Question Generation from Sentences

Task specification ***Version 2****[*], 24 January 2010*

Brendan Wyse, Paul Piwek and Svetlana Stoyanchev

Contact: *bjwyse@gmail.com*

TASK TITLE: Question Generation (Sentence Level)

BRIEF DESCRIPTION: Participants are given a set of inputs. For each input, consisting of an input sentence + question type, their system should generate 2 questions.

TASK DETAILS: Each input consists of:

- a single sentence and
- a specific target question type (e.g., WHO?, WHY?, HOW?, WHEN?, etc.; the possible question types are described in the next section).

For each input, participants are asked to generate 2 questions.

Sentences will be selected from the primary data sources for the QGSTEC 2010 (Wikipedia, OpenLearn and Yahoo! Answers or similar). Sentences which are extremely long or extremely short will be avoided.

60 inputs from each data source will be provided and a development data set will be published in advance of the actual evaluation data set.

The questions generated will be judged based on a range of criteria (these are discussed below). We will be reporting how systems perform against each of the individual criteria and also compute a score which combines the scores for all of the criteria. The latter combined score will allow us to provide an overall ranking of the participating systems.

Here are three examples of an input (involving the same input sentence, but different target question types) and a possible output from a QG system for each:

[*] This is not the final version of the Task Specification. This document is intended for soliciting feedback from the Question Generation community. The final version will explicitly state that it is the "final version".

INPUT SENTENCE:

Abraham Lincoln (February 12, 1809 – April 15, 1865), the 16th President of the United States, successfully led his country through its greatest internal crisis, the American Civil War (1861 – 1865).

TARGET QUESTION TYPE: WHEN?

OUTPUT:

(1) In what year was Abraham Lincoln born

(2) In what year did the American Civil War commence?

INPUT SENTENCE:

Abraham Lincoln (February 12, 1809 – April 15, 1865), the 16th President of the United States, successfully led his country through its greatest internal crisis, the American Civil War (1861 – 1865).

TARGET QUESTION TYPE: HOW MANY/LONG?

OUTPUT:

(1) What age was Lincoln when he died?

(2) How many years did the American Civil War last?

INPUT SENTENCE:

Abraham Lincoln (February 12, 1809 – April 15, 1865), the 16th President of the United States, successfully led his country through its greatest internal crisis, the American Civil War (1861 – 1865).

TARGET QUESTION TYPE: WHICH?

OUTPUT:

(1) Which President of the United States led his country through the American Civil War?

(2) Which war was the greatest internal crisis of the United States?

QUESTION TYPES:

The list of question types below is final for this task. We acknowledge that there are other question types which we have not included in our challenge at this stage but we would rather see participants focus on this particular set of question types.

WHO?

> The answer to the generated question is a person (e.g. Abraham Lincoln) or group of people (e.g. the American people) named in the input sentence.

WHERE?

> The answer to the generated question is a placename (e.g. Dublin, Mars) or location (North-West, to the left of) which is contained in or can be derived from the input sentence.

WHEN?

> The answer is a specific date (e.g. 3rd July 1973, 4th July), time (e.g. 2:35, 10 seconds ago), era or other representation of time.

WHICH?

> The answer will be a member of a category (e.g. Invertebrate or Vertebrate) or group (e.g. Colours, Race) or a choice of entities (e.g. Union or Confederacy) given in the input sentence.

WHAT?

> The question might describe a specific entity mentioned in the input sentence and ask what it is. The question may also ask the purpose, attributes or relations of an entity as described in the input sentence.

WHY?

> The question asks the reasoning behind some statement made in the input sentence

HOW MANY/LONG?

> The answer will be a duration of time or range of values (e.g. 2 days) or a specific count of entities (e.g. 32 counties) within the input sentence.

YES/NO

> The generated question should ask whether a fact contained in the input sentence is either true or false (e.g. Are mathematical co-ordinate grids used in graphs?).

SUBMISSION EXAMPLES:

The following examples show a selection of inputs (sentence + question type) and accompanying sample questions. Where we use the same sentence for several inputs, we have grouped these together. ***In the actual task, each input always consists of a single sentence and a single question type for which 2 questions need to be generated.***

Note that most sample questions are deliberately constructed using the entire phrases from the original input sentence and consequently do not always read fluently. This means that these questions do not necessarily receive the top score under the evaluation criteria that we present further on in this document.

In the very first example below we provide the sample question #1 as '*Who was the artist who was raised amid the heroism and turmoil of Napoleon's regime but whose artistic career began in earnest after Waterloo?*'. Systems capable of generating the more natural '*Who was the artist who was raised amid Napoleon's regime?*' will receive a better score for fluency in the evaluation.

Example 1

Taken from the OpenLearn study unit about the artist Delacroix

> *Ferdinand-Victor-Eugène Delacroix (1798–1863) was an artist raised amid the heroism and turmoil of Napoleon's regime but whose artistic career began in earnest after Waterloo.*
>
> Target Question Types: WHO?, WHEN?, HOW MANY?

Possible generated questions:

1. *Who was the artist who was raised amid the heroism and turmoil of Napoleon's regime but whose artistic career began in earnest after Waterloo? (WHO?)*
2. *Delacroix was raised amid the heroism and turmoil of who's regime? (WHO?)*
3. *In what year was Delacroix born? (WHEN?)*
4. *When did Delacroix's career begin in earnest? (WHEN?)*
5. *What age was Delacroix when he died? (HOW MANY/LONG?)*[†]
6. *How old was Delacroix when he passed away? (HOW MANY/LONG?)*

[†] Note that the answer requires the subtraction of two cardinal numbers which appear in the input sentence. The question is actually asking 'how many' years between Delacroix's birth and death as given in the input sentence.

