Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6521–6529
Marseille, 20-25 June 2022

VIMQA: A Vietnamese Dataset for Advanced Reasoning and Explainable Multi-hop Question Answering

Nguyen-Khang Le[1,⋆], Dieu-Hien Nguyen[1,⋆], Tung Le[1,2,3], Minh Le Nguyen[1]
[1]Japan Advanced Institute of Science and Technology, Ishikawa, Japan
[2]Faculty of Information Technology, University of Science, Ho Chi Minh city, Vietnam
[3]Vietnam National University, Ho Chi Minh city, Vietnam
{lnkhang, ndhien}@jaist.ac.jp; lttung@fit.hcmus.edu.vn; nguyenml@jaist.ac.jp

Abstract

Vietnamese is the native language of over 98 million people in the world. However, existing Vietnamese Question Answering (QA) datasets do not explore the model's ability to perform advanced reasoning and provide evidence to explain the answer. We introduce VIMQA, a new Vietnamese dataset with over 10,000 Wikipedia-based multi-hop question-answer pairs. The dataset is human-generated and has four main features: (1) The questions require advanced reasoning over multiple paragraphs. (2) Sentence-level supporting facts are provided, enabling the QA model to reason and explain the answer. (3) The dataset offers various types of reasoning to test the model's ability to reason and extract relevant proof. (4) The dataset is in Vietnamese, a low-resource language. We also conduct experiments on our dataset using state-of-the-art Multilingual single-hop and multi-hop QA methods. The results suggest that our dataset is challenging for existing methods, and there is room for improvement in Vietnamese QA systems. In addition, we propose a general process for data creation and publish a framework for creating multilingual multi-hop QA datasets. The dataset and framework are publicly available to encourage further research in Vietnamese QA systems.

Keywords: Multi-hop Question Answering, Dataset, Machine Reading Comprehension

1. Introduction

Question Answering (QA) task is one of the essential tasks in Natural Language Processing (NLP). In particular, most popular QA systems digest a question and contexts to reveal a correct answer. QA models play a vital role in a wide range of NLP applications such as search engines, intelligent agents, chatbots. Traditionally, most popular datasets are designed to evaluate the ability of the systems to understand and answer an input question within a single document. In this type of QA task, the answer is only related to a specific context such as a keyword and a sentence, which is referred to as single-hop reasoning. To this end, there are some datasets in the English language. Particularly, Rajpurkar et al. (2016) proposed SQuAD, one of the large-scale span-extraction QA datasets where questions can be answered by only a single paragraph. Other single-hop datasets, TriviaQA (Joshi et al., 2017), and SearchQA (Dunn et al., 2017), have a more challenging setting where contexts are constructed from multiple documents retrieved given existing question-answer pairs. In these above datasets, complex reasoning over the coordinating context is not adequately addressed, which essentially acquires new kinds of QA tasks to accentuate the multiple associations in text understanding.

Accordingly, Multi-hop QA is highly essential and challenging as a potential solution in this trend. The difference in this task comes from a difficult context which requires a QA systems' ability to gather information from multiple documents and arrive at the answer. Recently, significant progress in this direction has been made in resource-rich languages such as English and Chinese. Correspondingly, HotpotQA (Yang et al., 2018) is one of the largest multi-hop QA datasets, which requires the system to reason over multiple paragraphs and provide supporting facts at the sentence level to support the answer. Other multi-hop QA datasets constructed using existing knowledge bases are also proposed, such as QAngaroo (Welbl et al., 2018), ComplexWebQuestions (Talmor and Berant, 2018). Unfortunately, all of these above datasets are in the English language. As a result, there are few datasets in low-resource languages, especially in Vietnamese. It is too challenging to deploy a Multi-hop QA system in Vietnamese with the current situation.

In the Vietnamese language, the number of span-extraction datasets is limited. One of the largest single-hop span-extraction datasets is the UIT-ViQuAD (Nguyen et al., 2020), which consists of about 23,000 question-answer pairs and has a lot in common with the SQuAD dataset. Unfortunately, similar to SQuAD, UIT-ViQuAD still deals with the lack of reasoning from multiple contexts. Accordingly, a multi-hop QA dataset in the Vietnamese language is highly essential to develop and evaluate Vietnamese QA systems' ability to perform complex reasoning and provide explainable answers. To address this challenge, we introduce the Vietnamese Multi-hop Question Answering Dataset (VIMQA), which requires multi-hop reasoning and provides supporting facts to guide the QA system to perform explainable inference. VIMQA is col-

⋆ The first two authors contributed equally to this work

lected by crowdsourcing based on Wikipedia articles. To ensure the multi-hop questions in VIMQA are natural and not constraint to any pre-existing knowledge base, we show the crowd workers multiple supporting paragraphs and ask them to think up the question requiring reasoning over all of the paragraphs (Yang et al., 2018). Crowd workers are also asked to provide the answers and evidence in the paragraphs that support the answers. VIMQA is publicly available at `https://github.com/vimqa/vimqa`. Figure 1 presents an example in our dataset.
Our main contribution in this paper is as follows.

