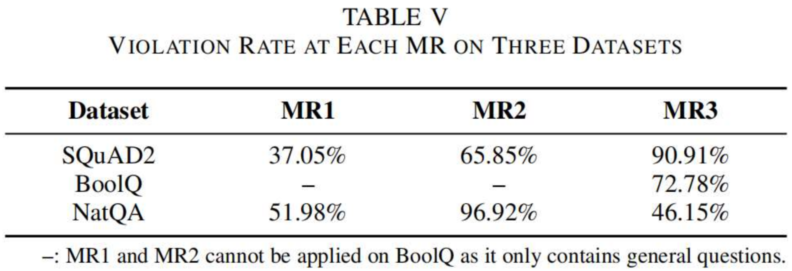


TABLE V
VIOLATION RATE AT EACH MR ON THREE DATASETS

Dataset	MR1	MR2	MR3
SQuAD2	37.05%	65.85%	90.91%
BoolQ	–	–	72.78%
NatQA	51.98%	96.92%	46.15%

–: MR1 and MR2 cannot be applied on BoolQ as it only contains general questions.

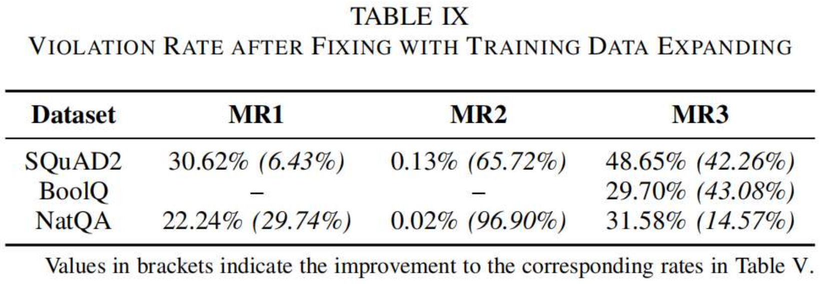
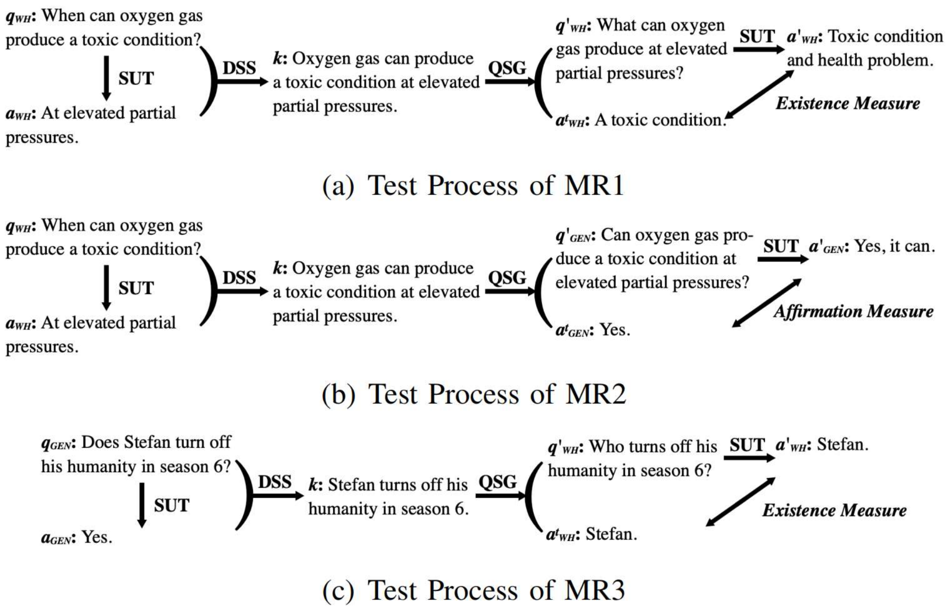


