MSc project plan: Intelligent Chat Agent for Q/A on Programming using StackOverflow as Knowledge base

130533 - Knut Lucas Andersen (knut.andersen@hig.no)

Master's Thesis Project Description
Master of Science in Applied Computer Science
5 ECTS
Department of Computer Science and Media Technology
Gjøvik University College, 2015

Avdeling for
informatikk og medieteknikk
Høgskolen i Gjøvik
Postboks 191
2802 Gjøvik

Department of Computer Science
and Media Technology
Gjøvik University College
Box 191
N-2802 Gjøvik
Norway

Revision history

Version #	Description of change (why, what where - a few sentences)
0.1	First version: Problem description (Chapter 1)
0.2	Second version: Added sections to Chapters 2, 3 and 4. Added Appendices for the Survey and Interview format used in IMT5251 Advanced Project Course (see p. 21 and p. 23). Updated Section 1.6: Planned contributions. Added content to the following Chapters: 4.

Abstract

When first starting to learn how to program, there is a lot of information to take in. There are a lot of different programming languages, some which have their own rules on how to build and execute the developed program. There are tons of different algorithms and different ways a problem can be solved, not to forget all the different terminologies and semantics that exists within the field. There is also a great amount of online resources on the Internet, ranging from encyclopaedias (e.g. Wikipedia), tutorial sites (e.g. TutorialsPoint, W3Schools, HackThisSite, etc), to online communities (e.g. StackOverflow, CodeProject, etc).

When looking for information, searching for an answer, or looking for a solution to a problem, it is not always that easy to come up with a good question (what defines a good question?). When learning to program in class, the questions may not come right away. Seeing something explained on the blackboard is quite different from actually understanding and doing it yourself. Looking for answers online can become quite time-consuming, since the relevance of the results returned by the search engine vary in a large degree. The answer to the question asked may not appear until page 10 of the returned results.

To help students learning programming, the goal of this thesis is to create a plug-in for the Learning Management System (LMS) Open Edx. The plug-in to be created is an Artificial Intelligent Chat Agent (aka. a ChatBot), and the students can then ask this Chat Agent questions related to programming. The Chat Agent will use StackOverflow as its Knowledge base, meaning the answers presented to the user will be based on the answers posted by users on StackOverflow. By using this Chat Agent, students can ask questions in the same way they would ask a teacher or a classmate, instead of having to use keywords when using a search engine. The goal of this thesis is to see if the Chat Agent can aid students that are learning programming, but also do research on it (e.g. trying to humanize the Chat Agent (Turing Test)). An example of the experiment would be to have two test groups (A/B testing), where the comparison would be on the grades of the students using the Chat Agent vs. those not using it, to see if the Chat Agent had any effect on the students grade.

Contents

1 Introduction

1.1 Topic covered by the project

The goal of this thesis is to develop an Artificial Intelligent Chat Agent (aka. ChatBot), which will function as a plug-in in the LMS Open Edx[1]. This Chat Agent is targeted at students learning programming, and will therefore be used to answer the students questions related to programming. The answers from the Chat Agent will be based on the content found in the online community StackOverflow[2]. Since StackOverflow is one of the many communities belonging to the StackExchange network[3], the Chat Agent can later on be expanded to cover information from other communities.

The hope is that this Chat Agent can aid the students in their learning progress, since they can ask the Chat Agent questions in the same way they would ask their teacher or their classmate. They do not have to think about keywords or read through a lot text, since the Chat Agent will try to answer their questions based on the answers it finds in the posted answers on StackOverflow.

From a research perspective, it would be interesting to see if the Chat Agent will have any effect on the learning outcome and what the students think of the Chat Agent, e.g. is it useful or would they just prefer to continue using search engines and look for the answer(s) themself. As for the Artificial Intelligence (AI) side, it would be interesting to see what can be done to "humanize" the Chat Agent so that it could pass the Turing test[4]. It would also be interesting to look more deeply into the AI algorithms, to see in what way they can be improved or extended to cover a larger base of linguistics.

1.2 Keywords

Intelligent Agent, Chatbot, Natural language processing, Human–computer interaction, Education, Question-answering

1.3 Problem description

Can we find the answer if we do not know the question? The issue with most search engines today is that they are based on taking each search word (the Term) to create what is called a Dictionary (which contains all the words/terms searched for). It then looks through its content (documents, files, multimedia, etc) and searches for each term, and returns those that contain at least one of the terms, ranking the results according to frequency of the given term. Although a given amount of search results is returned, there is no guarantee that the answer searched for are among the returned results. While programming, if an issue occurs, sometimes you can find the answer by using a few keywords, or simply copy/pasting the error message. But what do you search for when

[1] Open Edx: `https://open.edx.org/`

[2] StackOverflow: `http://stackoverflow.com/`

[3] You can see all the StackExchange communities here: `http://stackexchange.com/sites`.