1. We propose VIMQA, a Vietnamese Dataset for advanced reasoning and explainable Multi-hop QA.

2. We also deploy a framework for collecting multilingual multi-hop QA datasets. Especially, we calibrate the framework for the Vietnamese language.

3. We further provide insight into the dataset through analysis of different linguistic aspects.

4. In addition, our dataset is also evaluated by the current baseline and the state-of-the-art methods in QA to highlight its quality and robustness.

The remaining of this paper is organized as follows. Details of our data collecting schema and approach are shown in Section 2. We provide insights into our dataset through data analysis in Section 3. Data splits and benchmark settings are discussed in Section 4. The detailed experiments of the baseline and state-of-the-art models in our dataset are presented in Section 5. Finally. Section 6 concludes our work and discuss directions for future works.

Paragraph 1, John O'Shea:
[1] John Francis O'Shea (sinh ngày 30 tháng 4 năm 1981) là một cựu cầu thủ bóng đá người Ireland và hiện là huấn luyện viên đội một cho Reading. [2] Sinh ở Waterford, O'Shea gia nhập Manchester United năm anh 17 tuổi và được đánh giá như một trong những cầu thủ đa năng nhất ở Premier League. [3] Anh đã từng chơi ở mọi vị trí cho Manchester United, bao gồm cả thủ môn trong một trận đấu gặp Tottenham Hotspur.
Translation: *[1] John Francis O'Shea (born 30 April 1981) is an Irish former footballer and current first team coach for Reading. [2] Born in Waterford, O'Shea joined Manchester United at the age of 17 and is widely regarded as one of the most versatile players in the Premier League. . [3] He played in every position for Manchester United, including as a goalkeeper in a match against Tottenham Hotspur.*
Paragraph 2, Manchester United F.C.:
[4] Câu lạc bộ bóng đá Manchester United (tiếng Anh: Manchester United Football Club, hay ngắn gọn là MU hay Man Utd) là một câu lạc bộ bóng đá chuyên nghiệp có trụ sở tại Old Trafford, Đại Manchester, Anh. [5] Câu lạc bộ đang chơi tại Giải bóng đá Ngoại hạng Anh, giải đấu hàng đầu trong hệ thống bóng đá Anh.
Translation: *[4] Manchester United Football Club (English: Manchester United Football Club, or simply MU or Man Utd) is a professional football club based at Old Trafford, Greater Manchester, England. [5] The club plays in the English Premier League, the top division in English football.*
Question: Câu lạc bộ John O'Shea gia nhập năm 17 tuổi có trụ sở ở đâu? *(Where is the club in which John O'Shea joined when he was 17 years old based?)*
Answer: Old Trafford
Supporting facts: 2, 5

Figure 1: Example of the multi-hop questions in VIMQA. Supporting facts are also a part of the dataset and are highlighted in blue. The translation is in *italic*

2. Data Collection

In this section, we describe the details of our data collecting pipeline. Inspired by the work of Yang et al. (2018), we aim to design a framework to collect multilingual explainable QA datasets that require multi-hop reasoning. Although we primarily deploy our framework in the Vietnamese language, our designed framework is general enough to be tweaked to adapt it into every language. Despite the existence of a few Multi-hop QA datasets, our framework is a beginning of convenience and simplicity in the development of Multi-hop QA.
Traditionally, one way to collect multi-hop datasets is through reasoning chains based on a knowledge base. However, the resulting dataset might be constrained to the knowledge base and not diverse (Yang et al., 2018). Following the approaches of Rajpurkar et al. (2016) and Yang et al. (2018) to collect a text-based QA dataset, we design our framework with a fewest modification. A typical QA sample consists of some contexts and a question. In this setting, an answer should be extracted by a span of text from the context. At the same time, a question requires multi-hop reasoning over multiple contexts.
Our target dataset must require advanced reasoning over multiple paragraphs and provide supporting facts for explainable predictions. Moreover, the data collecting pipeline should be flexible and easy to adapt to any language with as few changes as possible. To address our motivation, we propose an overall pipeline of data collection in Figure 2. Firstly, a title is randomly sampled from the curated list of feasible titles. Then, using this title, we randomly choose a paragraph pair from the Wikipedia graph. The pair is then shown to the crowd workers to collect questions, answers, and supporting facts. Finally, the annotated sample is processed and normalized by our configuration. The details of each component and processing step are presented in the following parts.

