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 90 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**

- **WikiGraph**: Hyper-links are useful for building graph-based reasoning questions. We construct a directed graph from each article, where vertices are articles, and each edge $(u, v)$ indicates $u$ mentions $v$. Each article contains links to articles related by multiple types of edges.
- **Paraphrase Sets**: General concepts are collected from the dataset. Articles about particular people, events, papers, and titles are selected and linked. We then gather all articles straightforward to collect multi-hop questions.

**Processing and Normalizing**

- **Accent Unicode encoding**: Accented letters in Vietnamese such as "lá" can be encoded using either a single Unicode point (U+0061) 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**: "họ" and "họa" 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.

**Example of VIMQA**

- **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 multi-hop questions from the two given paragraphs. Three crowd workers who are researchers with Vietnamese native language annotated the VIMQA dataset. The dataset is then validated by more than one worker added to the dataset.

**Data Analysis**

- **Distribution of question lengths in VIMQA**: The distribution of question lengths shows a positive skew, indicating that questions tend to be longer than the average.
- **Types of answers in VIMQA**: The distribution of question types in VIMQA shows that most questions are closed-ended, with only a small fraction being open-ended.

**Experimental Results**

- **Gold Only**: The Gold Only setting is a zero-shot method. The models are evaluated in this VIMQA Gold Only setting, where only the gold answers are predicted. The result indicates that VIMQA is more challenging than QA datasets in English, which is the reason for the new multi-hop reasoning on the VIMQA dataset.

**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.