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

Nguyen-Khang Le, Dieu-Hien Nguyen, Tung Le, Minh Le Nguyen

Summary

Vietnamese is the native language of over 98 million people in the world. However, existing Vietnamese Question Answering (QA) datasets do not explore the ability of the models 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) Questions require advanced reasoning over multiple paragraphs

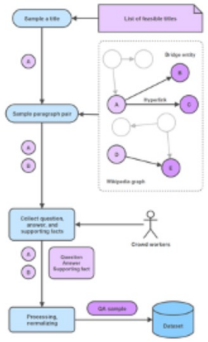
(2) Sentence-level supporting facts

(3) Various types of reasoning

(4) The dataset is in Vietnamese, a low-resource language

Data Collection

- **Wikipedia Graph**: Hyper-links are useful for multi-hop reasoning. The summary contains the most information that facilitates meaningful questions. We consider the Vietnamese Wikipedia corpus a directed graph where each vertex is an article, and each edge (u, v) denotes a hyperlink from article u to article v. Only the summary passage of each article is considered.
- **Feasible Titles List**: General concepts are difficult to create multi-hop questions. Articles about particular people, events, places are easier. we manually select a list of feasible article titles that are straightforward to collect multi-hop questions.



Example of VIMQA

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 he 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

- **Paragraph Pairs Selection**: 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. A and B are then presented to the crowd workers to create QA data. Multi-hop questions can be created by asking question about the *bridge entity* B.
- **Annotation by Crowd Workers**: A working interface is provided for the crowd workers to create multi-hop questions from the given two paragraphs. Three crowd workers who are researchers with Vietnamese native language annotated the VIMQA dataset. Only examples that are verified by more than one worker are added to the dataset.

Processing and Normalizing

- Accent Unicode encoding: 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) depending on the encoding software. Different encodings look the same to humans but are interpreted differently by computer models. All accented Vietnamese letters are normalized to single Unicode points in VIMQA.
- Accent position in words: "hoà" and "hòa" are the same word in Vietnamese, but the accent is put at different characters and can be interpreted differently by computer models. These words are normalized based on the official dictionaries.

Data Analysis

Figure: Distribution of question types in VIMQA

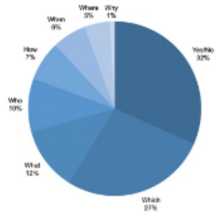
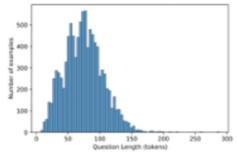


Figure: Distribution of question lengths in VIMQA



Type of reasonings

I. Inferring the bridge entity to complete the 2nd-hop question
II. Locating the answer entity by checking multiple properties
III. Reasoning types that require more than two supporting facts
IV. Comparing two entities
V. Identifying the Negation factor to answer Yes/No questions
VI. Identifying the Entity Swap to answer Yes/No questions

Types of answers in VIMQA

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 Trần Nhân Tông)*
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, tấn hố mang chìa *(chemistry, King cobra)*
Other	2	bằng thiết bị kết nối Internet *(with an Internet-connected device)*

List of Vietnamese Central Question Words

Group	English CQW	Vietnamese CQW
Yes/No	Copulas (is, are)	Phải không, Đúng không
	Aux (does, did)	
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

Benchmark settings

- **Gold Only setting** tests the ability to perform multi-hop reasoning to output the answer and sentence-level supporting facts to explain its answer. The models are provided with two gold paragraphs and a question requiring multi-hop reasoning
- **Distractor setting** tests the ability to find the answer and supporting facts when there are noises from the distracting paragraphs. 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.

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

Data splits of VIMQA

Experimental Results

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**

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

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**

Comparison of the models on VIMQA (Gold Only setting) and UIT-ViQUAD. The models are evaluated 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	51.45	**11.07**	21.94
BM25 + XLM-R$_{Large}$	32.30	43.80	20.64	52.86	10.97	**22.14**
BM25 + mBERT	**39.08**	**49.34**	18.04	51.33	7.87	18.30
Human	87.40	91.26	72.20	79.39	72.20	77.12

Comparison of existing methods in three sets of metrics on 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.

Conclusions

We propose VIMQA, a multi-hop QA dataset in the Vietnamese language. We also propose a pipeline for collecting multi-hop QA examples that can be generalized for all languages. The efficiency of the pipeline is proved via the detailed analysis in VIMQA. The experimental results indicate that VIMQA is challenging for competitive approaches in both single and multiple hop QA, and that our VIMQA dataset is a good resource for Vietnamese and cross-lingual QA models.