2.1. Wikipedia Graph

Despite the ease of integration in our framework, we deploy it on the Vietnamese Wikipedia. Coming from

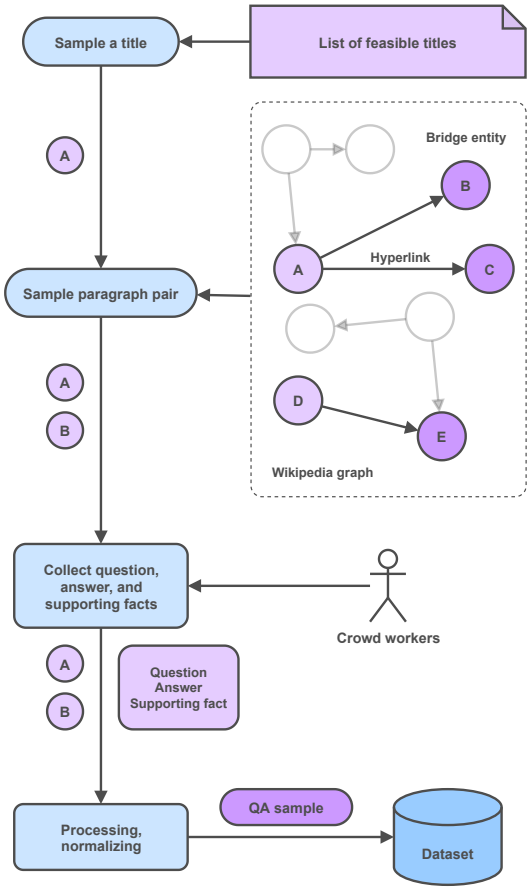

Figure 2: Overall data collecting pipeline of VIMQA

the architecture of Wikipedia, our VIMQA shares the same characteristics as Yang et al. (2018) observed in the English Wikipedia as follows: (1) Hyper-links in the Wikipedia articles are often useful for multi-hop reasoning because they entail an association between two entities in the graph. (2) The summary (the first passage of each Wikipedia article) often contains the most information that facilitates meaningful questions. Therefore, we consider the Vietnamese Wikipedia corpus a directed graph where each vertex is an entity (represented by a unique article title). Each edge (u, v) denotes a hyperlink from article u to article v. We only consider the summary passage of each article.

2.2. Feasible Titles List

The Vietnamese Wikipedia consists of about 1,200,000 articles. Its size is about one-fifth of the English Wikipedia. However, not all articles are eligible for creating multi-hop questions. For example, in our pilot studies, we found that articles about general concepts such as "football", "city", "music" are difficult to come up with multi-hop questions. On the other hand, articles about a particular person, event, or place are easier to create multi-hop questions. In addition, it is also challenging for the crowd worker to create meaningful questions when the articles are highly technical such as "Binary search tree", "TCP/IP". To address this problem, we manually select a list of feasible article titles that are straightforward to collect meaningful multi-hop questions. Although we also provide the tool to gather all titles in a specific Wikipedia, users should filter the list of feasible titles via their goals.

2.3. Paragraph Pairs Selection

The crowd workers are given a pair of paragraphs to think up a question requiring multi-hop reasoning. Our method to select the paragraph pairs is followed to HotpotQA (Yang et al., 2018). We consider the question in the example of Section 1 "Where is the club in which John O'Shea joined when he was 17 years old based?". Naturally, to answer this question, we first need to perform reasoning from the first paragraph to know that "the club in which John O'Shea joined when he was 17 years old" is "Manchester United". Then, we can then find where the club is based in the second paragraph. As proposed by Yang et al. (2018), the "Manchester United" in our example can be considered the *bridge entity*, and it often connects the contexts of the two paragraphs. Therefore, to sample the paragraph pairs, we first get a title A from the feasible titles list and then sample an edge (A, B) in the Wikipedia graph where B is also in the feasible titles list. The pair of paragraphs A and B is then presented to the crowd workers to create QA data.

Besides creating questions using the bridge entity, comparing two different entities in the same category also leads to interesting questions (Yang et al., 2018) such as "Does Cristiano Ronaldo has more titles than Ryan Giggs?". For comparison questions, we manually collect lists of similar entities. Each list contains entities in a same category such as "Fooballers", "Musicians", "Scientists", "Organizations", "Countries", etc. To sample a paragraph pair for comparison questions, we randomly select two paragraphs in the same list and provide them to the crowd worker to create QA data.

2.4. Annotation by Crowd Workers

To create a QA sample, the crowd worker is given a pair of paragraphs and must provide a multi-hop question, an answer, and supporting facts. We develop a working interface for the crowd workers to do this task. Figure 3 shows the screenshot of the crowd working interface during data collection. The interface provides step-by-step instruction and only allows the crowd worker to submit the result when all the requirements are satisfied. This helps to prevent human error when collecting the data. The crowd worker is also hinted that the multi-hop question can be made by asking questions about the bridge entity.

We have three crowd workers who are researchers with Vietnamese native language annotate the VIMQA dataset. At the end of each day, the crowd workers verify each other examples. Only examples that are verified by more than one worker are added to the dataset.

