**Link paper:** [**https://www.researchgate.net/publication/352365285\_Unsupervised\_Multi-hop\_Question\_Answering\_by\_Question\_Generation**](https://www.researchgate.net/publication/352365285_Unsupervised_Multi-hop_Question_Answering_by_Question_Generation)

**Summary:**

**Introduction**

* EQA (Extracting Question Answering) là task trả lời câu hỏi bằng cách trích xuất một đoạn (span) từ văn bản đầu vào.
* EQA gồm 2 loại, là single hop và multi hop QA (QA suy luận)
* Việc xây dựng bộ dữ liệu multi-hop QA hao tổn nhiều nhân lực, tiền bạc. Nhóm tác giả đề xuất mô hình MQA-QG (Multi-Hop Question Generator) giúp sinh ra câu hỏi từ các nguồn thông tin đầu vào, giải quyết được 2 vấn đề trong MQA là
  + Homogeneous (thông tin đồng nhất): thông tin ở dạng text
  + Heterogeneous (thông tin không đồng nhất): thông tin ở dạng table
  + Unsupervised Question Answering

**Related work**

**Unsupervised QA**

Các phương pháp trước đây được dựa thử nghiệm trên bộ SQuAD (single-hop và information input là text), khiến cho cặp QA thiếu sự tích hợp và suy luận vốn có của multi-hop QA, và không thỏa được các hệ tri thức đầu vào (knowledge base) vốn được tổ chức dưới dạng bảng.

**Multi-hop Question Generation**

Tuan el đề xuất multi-state attention để bắt chước khả năng suy diễn. Pan el sử dụng semantic graph trên đoạn văn đầu vào, từ đó tiến hành suy diễn qua các thực thể khác nhau. Nhưng cả hai cách trên đều mất thời gian để làm data.

**Methodogy**

**Operators**

* QGwithAns, QG withEnt: Từ dữ liệu tập train SQuAD, thu được tập D gồm context-question-answer (c-q-a).
  + Từ c, a -> q (Question generate with answer)
  + Từ c, e -> q (Question generate with entity)
* Describe Ent: Từ table T và entity e được xác định trong bảng, apply model GPT-TabGen2 (Model flatten table thành text, để thông tin được chính xác thì chỉ flatten hàng chứa entity được xác định
* BridgeBlend: Kết hợp 3 thành phần lần lượt là bridge entity, single-hope question that contains bridge entity và câu mô tả bridge entity s tạo thành một câu hỏi. Sau đó tiến hành che đi entity trong câu hỏi, dùng BERT-large để điền từ thay thế cho entity (dựa trên câu mô tả s)
* CompBlend: Từ hai câu hỏi single-hop q1 và q2 chứa các comparative property entities e1, e2 (thực thể tiềm năng, phân biệt với thực thể cầu nối (bridge entity), kết hợp lại để tạo thành comparison-type multi-hop question

**Reasoning graph**

**Question filtration (Lược bỏ những câu hỏi không liên quan và thiếu tự nhiên)**

Filtration: Loại bỏ những câu hỏi thiếu tự nhiên, mạch lạc bằng cách sử dụng model pretrained GPT-2. Các câu hỏi với PPL score (perplexity-độ hỗn loạn) thấp nhất được đưa vào tập train.

Paraphrasing: sử dụng model BART để tiến hành paraphrase các câu hỏi được tự nhiên hơ, nhưng điều này cũng tạo ra vấn đề semantic drift (chuyển nghĩa, có thể hàm ý xấu nghĩa hoặc tốt nghĩa), gây ảnh hưởng đến kết quả đánh giá của task QA.

**Experiments**

Tiến hành đánh giá trên 2 bộ dữ liệu là HotpotQA (text) và HybridQA (text n tabular)

HotpotQA ~100K multi-hop QA pairs crowd-sources.

HybridQA ~70K human-labeled multi-hop questions, ứng với mỗi Q sẽ là một table từ Wiki và những đoạn text linked với từng entities trong table.

**Task Question Generation:**

Lấy thông tin từ bảng và text trong bộ HybridQA, tạo ra 2 training set txt->tbl và tbl->txt bằng cách excute reasoning graph (truy vấn đồ thị suy diễn) là Text-to-Table và Table-to-Text với 170K cặp QA. Từ HotpotQA, tiến hành generate Q bge và Q com, sau đó merge và filter, từ đó tạo thành 100K cặp QA

**Task Question Answering:** Sử dụng model HYBRIDER trên HybridQA và SpanBERT trên HotpotQA.

**Baselines:** Tiến hành đánh giá MQA-QG trên baselines của cả 2 phương pháp học có giám sát và không giám sát. <https://qa.fastforwardlabs.com/no%20answer/null%20threshold/bert/distilbert/exact%20match/f1/robust%20predictions/2020/06/09/Evaluating_BERT_on_SQuAD.html>

**Ablation Study**

* Việc kết hợp training dữ liệu bằng MQA-QG và fine tune với một ít dữ liệu được gán nhãn bởi con người sẽ giúp mô hình đạt được độ chính xác cao với trường hợp nghèo dữ liệu huấn luyện, giảm thiểu sự phụ thuộc lên con người. Có thể coi như là một sự “khởi động” đối với một số hệ thống QA khi áp dụng lên các domain mới mà không phải tốn quá nhiều tài nguyên cho dữ liệu huấn luyện.
* Nhóm tác giả muốn câu hỏi tạo ra được lưu loát hơn bằng cách áp dụng kĩ thuật paraphrase, nhưng kết quả đưa ra thấp hơn, xảy ra vấn đề semantic drift. Vì thế họ ưu tiên độ chính xác của câu hỏi hơn là độ lưu loát, nhưng vấn đề lưu loát trong câu hỏi vẫn sẽ là mục tiêu trong tương lai của nhóm phát triển.