TABLE IX
VIOLATION RATE AFTER FIXING WITH TRAINING DATA EXPANDING

Dataset	MR1	MR2	MR3
SQuAD2	30.62% *(6.43%)*	0.13% *(65.72%)*	48.65% *(42.26%)*
BoolQ	–	–	29.70% *(43.08%)*
NatQA	22.24% *(29.74%)*	0.02% *(96.90%)*	31.58% *(14.57%)*

Values in brackets indicate the improvement to the corresponding rates in Table V.



q_{WH}: When can oxygen gas produce a toxic condition?
SUT
a_{WH}: At elevated partial pressures.
DSS
k: Oxygen gas can produce a toxic condition at elevated partial pressures.
QSG
q'_{WH}: What can oxygen gas produce at elevated partial pressures?
SUT
a'_{WH}: Toxic condition and health problem.
a^t_{WH}: A toxic condition.
Existence Measure
q_{WH}: When can oxygen gas produce a toxic condition?
SUT
a_{WH}: At elevated partial pressures.
DSS
k: Oxygen gas can produce a toxic condition at elevated partial pressures.
QSG
q'_{GEN}: Can oxygen gas produce a toxic condition at elevated partial pressures?
SUT
a'_{GEN}: Yes, it can.
a^t_{GEN}: Yes.
Affirmation Measure
q_{GEN}: Does Stefan turn off his humanity in season 6?
SUT
a_{GEN}: Yes.
DSS
k: Stefan turns off his humanity in season 6.
QSG
q'_{WH}: Who turns off his humanity in season 6?
SUT
a'_{WH}: Stefan.
a^t_{WH}: Stefan.
Existence Measure

(a) Test Process of MR1

(b) Test Process of MR2

(c) Test Process of MR3

Fig. 1. Proposed Recursive Metamorphic Relations

RQ4: Helpfulness to Fix the Revealed Answering Issues

How could the revealed violations help fix the QA software?

- Violation rate about all MRs decreases a lot (**By retrain a new model using training data augmented**)
- Reference-based test metric stays stable (original:0.5574 –> retrained:0.5483)

Other conclusions from the study?

- It is not that easy to repair all the issues revealed by QAASKER
- Retraining a model with the samples expanded **could not solve all** violation problems.
- The proposed MRs are helpful for improving the performance and the **improvement is quite substantial**
- QAASKER is **a testing method**, which is necessary for the reliability checking of QA software output and the in-depth problem revealing of QA software

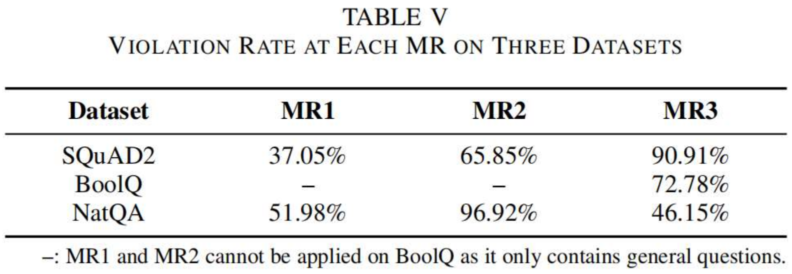


TABLE V
VIOLATION RATE AT EACH MR ON THREE DATASETS

Dataset	MR1	MR2	MR3
SQuAD2	37.05%	65.85%	90.91%
BoolQ	–	–	72.78%
NatQA	51.98%	96.92%	46.15%

–: MR1 and MR2 cannot be applied on BoolQ as it only contains general questions.

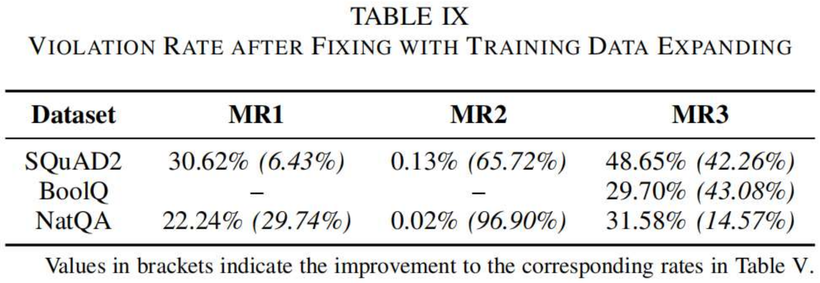


TABLE IX
VIOLATION RATE AFTER FIXING WITH TRAINING DATA EXPANDING

Dataset	MR1	MR2	MR3
SQuAD2	30.62% *(6.43%)*	0.13% *(65.72%)*	48.65% *(42.26%)*
BoolQ	–	–	29.70% *(43.08%)*
NatQA	22.24% *(29.74%)*	0.02% *(96.90%)*	31.58% *(14.57%)*