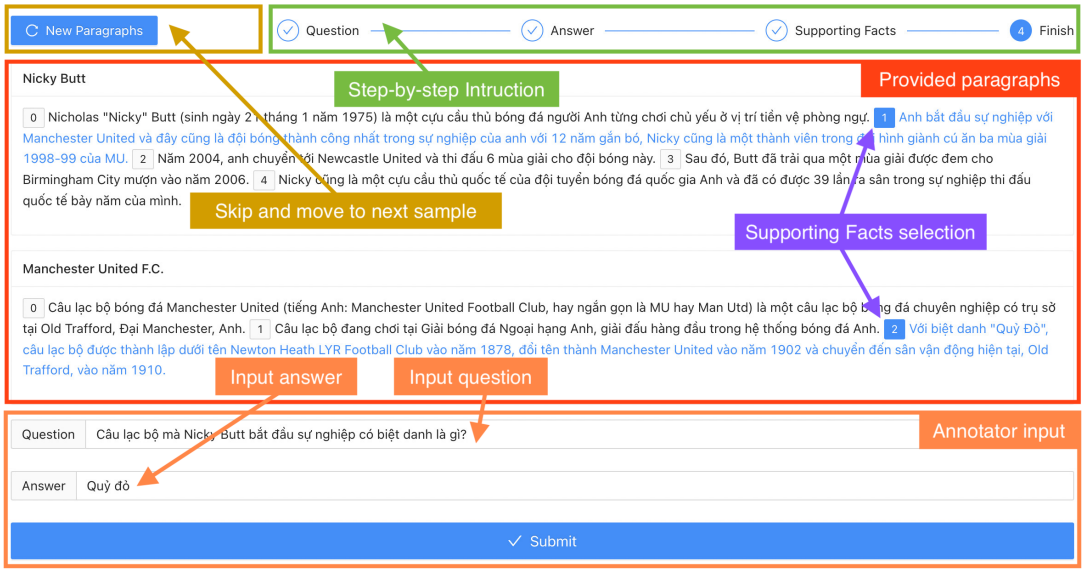

Figure 3: Screenshot of our interface for annotators

2.5. Processing and Normalizing

Each language has unique characteristics and needs to process and normalize differently. One of the normalizing problems in the Vietnamese language is a Unicode encoding of accents. For example, accented letters in Vietnamese such as "á" can be encoded using either a single Unicode point (U+00E1) or two Unicode points (combining acute accent - U+0301 and lower case letter A - U+0061). The reason for this phenomenon comes from the typical characteristics of the Vietnamese. Besides a simple character, some complex characters contain tonal symbols in Vietnamese. Since we collect text data from the crowd workers, the way of encoding depends on the crowd worker's encoding software. Different encodings look the same to humans but are interpreted differently by computer models. Our dataset normalizes all accented Vietnamese letters to single Unicode points.

Another problem in the Vietnamese language is the accent positions in words. For example, "hoà" and "hòa" is the same word in Vietnamese, but the accent is put at different characters and can be interpreted differently by computer models. Therefore, we normalize these words based on the official dictionaries.

Unfortunately, the post-processing and normalization depend on the typical characteristics of different languages. Although we also try to provide all convenient tools in Vietnamese, users should modify and design the other ones for their own language. However, our framework is flexible enough to adapt it into every language with a few modifications.

3. Data Analysis

3.1. Question Analysis

Our analysis focuses on revealing the typical distribution of length and types in questions. Firstly, we also point out the various question types in our VIMQA dataset. In particular, we define a list of central question words (CQW) in Vietnamese. It is used to divide questions into their specific categories in Table 1. If questions are not in the CQW list, they are manually classified into eight large categories.

Group	English CQW	Vietnamese CQW
Yes/No	Copulas (is, are) Aux (does, did)	Phải không, Đúng không
Which	Which	Nào
What	What	Là gì
	What ordinal number	Thứ mấy, Thứ bao nhiêu
Who	Who	Ai
	By whom	Bởi ai
How	How many	Bao nhiêu
	How often	Bao lâu một lần
	How long	Bao lâu
	How far	Bao xa
When	When	Khi nào
Where	Where	Ở đâu, Tại đâu
Why	Why	Vì sao, Tại sao

Table 1: List of Vietnamese central question words

Based on the above division, the distribution of question types is presented in Figure 4. Yes/No questions account for about a third of the total questions. Besides Yes/No questions, "Which", "What", and "Who" questions have the largest proportion in the dataset. This characteristic is also similar to the observation in HotpotQA (Yang et al., 2018).

We also analyze the distribution of question lengths in the dataset. Figure 5 shows the distribution of lengths of questions in VIMQA. The distribution indicates that questions vary remarkably in size.

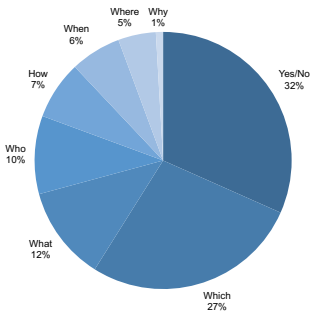

Figure 4: Distribution of question types in VIMQA

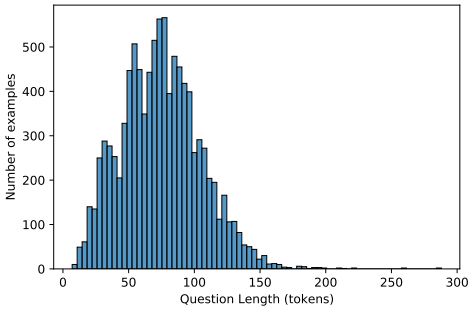

Figure 5: Distribution of question lengths in VIMQA

3.2. Answer Analysis

Following the configuration of HotpotQA in English, we also sample 100 examples from our dataset to analyze the answer types. Table 2 presents the types of answers. The distribution of answer types suggests that VIMQA covers various types of answers, which complements the previous analysis on question types. A majority of answer types are Yes/No (28%), location (15%), date/time (12%), and person (11%). It proves that our dataset is highly challenging and sufficient in the quality of multi-hop QA for Vietnamese.