Example 2

Taken from the OpenLearn study unit ‘Finding information in business and management’

> *Your approach to searching will depend to a great extent on what kind of person you are.*
>
> Target Question Type: WHAT?

Possible generated questions:

(1) What will your approach to searching depend on to a great extent? (WHAT?)

(2) Your approach to searching will depend to a great extent on what? (WHAT?)

Example 3

Taken from the OpenLearn study unit 'Europe and the law'

> *The Council of Europe was set up in 1949.*
>
> Target question type: WHEN?, WHICH?

Possible generated questions:

(1) When was the Council of Europe set up? (WHEN?)

(2) In what year was the Council of Europe set up? (WHEN?)

(3) Which council was set up in 1949? (WHICH?)

(4) In 1949, which council was set up? (WHICH?)

Example 4

Taken from the OpenLearn study unit 'Diagrams, charts and graphs'

> *Mathematical co-ordinate grids are used for maps, plans, geometric diagrams and also for graphs.*
>
> Target question type: WHAT?, YES/NO

Possible generated question:

(1) What are mathematical co-ordinate grids used for? (WHAT?)

(2) Apart from maps, plans and geometric diagrams, what are mathematical co-ordinate grids also used for? (WHAT)

(3) Are mathematical co-ordinate grids used for graphs? (YES/NO)

(4) Are mathematical co-ordinate grids used for maps? (YES/NO)

Example 5

Taken from the Wikipedia article for Phagocyte

> *Phagocytes are important throughout the animal kingdom, and are highly developed in vertebrates.*
>
> Target question type: WHICH

Possible generated question:

(1) *In which type of animals are phagocytes, which are important throughout the animal kingdom, highly developed?(WHICH?)*

(2) *Which type of animals have highly developed phagocytes?*

Example 6

Taken from the Wikipedia article for Sonia Sotomayor

> *In 1997, Sotomayor was nominated by President Bill Clinton to the U.S. Court of Appeals for the Second Circuit.*
>
> Target question type: WHO?, WHEN?

Possible generated questions:

(1) *Who was nominated in 1997 by President Bill Clinton to the U.S. Court of Appeals for the Second Circuit? (WHO?)*

(2) *Who nominated Sotomayor in 1997 to the U.S. Court of Appeals for the Second Circuit? (WHO?)*

(3) *When was Sotomayor nominated to the U.S. Court of Appeals for the Second Circuit? (WHEN?)*

(4) *In what year was Sotomayor nominated by Clinton to the US Court of Appeals for the Second Circuit? (WHEN?)*

DATA SOURCES: Wikipedia, OpenLearn, Yahoo!Answers and similar

Data Size: 60-60-60, i.e. 60 inputs (sentence + target question type) from each of the three sources above.

INPUT SENTENCES:

Input sentences will be provided as raw text. Annotations will not be provided. There are a variety of NLP open-source tools available to potential participants and the choice of tools and how these tools are used is a fundamental part of the challenge.

EVALUATION

Evaluation will involve:

- independent human raters,
- peer-rating, and
- possibly automated rating through pooling.

GUIDELINES FOR HUMAN JUDGES

We will ask judges to rate questions on the following criteria (mostly on a scale from 1 to 4 with 1 being the best score). (Additionally, if a question is missing in the output, judges are instructed to assign the worst score on all criteria to that question and also the worst score for VARIETY for the other output question).

Here are the criteria and some brief examples of how we should mark each criterion:

RELEVANCE

Questions should be relevant to the input sentence. This criterion measures how well the question can be answered based on what the input sentence says. Systems should aim to generate rank 1 questions as the score in relevance limits the possible overall score (i.e., a question will receive the worst overall score if it has rank 4 for relevance, even if it scores well on other criteria). This is to ensure that only questions related to the input sentence are rewarded.

Rank	**Description**
1	The question is completely relevant to the input sentence.
2	The question relates mostly to the input sentence.
3	The question is only slightly related to the input sentence.
4	The question is totally unrelated to the input sentence.

QUESTION TYPE

Questions should be of the specified target question type.

Rank	**Description**
1	The question is of the target question type.
2	The type of the generated question and the target question type are different.

SYNTACTIC CORRECTNESS AND FLUENCY

The syntactic correctness is rated to ensure systems can generate sensible output. In addition, those questions which read fluently are ranked higher.

Rank	**Description**	**Example**

1	The question is grammatically correct and idiomatic/natural.	In which type of animals are phagocytes highly developed?
2	The question is grammatically correct but does not read as fluently as we would like.	In which type of animals are phagocytes, which are important throughout the animal kingdom, highly developed?
3	There are some grammatical errors in the question.	In which type of animals **is** phagocytes, which are important throughout the animal kingdom, highly developed?
4	The question is grammatically unacceptable.	**On** which type of animals **is** phagocytes, which are important throughout the animal kingdom, developed?

AMBIGUITY

The question should make sense when asked more or less out of the blue. Typically, an unambiguous question will have one very clear answer.

Rank	**Description**	**Example**
1	The question is unambiguous.	Who was nominated in 1997 to the U.S. Court of Appeals for the Second Circuit?
2	The question could provide more information.	Who was nominated in 1997?
3	The question is clearly ambiguous when asked out of the blue.	Who was nominated?

VARIETY

Pairs of questions in answer to a single input are evaluated on how different they are from each other. This rewards those systems which are capable of generating a range of different questions for the same input.

Rank	**Description**	**Example**
1	The two questions are different in content.	Where was X born?, Where did X work?

2	Both ask the same question, but there are grammatical and/or lexical differences.	What is X for?, What purpose does X serve?
3	The two questions are identical.	

★ ★ ★