Values in brackets indicate the improvement to the corresponding rates in Table V.

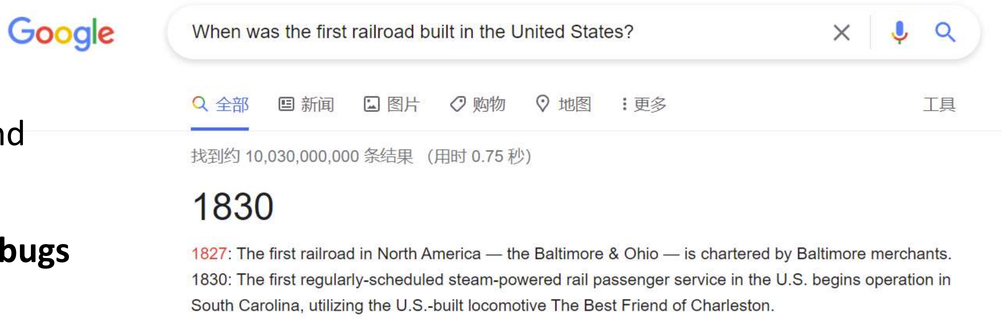
Small-Scale Trial by hand on Google

Why manual study?

- Google Search service can only answer wh-questions
- The returned results vary in forms (e.g., sometimes an exact phrase and occasionally a paragraph with one span in bold)
- This trial could show the **potential of QAASKER to reveal the real-life bugs**

How did they conduct the manual trial?

- Randomly pick 20 wh-questions from MKQA
- Get answers from Google by entering questions manually
- Run QAASKER to generate new questions and their target answers
- Input the new questions as queries and obtain answers
- Finally, 5 of 20 test cases trigger a violation

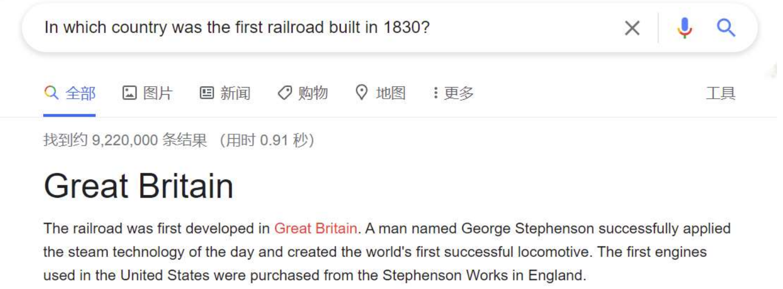


When was the first railroad built in the United States?

找到约 10,030,000,000 条结果 （用时 0.75 秒）

1830

1827: The first railroad in North America — the Baltimore & Ohio — is chartered by Baltimore merchants. 1830: The first regularly-scheduled steam-powered rail passenger service in the U.S. begins operation in South Carolina, utilizing the U.S.-built locomotive The Best Friend of Charleston.



In which country was the first railroad built in 1830?

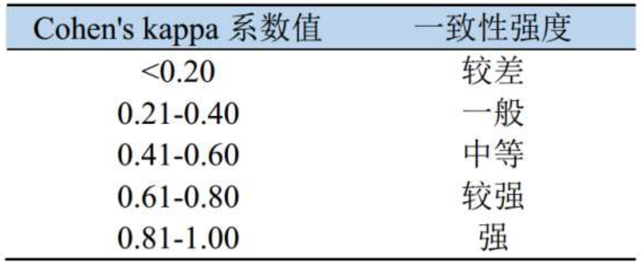
找到约 9,220,000 条结果 （用时 0.91 秒）

Great Britain

The railroad was first developed in Great Britain. A man named George Stephenson successfully applied the steam technology of the day and created the world's first successful locomotive. The first engines used in the United States were purchased from the Stephenson Works in England.