Answer Type	%	Example(s)
Yes/No	28	Đúng, Không *(Yes, No)*
Location	15	Tây Bắc Châu Âu, Nhật Bản *(Northwestern Europe, Japan)*
Date and time	12	1908, thời kỳ trị vì của Trần Nhân Tông *(1908, the reign of King Tran Nhan Tong)*
Person	11	Benjamin Franklin, Nguyễn Phú Trọng
Group / Org	6	The Beatles, Republic Records
Title / Nick name	5	Ông hoàng nhạc pop, Quỷ Đỏ *(King of Pop, Red Devils)*
Ordinal Number	4	hạng nhất, hạng tư *(first prize, fourth prize)*
Number	8	130 triệu; 45,5 tỷ bảng Anh *(130 million, 45.5 billion pounds)*
Proper noun	6	I'm Too Sexy, dân tộc Nùng *(I'm Too Sexy, Nung ethnic group)*
Common noun	3	hoá học, rắn hổ mang chúa *(chemistry, King cobra)*
Other	2	bằng thiết bị kết nối Internet *(with an Internet-connected device)*

Table 2: Types of answers in VIMQA

3.3. Multi-hop Reasoning Type Analysis

To provide a better insight into the types of multi-hop reasoning in VIMQA, we randomly sampled 100 examples from the development and test sets and manually categorized the kinds of reasoning needed to answer each question. Table 3 describes the details of different types of multi-hop reasoning in VIMQA with self-explanatory examples.

A majority of questions have Type I reasoning. In this type of reasoning, the reader must first identify the bridge entity in the question and locate it in the context (Frank Capra in the example), then perform the second-hop reasoning to answer the question (where Frank Capra was born?). This type of reasoning is also referred to as chain reasoning.

In Type II reasoning, the answer entity (The Byrds in the example) usually lies among a list of entities (The Byrds, Crosby, Stills & Nash). The reader needs to check multiple properties of the entity in the question to choose the correct entity in the list.

Type III needs more than two supporting facts to answer and often requires more complex inference like three-hop reasoning. Type IV requires the reader to comprehend the properties of the two entities in the question.

In addition, inspired by the work of Rajpurkar et al. (2018) about constructing unanswerable questions, we create a new Yes/No type of question (Type V, IV) that needs identifying the Negation/Entity Swap to arrive at the answer yes or no. Not only does this type of question needs multi-hop reasoning, but it also requires the ability to recognize negation and false entity in the contexts.

4. Benchmark Settings

4.1. Data Splits

In total, we collected and annotated 10,047 valid examples in VIMQA. For evaluation and experiments, we employ the configuration of HotpotQA (Yang et al., 2018), a multi-hop QA dataset in English, to divide our dataset into training, developing, and testing set. Firstly, it is necessary to perform cross-validation to choose the remarkable samples. In our process, the cross-validation is done by the HotpotQA model (baseline) 5 times. The detailed results are presented in Table 4. It is easily observed that the model correctly answers about 40% of the questions. We split out these 40% (correctly answered) questions and mark them as train-normal. This portion is used as part of the training set.

The other 60% of the questions are complex questions which the model fails to answer correctly. As we mentioned above, we aim to build a dataset to evaluate the model's ability to perform advanced and complex reasoning. Therefore, we make the validation split and the test split containing complex examples only. Using a similar approach as HotpotQA (Yang et al., 2018), we divide the complex examples (60% of the dataset) into