[4] A Turing Test is a test where a human is asked to converse with a unknown party, and then later on decide whether the party was a human or a computer ([p. 2][1]).

your question is more abstract? What do you search for when you have a question, but are struggling with phrasing it in a way that a search engine can understand? What do you do when you have a question that a teacher or a classmate could easily answer, but the search engine cannot?

With a Chat Agent, you do not have to think about keywords, phrases or "words best describing the problem". You can just ask the question you want an answer to. You also get anonymity with a Chat Agent. You can ask all sorts of questions, no matter how dumb you feel they are, because the Chat Agent is there to help.

A keypoint to remember is that the Chat Agent is intended to function as a help tool (e.g. FAQ FINDER [2] and Bzz [3]), and not as a replacement for the teacher (e.g. CALMsystem [4]).

1.4 Justification, motivation and benefits

With the advancement of programming and the increase of different languages, libraries and functionality, it can be hard for a teacher to cover all topics. It can also be hard for a student to grasp everything at once and understand everything the code does. Although there is a wast amount of online resources on the Internet, finding what you need can take a lot of time, but when using a Chat Agent, you can just ask the question and get the answer right away. Since the Chat Agent will be based on Hidden Markov Model (HMM), it can remember everything previously discussed in the conversation, increasing the chance of it helping the student finding the desired answer. The Chat Agent can also help the teachers, e.g. when running exercises in class, students can use the Chat Agent to find answer to the simpler questions, and then ask the teacher for help on the more advanced and problematic issues.

One example is the paper by [3] were they used an existing ChatBot called Bzz for answering adolescents' questions related to sex, drugs and alcohol. Their study showed that the users used the ChatBot a lot[5], and that the users felt it was faster and better then information lines and search engines. [5] says that a ChatBot can be easier to converse with due to its anonymity. Furthermore, teachers can look at the conversation logs to see what the students have discussed, to be able to map the problems and see how students learn.

[6, p. 268] did two case studies where they compared the use of ChatBots and e-learning in relation to Information Security. They measured the ChatBot experience qualitatively, and 70% of the users found the ChatBot useful and would use one in the future. However, quantitatively they found no significant difference between those using a ChatBot and those using e-learning. CSIEC (documented in [7]) is a ChatBot developed for learning english, and was tested in [8]. The students achieved a very high score at the exam, but as the authors note, CSIEC was tested only between two tests (and there is also a chance of bias, since one of the authors is also the developer of CSIEC).

[5] "42,217 conversations with the chatbot; thus, an average of 11.3 conversations with each lasting 3 minutes and 57 seconds" [p. 516][3].

1.5 Research questions

- How was the Chat Agent perceived by the user (e.g. using the Chat Agent vs. using a search engine)?
- Did the Chat Agent help the user understand/learn more about programming?
- By using A/B testing, is there a (statistical) improvement from the students using the Chat Agent, vs. those not using it?
- In a given amount of executed queries, how many correct results were presented to the user?
- In what way can the conversation pattern (and algorithms used) be improved to pass a Turing test?
- In what way can the technology (i.e. the plug-in) be improved?
- When retrieving question-answer(s) from StackOverflow, some questions may be closed due to it being a duplicate. Can this data be used in any way to see what is defined by the StackOverflow community as a "good" question?

1.6 Planned contributions

To create an Artificial Intelligent Chat Agent for Question-Answering (Q/A) on programming by using StackOverflow as its knowledge base. This chat agent will function as a plug-in in the LMS system Open Edx. The external resources are content accessible on StackOverflow, libraries for content retrieval from StackOverflow, lexical word mappings (e.g. WordNet[6]), word filtering (to be decided) and available resources from Open Edx.

The prototype will be developed in the Advanced Project Course and tested during the Master Thesis. The goal is to see if the Chat Agent can aid students that are learning programming, but also do research on this it (e.g. trying to "humanize" the Chat Agent (Turing Test[7])). An example of the experiment would be to have two test groups (A/B testing), where the comparison would be on the grades of the students using the Chat Agent vs. those not using it, to see the use had any effect on the students grade.

Since this is such a narrow and specific field, the Chat Agent will be based on the Artificial Intelligence (AI) algorithms Hidden Markov Model (HMM) and Bayesian network (Bayes Net). These are chosen because they have been used for a very long time and there is a great deal of research out there on using these for linguistics and Chat Agents. It could of course be interesting to look at newer or different AI algorithms, but the issue is that these may not have an adequate amount of research and testing in relation to linguistics. It is therefore in my opinion safer to go with HMM and Bayes Net, to ensure the thesis is finished.

