Generating Question-Answer Hierarchies

Kalpesh Krishna & Mohit Iyyer

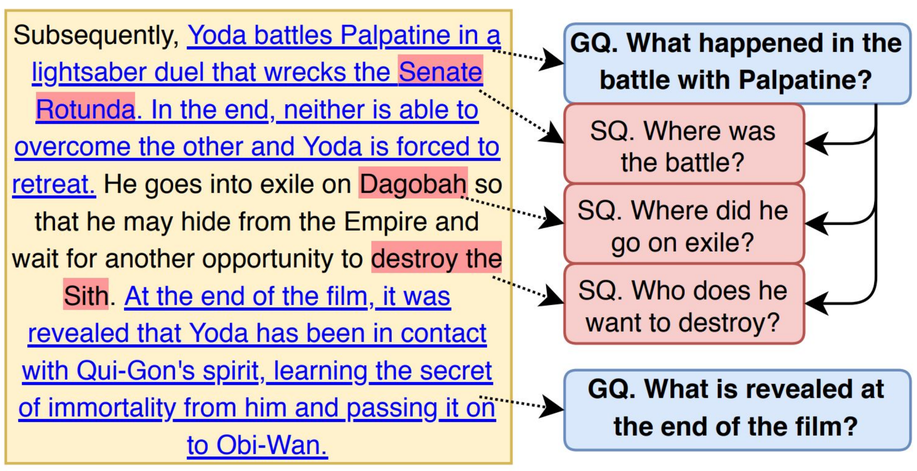
UMassAmherst
College of Information & Computer Sciences

code, dataset, demo at http://squash.cs.umass.edu

Q1. What is this paper about?

A novel text generation task, Specificity-controlled Question Answer Hierarchies (SQUASH)

Input - sequence of paragraphs
Output - hierarchy of Question-Answer (QA) pairs arranged according to their specificity



Q. What is the difference between blue and red QA?
Top level (blue) - general, broad, overview questions
Bottom level (red) - specific, drill-down questions

Q2. How did we construct a dataset to train a SQUASH system?

- Collecting large-scale annotated data is expensive
- Instead, label questions in existing reading comprehension datasets (SQUAD, CoQA, QuAC) according to their specificity

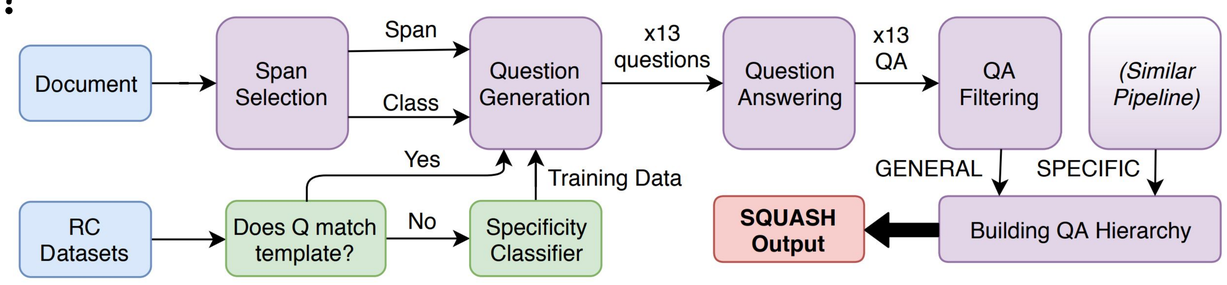
Q. What is the scheme for specificity labelling?
Templates from Lehnert 1978 (48.5%), hand-labelling (0.5%), distant supervision (51%) - crowdworkers agree

Lehnert 1978 category	Examples	Specificity
Causal Antecedent / Consequent, Enablement	*Why ..., What happened after / before ..., What led to ...*	general
Quantification	*How many ..., How long ...*	specific
Concept Completion / Feature Specification	*What is computer science?, Who invented the computer?*	general / specific