Reasoning Type	%	Example(s)
I. Inferring the **bridge entity** to complete the 2nd-hop question	54	**Question:** Đạo diễn phim It Happened One Night sinh ra ở đâu? *(Where was the director of It Happened One Night born?)* **Paragraph 1:** It Happened One Night là một bộ phim hài Mỹ ..., đạo diễn **Frank Capra**. *(It Happened One Night is a comedy film ..., directed by **Frank Capra**)* **Paragraph 2:** **Frank Capra** ... Sinh ra ở **Ý** và lớn lên ở Los Angeles ... *(**Frank Capra** ... Born in **Italy** and raised in Los Angeles ...)*
II. Locating the **answer entity** by checking multiple properties	28	**Question:** David Crosby từng là thành viên sáng lập của ban nhạc nào tan rã vào năm 1973? *(Which band did David Crosby founded broke up in 1973?)* **Paragraph 1:** David Van Cortlandt Crosby ... còn là thành viên sáng lập của các ban nhạc **The Byrds**, Crosby, Stills & Nash ... *(David Van Cortlandt Crosby ... was also a founding member of **The Byrds**, Crosby, Stills & Nash ...)* **Paragraph 2:** **The Byrds** là ban nhạc rock ... cho tới khi tuyên bố tan rã vào năm 1973. *(**The Byrds** were a rock band ... until their disbandment in 1973.)*
III. Other types of reasoning that require more than two supporting facts	4	**Question:** Giải đấu nào Fabien Barthez từng có một số danh hiệu được điều hành bởi Ligue de Football Professionnel? *(Which league did Fabien Barthez have several titles is run by the Ligue de Football Professionnel?)* **Paragraph 1:** Fabien Alain Barthez ... đã từng chiến thắng tại giải Cúp các đội vô địch bóng đá quốc gia châu Âu, một số danh hiệu tại **Giải vô địch bóng đá Pháp** và Giải bóng đá Ngoại hạng Anh. *(Fabien Alain Barthez ... has won the UEFA Champions League, several titles at **The French national football championship** and The English Premier League.)* **Paragraph 2:** **Giải bóng đá vô địch quốc gia Pháp** (tiếng Pháp: **Ligue 1**), ... Được điều hành bởi Ligue de Football Professionnel, Ligue 1 bao gồm ... *(**The French national football championship** (French: **Ligue 1**), ... Administrated by the Ligue de Football Professionnel, Ligue 1 consists of ...)*
IV. Comparing **two entities**	7	**Question:** **Daniel Sturridge** và **Frank Lampard** đều có chơi cho câu lạc bộ Chelsea phải không? *(Do **Daniel Sturridge** and **Frank Lampard** both play for Chelsea Football Club?)* **Answer:** **đúng** *(**yes**)* **Paragraph 1:** **Daniel Andre Sturridge** ... Anh rời Manchester City ... và gia nhập Chelsea theo dạng chuyển nhượng tự do. *(**Daniel Andre Sturridge** ... He left Manchester City ... and joined Chelsea as a free agent.)* **Paragraph 2:** **Frank James Lampard OBE** ... Anh được xem là một trong những cầu thủ xuất sắc nhất lịch sử của Chelsea và ... *(**Frank James Lampard OBE** ... He is considered to be one of Chelsea's greatest ever players and ...)*
V. Identifying the **Negation** factor to answer Yes/No questions	4	**Question:** Francesco Totti **chưa từng** thi đấu cho đội bóng nào ở Ý phải không? *(Have Francesco Totti **never** played for any Italian football club?)* **Answer:** **không** *(**no**)* **Paragraph 1:** Totti giải nghệ ngày 28 tháng 5 năm 2017 sau khi cùng **Roma** giành chiến thắng 3-2 trước Genoa ... *(Totti retired on May 28th 2017 after playing for **Roma** in a 3-2 win over Genoa ...)* **Paragraph 2:** **A.S. Roma** ... là một đội bóng thủ đô của Ý, ... *(**A. S. Roma** ... is an Italian capital professional football club, ...)*
VI. Identifying the **Entity** Swap to answer Yes/No questions	3	**Question:** Đội bóng của Nathan Dyer thành lập năm **1812** phải không? *(Was Nathan Dyer's football team founded in **1812**?)* **Answer:** **không** *(**no**)* **Paragraph 1:** Nathan Antone Jonah Dyer ... hiện đang chơi cho đội **Swansea City** ở vị trí tiền vệ cánh. *(Nathan Antone Jonah Dyer ... currently plays for **Swansea City** as a midfielder.)* **Paragraph 2:** **Swansea City** Association Football Club (thành lập năm **1912**) là một câu lạc bộ bóng đá chuyên nghiệp có trụ sở tại ... *(Swansea City Association Football Club (founded in **1912**) is a professional football club based in ...)*

Table 3: Types of multi-hop reasoning required to answer questions in the VIMQA. The English translations are provided in *italics*. The bridge entity is shown in **orange bold**. The blue indicates supporting facts for the answers. The answers are highlighted in **green bold**. Words that reflects the reasoning type are marked in **purple**

Fold	Answer		Sup Fact		Joint	
	EM	F1	EM	F1	EM	F1
1	31.3	36.1	13.4	43.8	5.5	17.7
2	31.7	36.5	25.4	59.7	9.9	23.7
3	31.0	37.0	21.8	55.6	8.36	22.65
4	36.6	42.0	13.6	42.3	5.4	18.9
5	32.5	37.7	28.5	62.6	10.7	25.2

Table 4: Result of 5-fold cross-validation on VIMQA

three subsets: train-hard, validation, test. The detail of our dataset is shown in Table 5.

Name	Desc.	Usage	# Examples
train-normal	normal questions	train	4,018
train-hard	hard questions	train	4,023
dev	hard questions	validation	1003
test	hard questions	test	1003
Total			10,047

Table 5: The data splits of VIMQA

4.2. Benchmark settings

Inspired by the work of Yang et al. (2018), we create two different benchmark settings for evaluation: the Gold Only and the Distractor setting. Both of them share the same samples in the test set with a little difference in the input components.

