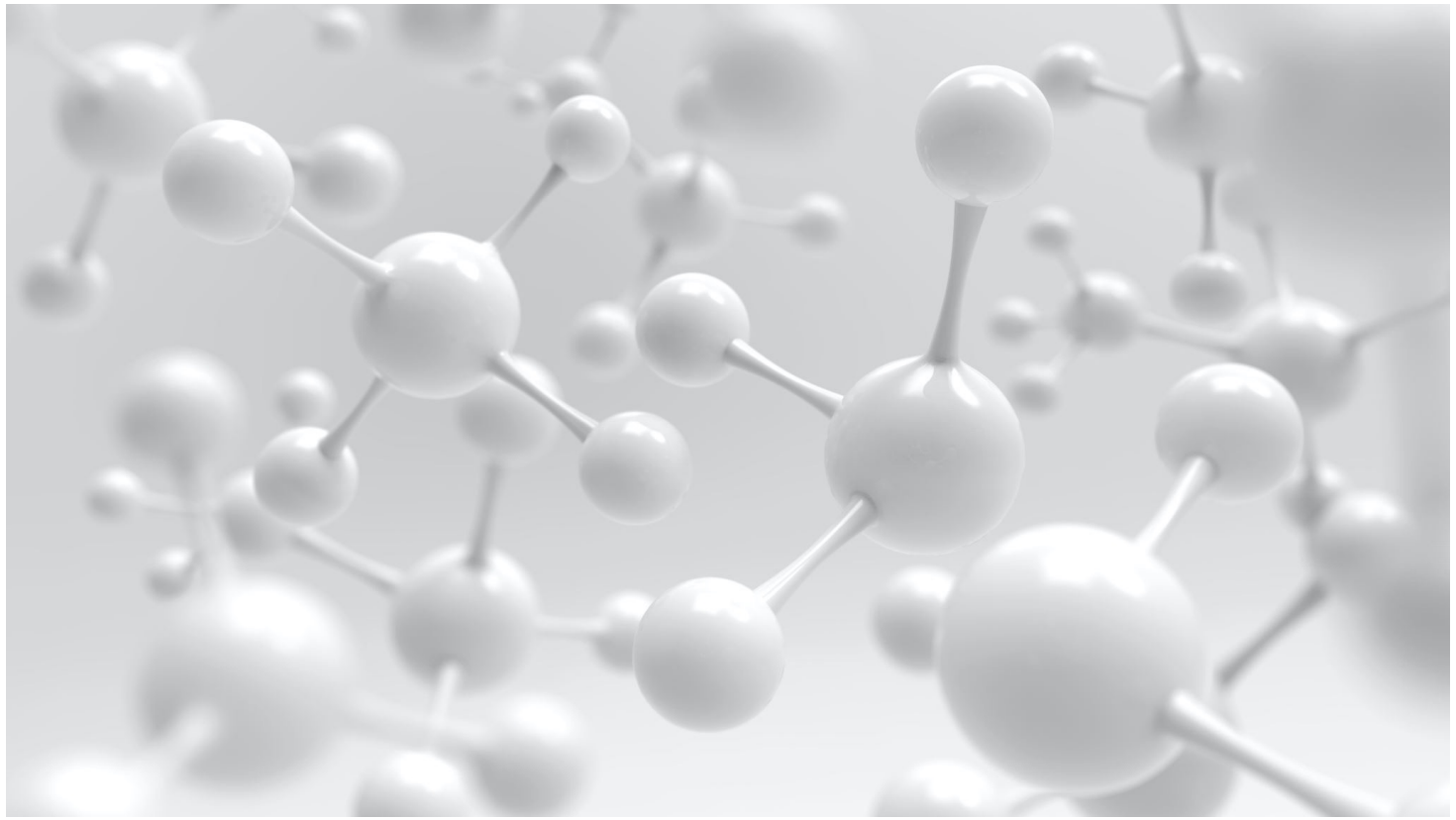


QT32: READING PAPER #2

nhóm

Bài báo

A Vietnamese Dataset for Evaluating Machine Reading Comprehension

https://www.aclweb.org/anthology/2020.coling-main.233.pdf

Kiet Van Nguyen1,2 , Duc-Vu Nguyen1,2 , Anh Gia-Tuan Nguyen1,2 , Ngan Luu-Thuy Nguyen1,2 1University of Information Technology, Ho Chi Minh City, Vietnam 2Vietnam National University, Ho Chi Minh City, Vietnam {kietnv, vund, anhngt, ngannlt}@uit.edu.vn