Threats to Validity

- **Representativeness of the test object(QA software) and the datasets:**
    - UnifiedQA is a state of-the-art QA algorithm (only method to unify the solutions)
    - QA software: Open-world QA and Closed-world QA (Google, UnifiedQA)
    - Datasets and Benchmarks: classic and have been widely used
- **Tools that we use to realize the proposed MR (results and implementation details)**
    - Design various methods to avoid the false positive violations (Wh-question generation and semantic similarity measurement are not perfect: Limited NLP techniques)
    - Inspect the factuality of the revealed violations: 80% is valid
- **Manual inspection and categorization of the revealed violations (subjective bias)**
    - Alleviate the bias from subjective cognition (deliver a tutorial and perform Cohen’s Kappa statistics)
    - The agreement rate between two inspectors is substantial (0.79)



Cohen's kappa 系数值	一致性强度
<0.20	较差
0.21-0.40	一般
0.41-0.60	中等
0.61-0.80	较强
0.81-1.00	强

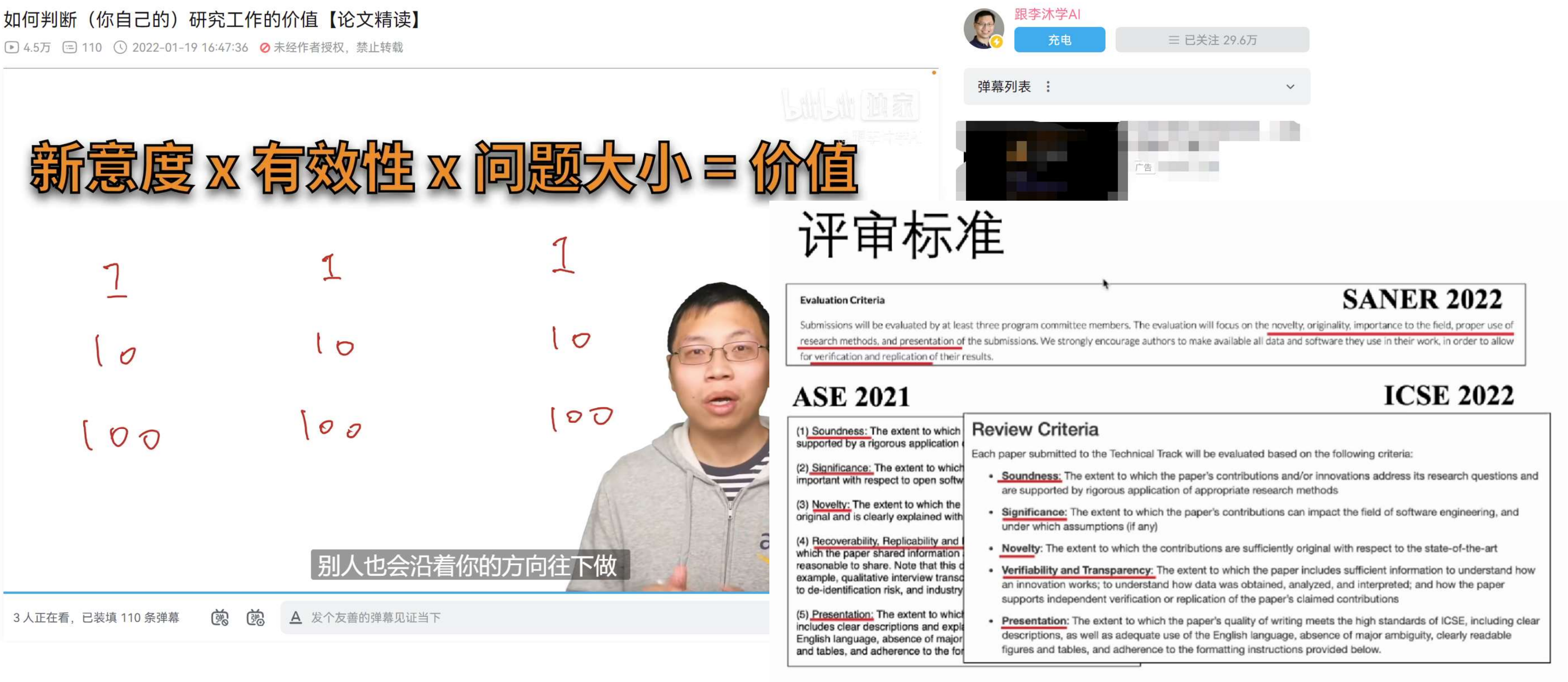
Conclusion

Advantage

Metamorphic testing

- *Break the reliance on the pre-annotated labels of test cases*
- *Enable the flexible just-in-time test and the extensible test that can leverage the massive unlabeled data in real-life usages*
- *A general method which could test all kinds of QA software*

Why is this paper distinguished



如何判断（你自己的）研究工作的价值【论文精读】

4.5万 110 2022-01-19 16:47:36 未经作者授权，禁止转载

跟李沐学AI

新意度 x 有效性 x 问题大小 = 价值

1	1	1
10	10	10
100	100	100

别人也会沿着你的方向往下做

3 人正在看，已装填 110 条弹幕

发个友善的弹幕见证当下

评审标准

Evaluation Criteria SANER 2022

Submissions will be evaluated by at least three program committee members. The evaluation will focus on the novelty, originality, importance to the field, proper use of research methods, and presentation of the submissions. We strongly encourage authors to make available all data and software they use in their work, in order to allow for verification and replication of their results.

ASE 2021

(1) Soundness: The extent to which
supported by a rigorous application

(2) Significance: The extent to which
important with respect to open softw

(3) Novelty: The extent to which the
original and is clearly explained with

(4) Recoverability, Replicability and
which the paper shared information
reasonable to share. Note that this d
example, qualitative interview transc