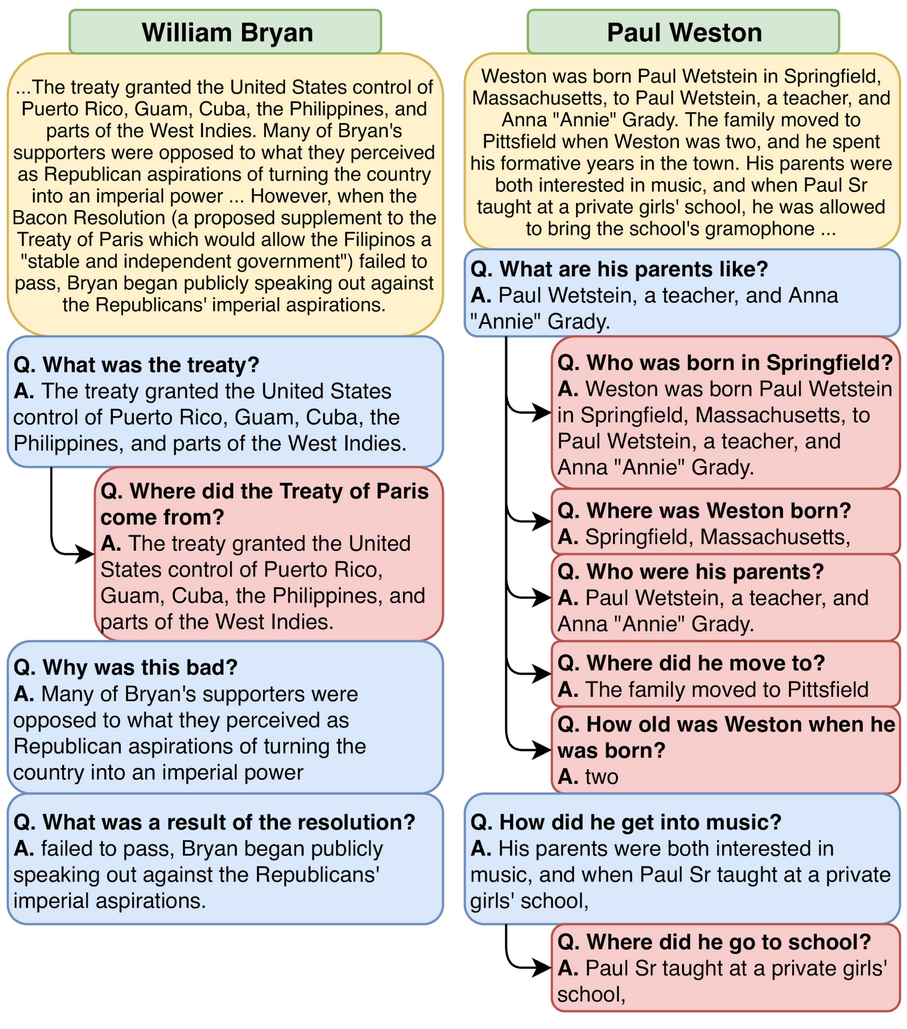
Q. What does the final dataset look like?
A total of 277098 QA pairs (27.8% general, 54.2% specific, 18% yes/no questions)

