# 15-112 Project proposal: A Question Answering System

Yunze Xiao

November 8, 2022

## 1 Introduction

The project is an adoption of the end of course project on 11-411 Natrual Language Processing.In the upcomming sections, we will describe the project and talk about the structure of the solutions. We will also mention some of the obstacles of the project and our approch to solve them. The system may or may not have all the parts mentioned in the description however, but is guarenteed to have at least 1 part fully implemented with Graphical User Interface.

## 2 Project Description

### 2.1 Question generation

Given a document (for example, the text1 of a document from Wikipedia), generate a set of questions relevant to that document. For example, if the document contains the sentence:

> 15-112,one of the best CS introduction courses in the World, was created by David Kosbie, a CMU professor in 2011

The question Generation system may generate these questions:

1. What is 15-112?
2. Who created 15-112?
3. When was 15-112 created

The goal is to generate questions that are **fluent** and **reasonable-** meaning that they have to have answers and be gramatically correct.

## 3 Competitive Analysis

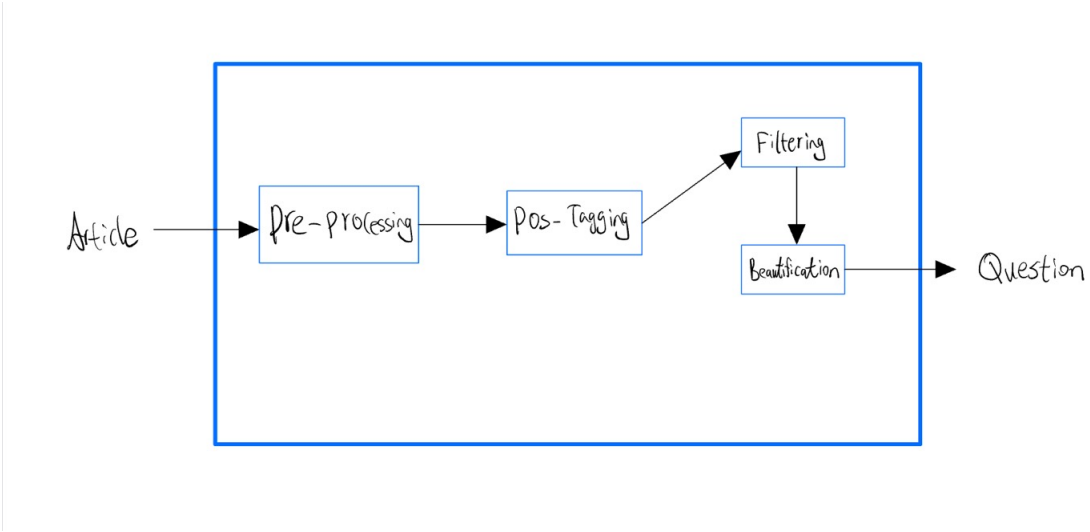
Since the project is a partial adaptation of 11-411, the structure and algorithm used in the code will be utterly similar to the counter parts project in 11-411. However, the major difference is that this project has some more realistic and applicable impelmentation. Specifically, the 15-112 project iteration are not scoring for competition but rather put into practice by allowing regular users to answer the generated questions

## 4 Structural Plan

The code will be structured in such way

- Question Generation
  - Binary questions
  - What/Who Question
  - Why
  - How many
  - Whose
  - When
  - Where
- GUI
  - Main Frame
  - Admin Version
    - allows the user to use generate questions
    - allows the user to overview the text corpus
  - regular version
    - allows the user to answer questions
- Corpus Visualization
  - Shows the user the word demography
    - Top 10

# 5 Algorithmic Plan




## 5.1 Preprocessing

- Sentence Split and Vectorization (through stanza)
- Part of speech tagging through stanza
- Clause Extraction

## 5.2 Question Generation

- Identify the subject(noun) and verb Phrases (named NP and VP)
- Restructure the sentence based on question type or NP/VP
- Able to do multiple question for the same sentences

## 5.3 Filtering

- Remove similar questions by finding intersection
- Beautification: implement correct syntax
- Tense problem
- Casing

# 6 Version Control

The code will be posted on Github

# 7 Timeline

**Oct.21** Finish question Generation implementation

**Oct.25** Finish GUI

**Oct.26** Final Test

# 8 Module List

- Stanza
- HuggingFace(used for dataset )
- nltk (to manipulate the parse tree)
- matplotlib (show text corpus demographics)

# 9 TP2 updates

## 9.1 New Features

- Created User Interface for the system
- Utilizes time and os libaries to add additional features
- Bug Fix

## 9.2 On Progress

- Still need to finish the answering part of the question
- Question ranking and beautification are still pending.

# 10 TP3 updates

## 10.1 New Features

- Added where question
- Bug Fix
- Beautification implemented

- More organized file generation
- Added citations and readme, and requirements.txt
- Shuffle implemented as a replecement for question ranking