The Gold Only setting tests the model's capability to perform multi-hop reasoning to output the answer and sentence-level supporting facts to explain its answer. In this setting, the models are provided with two gold paragraphs (two paragraphs used to create the question-answer pair) and a question. The question requires advanced multi-hop reasoning to arrive at the answer.

The Distractor setting tests the model's ability to find the answer and supporting facts when there are noises from the distracting paragraphs. In this benchmark, the models are presented with ten paragraphs (two gold paragraphs and eight distractors) and must locate the answer and supporting facts in the correct paragraphs. To create this benchmark, for each QA example, we use the question as a query and employ TF-IDF (Chen et al., 2017) to select eight summary paragraphs from Wikipedia. Combining with the two gold paragraphs from which the question and answer are collected, we have a total of ten paragraphs for each example in the distractor sets. These ten paragraphs are shuffled before being used.

5. Experiments

5.1. Experimental Settings

Most previous approaches of multi-hop QA are deployed in English. Therefore, we re-implement state-of-the-art multilingual QA models and conduct experiments on the VIMQA dataset in Vietnamese. These approaches have been proved on several QA benchmarks in English (e.g SQuAD (Rajpurkar et al., 2016)) and Vietnamese (e.g. UIT-ViQUAD (Nguyen et al., 2020)). The details of our competitive baselines are presented as follows:

- BERT (Devlin et al., 2019): is the most popular approach in many NLP tasks. Our evaluation utilizes the multilingual BERT (mBERT), which is pre-trained in 104 languages, including Vietnamese. Only $\text{mBERT}_{\text{Base}}$ is available for multilingual configuration.
- XLM-RoBERTa (Conneau et al., 2020): gains significant performance for a wide range of cross-lingual transfer tasks. In our experiments, we evaluate two versions of this model, $\text{XLM-RoBERTa}_{\text{Base}}$ and $\text{XLM-RoBERTa}_{\text{Large}}$.
- InfoXLM (Chi et al., 2021): is an Information-Theoretic framework for cross-lingual language model sharing the same architecture as XLM-RoBERTa with an improvement in the cross-lingual transfer-ability. Our experiments evaluate two versions of this model, $\text{InfoXLM}_{\text{Base}}$ and $\text{InfoXLM}_{\text{Large}}$.

For Yes/No questions, we add two special tokens representing *Yes* and *No* at the beginning of the contexts to make the Yes/No answer span appear in the contexts. This method allows the model to extract answers for Yes/No questions.

In addition, to show that the questions in VIMQA require a higher level of reasoning than the existing QA dataset in Vietnamese, we also conduct experiments to compare VIMQA with UIT-ViQUAD (Nguyen et al., 2020), one of the largest Wikipedia-based Vietnamese span-extraction QA datasets. For XLM-RoBERTa and mBERT, we use the results reported in (Nguyen et al., 2020) for comparison. For InfoXLM, we run our implementation on the UIT-ViQUAD dataset and use the result for comparison.

Following the benchmark settings in Section 4, we evaluate the models in two settings of VIMQA (Gold Only and Distractor). For the Distractor setting, we first use BM25 to retrieve two out of ten provided paragraphs using the question as the query. The two retrieved paragraphs are then fed to the QA model to extract the answer. For the Gold Only setting, we only utilize the QA model to extract the answer span of each sample.

Finally, to evaluate the whole Multi-hop QA system, we also re-implement the baseline model proposed by Yang et al. (2018) for evaluating our VIMQA dataset in three sets of metrics for multi-hop QA: answer, supporting facts, and joint. As a baseline to assess the supporting facts metrics of the above QA approaches, we consider the sentences containing the answer span as supporting facts.

Following the work of Rajpurkar et al. (2016) and Yang et al. (2018), we employ two evaluation metrics: exact

Settings	Methods	Answer EM		Answer F1	
		Dev	Test	Dev	Test
Gold Only	mBERT	**56.63**	**55.03**	**71.27**	**70.50**
	XLM-RoBERTa$_{Base}$	47.35	43.76	62.70	59.38
	XLM-RoBERTa$_{Large}$	50.14	49.75	66.42	65.64
	InfoXLM$_{Base}$	50.54	49.05	67.68	65.76
	InfoXLM$_{Large}$	50.65	49.75	66.09	65.29
Distractor	BM25 + mBERT	**41.77**	**39.08**	**51.17**	**49.34**
	BM25 + XLM-RoBERTa$_{Base}$	29.31	29.11	40.04	39.47
	BM25 + XLM-RoBERTa $_{Large}$	32.20	32.30	42.33	43.80
	BM25 + InfoXLM$_{Base}$	36.19	34.39	47.59	45.82
	BM25 + InfoXLM$_{Large}$	31.40	31.10	43.24	42.53
	Human	**87.40**		**91.26**	

Table 6: Performance of the evaluated methods on the dev and test set of VIMQA in two benchmark settings.

match (EM) and F1 to evaluate the answer. Furthermore, we also use two sets of metrics proposed by Yang et al. (2018) to assess the explainability of the models. The first set is EM and F1 on the set of supporting fact sentences compared to the gold set. The second set is joint metrics that combine the evaluation of answer spans and supporting facts.