Abstract

Over 97 million inhabitants speak Vietnamese as the native language in the world. However, there are few research studies on machine reading comprehension (MRC) in Vietnamese, the task of understanding a document or text, and answering questions related to it. Due to the lack of benchmark datasets for Vietnamese, we present the Vietnamese Question Answering Dataset (UIT-ViQuAD), a new dataset for the low-resource language as Vietnamese to evaluate MRC models. This dataset comprises over 23,000 human-generated question-answer pairs based on 5,109 passages of 174 Vietnamese articles from Wikipedia. In particular, we propose a new process of dataset creation for Vietnamese MRC. Our in-depth analyses illustrate that our dataset requires abilities beyond simple reasoning like word matching and demands complicate reasoning such as single-sentence and multiple-sentence inferences. Besides, we conduct experiments on state-of-the-art MRC methods in English and Chinese as the first experimental models on UIT-ViQuAD, which will be compared to further models. We also estimate human performances on the dataset and compare it to the experimental results of several powerful machine models. As a result, the substantial differences between humans and the best model performances on the dataset indicate that improvements can be explored on UIT-ViQuAD through future research. Our dataset is freely available to encourage the research community to overcome challenges in Vietnamese MRC.

Anthology ID: 2020.coling-main.233
Volume: Proceedings of the 28th International Conference on Computational Linguistics
Month: December
Year: 2020
Address: Barcelona, Spain (Online)
Venue: COLING
SIG: –
Publisher: International Committee on Computational Linguistics
Note: –
Pages: 2595–2605
Language: –
URL: https://www.aclweb.org/anthology/2020.coling-main.233

Mục tiê

- Bài toán là gì?
- Problem là gì?
- Ý tưởng của kỹ thuật giải quyết là gì?