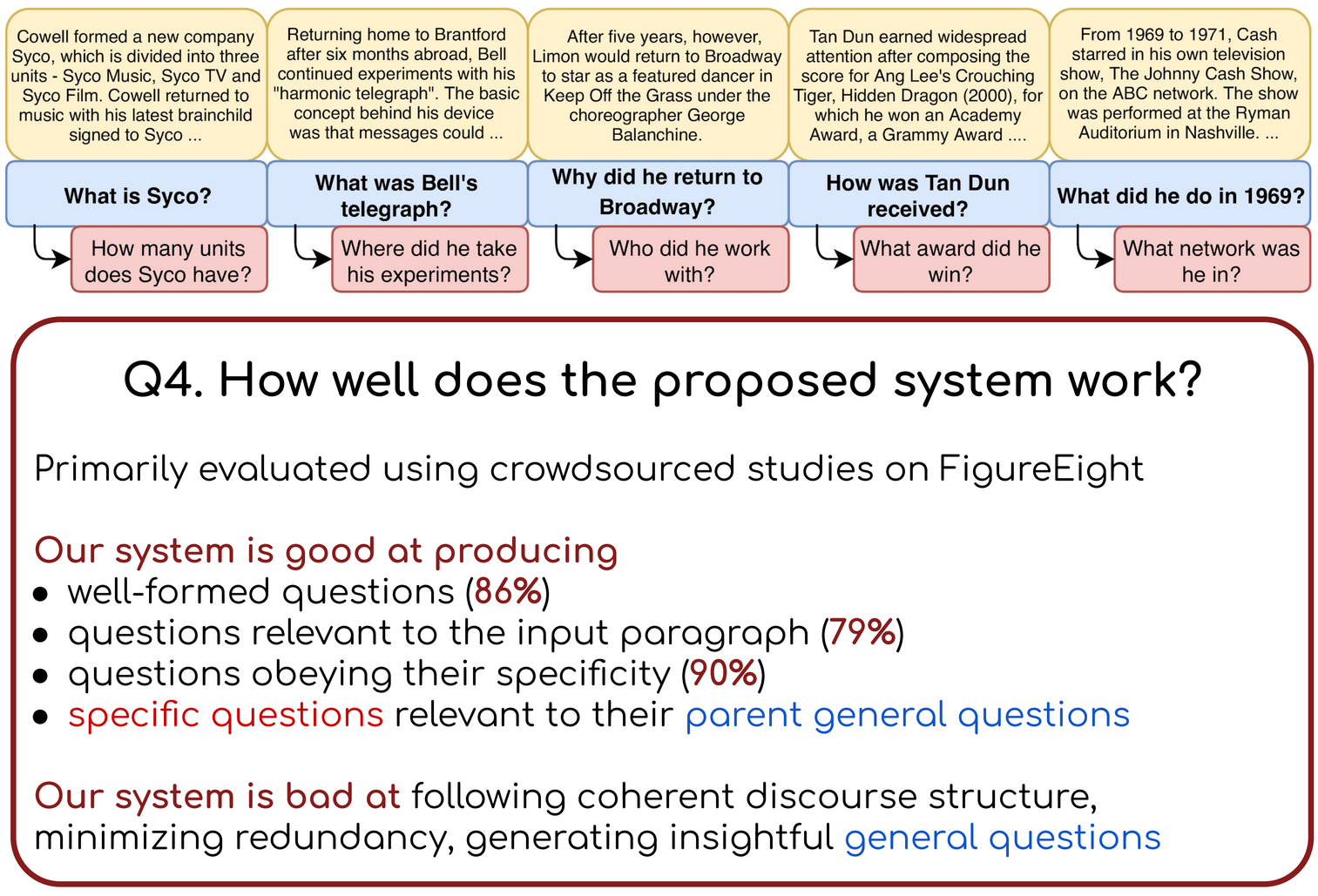
Q3. How do we SQUASH input paragraphs?

1. Train a specificity and answer conditioned neural question generation model
2. Generate general questions using full sentences and specific questions using entities and noun-phrases



3. Answer each generated question using QA system and filter bad questions (low answer overlap, unanswerable)
4. Construct hierarchy based on answer overlap and position of the predicted answers





Q4. How well does the proposed system work?

Primarily evaluated using crowdsourced studies on FigureEight

Our system is good at producing
- well-formed questions (86%)
- questions relevant to the input paragraph (79%)
- questions obeying their specificity (90%)
- specific questions relevant to their parent general questions

Our system is bad at following coherent discourse structure, minimizing redundancy, generating insightful general questions

Q5. Does SQUASH improve pedagogy?

- Support for FAQs, hierarchies and QA mode of communication in HCI, CogSci, Socrates' teaching
- Needs user-studies with accurate SQUASH outputs

Q6. What's new in our live demo?

An improved system with GPT-2 and BERT pre-training, coreference-resolved questions and customizable inference hyperparameters for generation and filtering