to de-identification risk, and industry

(5) Presentation: The extent to which
includes clear descriptions and expl
English language, absence of major
and tables, and adherence to the for

ICSE 2022

Review Criteria

Each paper submitted to the Technical Track will be evaluated based on the following criteria:

- **Soundness:** The extent to which the paper's contributions and/or innovations address its research questions and are supported by rigorous application of appropriate research methods
- **Significance:** The extent to which the paper's contributions can impact the field of software engineering, and under which assumptions (if any)
- **Novelty:** The extent to which the contributions are sufficiently original with respect to the state-of-the-art
- **Verifiability and Transparency:** The extent to which the paper includes sufficient information to understand how an innovation works; to understand how data was obtained, analyzed, and interpreted; and how the paper supports independent verification or replication of the paper's claimed contributions
- **Presentation:** The extent to which the paper's quality of writing meets the high standards of ICSE, including clear descriptions, as well as adequate use of the English language, absence of major ambiguity, clearly readable figures and tables, and adherence to the formatting instructions provided below.

Why is this paper distinguished

- **Significance:** QA Software Testing (AI software testing)
- **Novelty:** Metamorphic testing + Sentence transformation + Question Generation
- **Soundness:** Sufficient introduction of procedures; Full of **examples** in the paper
- **Reproducibility:** Dive into the experiment results; Categorize and Analyze
- **Presentation:** Concrete (with sufficient figures and tables)

My Thoughts

My Thoughts

My Thoughts

- Metamorphic testing seems to be useful in Machine Learning Testing (difference with data augmentation)

My Thoughts

- Metamorphic testing seems to be useful in Machine Learning Testing (difference with data augmentation)

- A simple, clear and creative idea is important

My Thoughts

- Metamorphic testing seems to be useful in Machine Learning Testing (difference with data augmentation)
- A simple, clear and creative idea is important
- A simple, clear and creative idea is not enough

Thanks

Zhu Jie

2022.03.17