Problem

A Vietnamese Dataset for Evaluating Machine Reading Comprehension

Kiet Nguyen, Vu Nguyen, Anh Nguyen, Ngan Nguyen

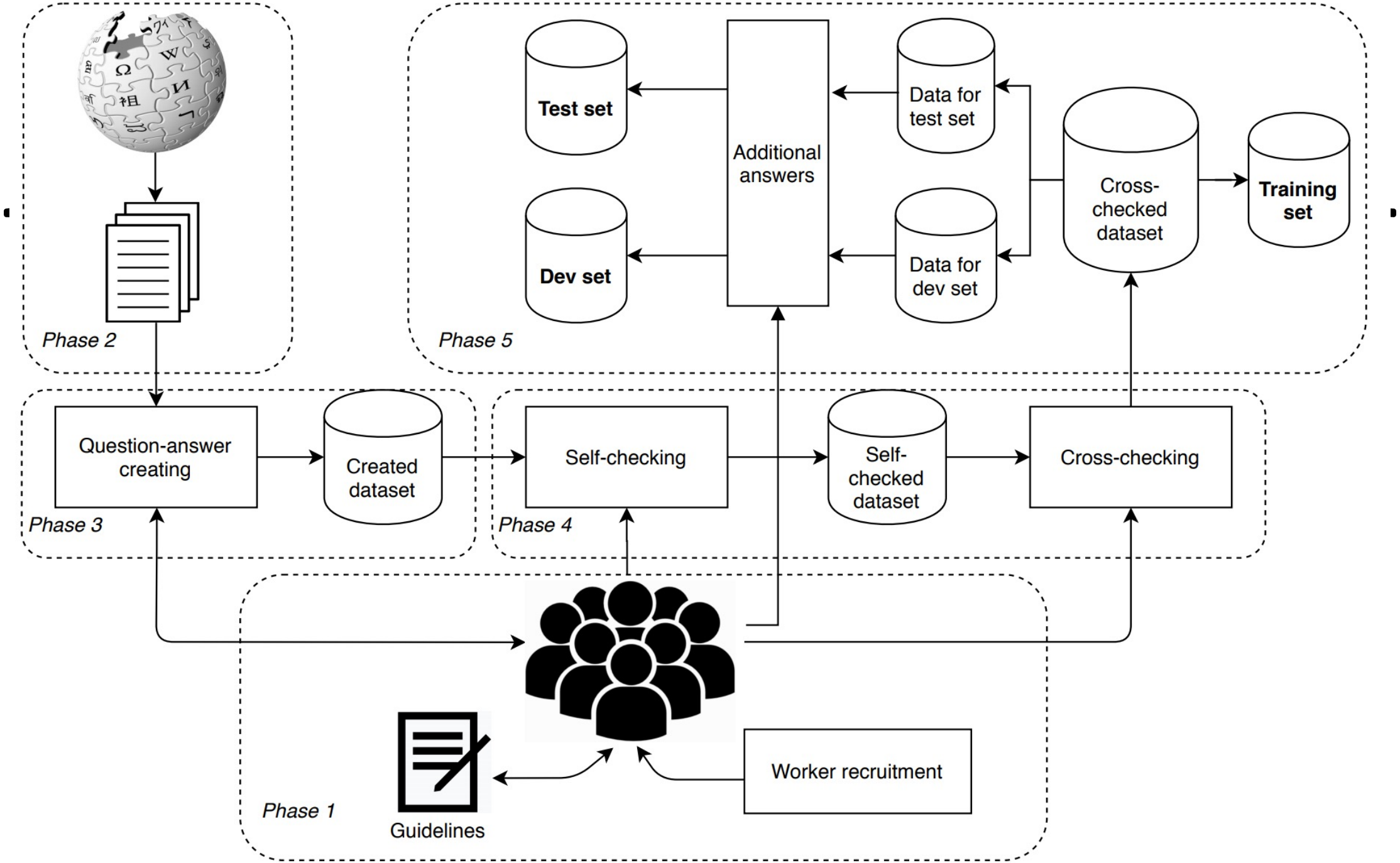
Bài toán

Vấn đề gì?

Abstract

Over 97 million inhabitants speak Vietnamese as the native language in the world. However, there are few research studies on machine reading comprehension (MRC) in Vietnamese, the task of understanding a document or text, and answering questions related to it. Due to the lack of benchmark datasets for Vietnamese, we present the Vietnamese Question Answering Dataset (UIT-ViQuAD), a new dataset for the low-resource language as Vietnamese to evaluate MRC models. This dataset comprises over 23,000 human-generated question-answer pairs based on 5,109 passages of 174 Vietnamese articles from Wikipedia. In particular, we propose a new process of dataset creation for Vietnamese MRC. Our in-depth analyses illustrate that our dataset requires abilities beyond simple reasoning like word matching and demands complicate reasoning such as single-sentence and multiple-sentence inferences. Besides, we conduct experiments on state-of-the-art MRC methods in English and Chinese as the first experimental models on UIT-ViQuAD, which will be compared to further models. We also estimate human performances on the dataset and compare it to the experimental results of several powerful machine models. As a result, the substantial differences between humans and the best model performances on the dataset indicate that improvements can be explored on UIT-ViQuAD through future research. Our dataset is freely available to encourage the research community to overcome challenges in Vietnamese MRC.

Ý tưởng của kỹ thuật giải quyết


Phase 2
Phase 5
Test set
Dev set
Additional answers
Data for test set
Data for dev set
Cross-checked dataset
Training set
Question-answer creating
Created dataset
Phase 3
Self-checking
Self-checked dataset
Cross-checking
Phase 4
Worker recruitment
Guidelines
Phase 1

THANKS