Progress Report I

1. Team Members

Rainer Nunez (rnunez), Zhenzhen Weng (wengz), Xi Liu (xiliu1), Yiting Hao (yitingh)

2. Progress

In the past weeks we worked on question generation, and completed question generation modules covering most of easy and medium questions. Now we are performing unit testing on each module to fix problems , and setting up pipeline for parsing documents, generating questions, and selecting questions. We broke down the question system into modules and assigned tasks accordingly.

Task	Developer	Tester
Who/what	Yiting	Rainer
When/where	Xi	Zhenzhen
Why/how many	Zhenzhen	Rainer
Yes/no	Rainer	Xi

So far we could generate following types of questions with strategies.

1) Yes/no:

For these kinds of questions, we first determined what the first words in the questions should be 'Does', 'Did', 'Is', 'Are', 'Has', or 'Have'". Using nltk POS tagging, we are able to find such helping verbs (MD) and adjust the string accordingly. In order to generate questions starting with 'Did' or 'Does', we first check if the sentence starts with either a preposition, a singular proper noun, or a plural proper noun, then we remove '-ed'/'-s' for a VBD/VBZ. If the verb contains the suffix 'ed', we delete the 'ed', put 'did' to the front of the statement, and substitute the period with a question mark.

Example
input:
["David Beckham has got a nice car.", "John's parents always traveled to Spain for vacation.", "Ronaldo currently plays for Real Madrid FC."]
output:
["Has David Beckham got a nice car?", "Did John's parents always travel to Spain for vacation?", "Does Ronaldo currently play for Real Macrid FC?"]

2) Who/what

Who/what questions could be generated if the sentences contain subjects (all most every sentences have subjects). The first step to do this is to tag each word in each sentence. Then find the nouns or pronouns in each sentence. If the nouns is labels as ‘NNP‘, then this is the answer for our question. If the nouns is not labels as ‘NNP’, instead with some pronouns, we find its nearest sentence that contains NNP nouns and treat it as the answer.Here are some examples for “Who” question:

Example
Input:
[“Jim is a student.”,” He should go to school.”, “Tom is the friend of Jim.”, “He is calling Jim to come to school.”]
Output:
[“Who is a student?”, “Who should go to school?”, “Who is the friend of Jim?”, “Who is calling Jim to come to school?”]

For “What” question generation, we use the same method.

Example
Input:
[“Both Earth and Mars are in solar system.”, “Banana, apple, grapes and oranges are all my favorite fruits.”]
Output:
[“What are in solar system?”, “What are all my favorite fruits?”]

However, here we can’t handle the case that negation occurs in the sentence and we can’t distinguish nouns are for things or for people’s name. See more details in our Problem section.

3) When/where

When/ where questions could generated based on yes/no questions. We could transform the statement into a yes/no question first, then identify a location or time in the question by Name entity recognition. Then we could substitute PREP + TIME with when, PREP + LOCATION/ORGANIZATION with where, and move the substitution to the front. We utilized Stanford NER tagger and POS tagger to find this pattern. Adverbial clause is also very common for describe time or locations. To handle these cases, we simply look for keywords like “when”, “where”, “before”, “after”, “since”, “during” and remove everything before a comma or question mark. If there are more than one phrases or clause for time/location, we treat them as one piece.

Example
input:
[“Tom worked at Carnegie Mellon University in Pittsburgh after he graduated from MIT.”]
output:
[“Where did Tom work after he graduated from MIT?”]
[“When did Tom work at Carnegie Mellon University in Pittsburgh?”]

4) Why

Why questions should be generated when we see subordinating conjunctions such as “because”, “since”, “so” and “due to”. Depending on which subordinating conjunction we see, we can determine whether the part of sentence that comes before it is the reason or the consequence. We will then ask the question “Why (consequence)?” We will need to alter the word orders in the (consequence) to make the question grammatically correct. Here is one example.

Example
input:
[“He attended church every week with his parents, because that was the only way he could play football for their team.”]
output:
[“Why did he attend church every week with his parents?”]

5) How many

How many questions will be generated if the sentence contains a quantity. For example, if the generator sees CD+NN/NNS/NNP/NNPS, then it will extract the noun that comes before the cardinal number and ask “How many +NN/NNS/NNP/NNPS + (the rest of the sentence reordered)?” The generator can now handle the following example.

Example
input:
[“Cancer contains two stars.”]
output:
[“How many stars does cancer contain?”]

One further consideration would be to differentiate the plural mass nouns and singular nouns and generate “How much” questions when it is mass noun.

3. Problems

In unit testing, we found the following problems and plan to solve them.

1) Appositions

We found difficult to generate proper questions with statements containing appositions. for example, in statement “Tom, a student at CMU, studies very hard.”, “Tom” and “a student at CMU” refer to the same person and we should treat them the same, not as two different persons. Also, we should be able to generate “Who is Tom?” as question, and “A student at CMU” as an answer for it.

We plan to utilize Stanford Parser to generate the syntax tree to identify appositions. After handling appositions, we should be able to generate hard questions.

2) Who/what:

We noticed some issues that we will be facing with generating questions involving ‘Who’ and ‘What’. We need to make sure that our system can identify when a statement’s noun is referring to a person before we stick ‘who at the beginning of a question generated based on the statement. We found that nltk has a HumanName tagger that would solve this problem for us, but we would still have to find a way to address situations like “The student”, or “The professor” where a question with ‘who’ is still suitable.

Right now we have an issue with our function that generates ‘what’ questions where it just substitutes the nouns with the word ‘what’. A sample statement would be “Earth and Mars are planets.” Our function right now would generate “What are planets?” This should be fairly simple to fix, by just reversing the logic for classifying what is the subject of the question.

4. Plan

We set up a roadmap to guide future development.

Date	Goal	Task Break-downs
29 Feb 2016	Complete question generation modules and setup a pipeline	Fix who/what module: Rainer, Yiting Setup pipeline: Xi, Zhenzhen Unit testing and improvements: in pair
7 Mar 2016	Complete question selection and ranking	Survey question selection strategies Predication selection Question evaluation and selection
14 Mar 2016	Complete question system	Fix problems Question quality improvement
17 Mar 2016	Progress Report II and video	Video, report
21 Mar 2016	Complete answer generation modules	Question classification Question-statement matching Answer selection
28 Mar 2016	Setup answer pipeline and improve precisions	Setup pipeline Improve precisions
4 April 2016	Dry run, improve question qualities, and answer precision	TBD
15 April 2016	Code submission	N/A
22 April 2016	Final video	Video