5.2. Human Performance

For human performance, we randomly sampled 500 examples from the VIMQA development and test sets (Distractor setting) and asked three additional native Vietnamese researchers to provide answers and supporting facts for these examples. We compare the gold annotation (collected during data collection) and human predictions (collected during establishing human performance) by evaluation metrics in answers, supporting facts, and joint. This result is considered as the human performance on the VIMQA dataset.

5.3. Results

Table 6 shows the performance of the evaluated models on the development and test sets of VIMQA along with human performance. The result suggests that our dataset is challenging for existing QA models and the Distractor setting is more complex than the Gold Only setting. mBERT has the highest performance, but it is still significantly lower than human performance.

Method	Split	VIMQA		UIT-ViQuAD	
		EM	F1	EM	F1
XLM-RoBERTa $_{Base}$	dev	47.35	62.70	63.87	81.90
	test	43.76	59.38	63.00	81.95
XLM-RoBERTa $_{Large}$	dev	50.14	66.42	69.18	87.14
	test	49.75	65.64	68.98	87.02
mBERT	dev	**56.63**	**71.27**	62.20	80.77
	test	**55.03**	**70.50**	59.28	80.00
InfoXLM $_{Base}$	dev	50.54	67.68	65.94	82.81
	test	49.05	65.76	64.36	82.39
InfoXLM $_{Large}$	dev	50.65	66.09	**72.52**	**88.85**
	test	49.75	65.29	**69.34**	**87.43**

Table 7: Comparing the performance of the models on VIMQA (Gold Only setting) and UIT-ViQUAD

Table 7 compares the performance of the models on our dataset (VIMQA) and UIT-ViQUAD (Nguyen et al., 2020). To compare the results fairly, we evaluate the models in the VIMQA Gold Only setting, where only two gold paragraphs are provided. The result indicates that VIMQA is more challenging than UIT-ViQUAD, one of the largest Vietnamese span-extraction datasets. The result shows that VIMQA is more challenging for existing methods than the UIT-ViQUAD dataset.

Method	Answer		Sup Fact		Joint	
	EM	F1	EM	F1	EM	F1
Baseline	16.95	27.92	**25.12**	**53.42**	4.89	16.88
BM25 + InfoXLM $_{Large}$	31.10	42.53	19.34	31.45	**11.07**	21.94
BM25 + XLM-R$_{Large}$	32.30	43.80	20.64	32.86	10.97	**22.14**
BM25 + mBERT	**39.08**	**49.34**	18.04	31.33	7.87	18.30
Human	87.40	91.26	72.20	79.39	72.20	77.12

Table 8: Comparing existing methods in three sets of metrics on the Distractor test set of VIMQA

Finally, Table 8 compares the performance of selected methods, the baseline method, and human performance in the Distractor test set of VIMQA. The result suggests that the selected models have higher performance than the baseline method but is dramatically lower than human performance in all three sets of metrics.

6. Conclusion

In this work, we propose VIMQA, a multi-hop Vietnamese QA dataset. It is highly necessary and important to facilitate the development of Vietnamese QA models that can perform advanced reasoning and provide explainable answers with supporting facts. Then, we also propose a pipeline for collecting multi-hop QA examples that can be generalized for all languages. We also prove the efficiency of our pipeline via the detailed analysis in our VIMQA dataset. The experimental results indicate that VIMQA is challenging for competitive approaches in both single and multiple hop QA. It reveals that our VIMQA dataset is a good resource for Vietnamese and cross-lingual QA models, especially in Vietnamese Multi-hop QA tasks for reasoning and explaining the comprehension and coherence of text understanding.

7. Bibliographical References

Chen, D., Fisch, A., Weston, J., and Bordes, A. (2017). Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada, July. Association for Computational Linguistics.

Chi, Z., Dong, L., Wei, F., Yang, N., Singhal, S., Wang, W., Song, X., Mao, X.-L., Huang, H., and Zhou, M. (2021). InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online, June. Association for Computational Linguistics.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Dunn, M., Sagun, L., Higgins, M., Guney, U., Cirik, V., and Cho, K. (2017). Searchqa: A new qa dataset augmented with context from a search engine. 04.

Joshi, M., Choi, E., Weld, D., and Zettlemoyer, L. (2017). TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada, July. Association for Computational Linguistics.

Nguyen, K., Nguyen, V., Nguyen, A., and Nguyen, N. (2020). A Vietnamese dataset for evaluating machine reading comprehension. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2595–2605, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November. Association for Computational Linguistics.

Rajpurkar, P., Jia, R., and Liang, P. (2018). Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia, July. Association for Computational Linguistics.

Talmor, A. and Berant, J. (2018). The web as a knowledge-base for answering complex questions. In *NAACL*.

Welbl, J., Stenetorp, P., and Riedel, S. (2018). Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R., and Manning, C. D. (2018). HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium, October-November. Association for Computational Linguistics.