[6] WordNet: `https://wordnet.princeton.edu/`

[7] For the record, it should be noted that most ChatBots fail the Turing test, and this is extremely hard to achieve. Therefore, being able to pass the test might not be possible, but I want at least the Chat Agent to be humanoid enough so that the users feel comfortable conversing with it.

2 Related work

2.1 Chat Agent vs. Search Engine

2.2 Chat Agents for Learning and Education

2.3 Turing Test

2.4 Question-Answering (Q/A): What defines a good question?

3 Choice of methods

3.1 Survey and Interview

3.2 A/B testing as an experimental method

3.3 Question-Answering (Q/A) model

3.4 Quantitative comparison of the students results

4 Milestones, deliverables and resources

4.1 Table of Contents: Master thesis

[1] Although it is called "Topic covered" in this report, it may be more appropriate to call it "Research area" in the Master thesis.

- 3.5. Research Design

4. A/B Testing, Surveys and Results
   - 4.1. A/B Testing
   - 4.2. Interaction with the Chat Agent
   - 4.3. Statistical comparison of the students results
5. Discussions
   - 5.1. Data and Testing
   - 5.2. Artificial Intelligence (AI) Methods
   - 5.3. Implementation Architecture
   - 5.4. Chat Agent vs. Search Engines
6. Conclusion/Summary[2]
   - 6.1. Overview of main results
   - 6.2. Further work

Bibliography

- A. Data sets/Statistical Overview
- B. User Survey
- C. Interview Questionnaire Format
- D. Application Screenshots (Interaction with the Chat Agent)
- E. Miscellaneous information

4.2 Desired/necessary knowledge

The following list is a short summary of the knowledge needed to ensure completion of the master thesis:

- Understand how the Open Edx system works
- Understand how the StackExchange API (and community) works
- Be able to create an functional hybrid algorithm based on HMM and Bayes Net that can:
  - Convert user input to a query format
  - Be able to communicate with external resources (e.g. WordNET, StackExchange)
  - Be able to use this query format to retrieve data from the StackExchange community site
  - Process the retrieved data to only select the results that are relevant for the user

[2] Whichever is the appropriate format for the Applied Computer Science Master thesis.

  -

-

-

-

This part will be updated after user testing:
To get an early start, a prototype was developed in the course IMT5251 Advanced Project Work. To also have the ability to get user feedback and check if the current survey is good enough, students from the 1. and 2. year learning programming at Gjøvik University College (GUC) were invited to test the prototype. . .

4.3 Produced deliverables

What deliverables are to be produced (MSc thesis report, software,...)

For each of the activities identified, specify
1. the time you need to complete each activity both calendar time and 'man-hours'.
2. hours needed by you
3. things you need to buy (consumables)
4. equipment, lab space or facilities you need access to
5. contributions from others (e.g. survey/interview participants) and how much each will have to contribute in terms of resources (probably time)

4.4 Availability of deliverables

When are the various deliverables going to be available?

5 Feasibility study

Why can this project be completed in time?

E.g. project in Adv. proj. work, comparison to other projects, attempt to answer research question, etc.
How to solve the issues/problems in the project

6 Risk analysis

What can possibly go wrong when you do your project?
How do you intend to reduce impact of/solve these problems?

7 Ethical and legal considerations

The purpose of this chapter is to convince the reader and your self that your project activities are
- legal
- ethical, e.g. don't use/distribute/collect etc. data in such a way that individuals may suffer.

- A/B Testing (only half a class have access to the tool, giving them chance to learn more)
- No guarantee that students will learn (may just be confusing and time-consuming)
-

A Acronyms and abbreviations

A.1 Acronyms

AI Artificial Intelligence. 1

Bayes Net Bayesian network. 3, 10

GUC Gjøvik University College. 11

HMM Hidden Markov Model. 2, 3, 10

LMS Learning Management System. i, 1

Q/A Question-Answering. 3

B IMT5251 Advanced Project Course: Survey

C IMT5251 Advanced Project Course: Interview questionnaire format

Bibliography

[1] Russell, S. J. & Norvig, P. 2013. *Artificial Intelligence: A Modern Approach*. Pearson Education, 3 edition.

[2] Burke, R. D., Hammond, K. J., Kulyukin, V., Lytinen, S. L., Tomuro, N., & Schoenberg, S. 1997. Question answering from frequently asked question files: Experiences with the faq finder system. *AI magazine*, 18(2), 57.

[3] Crutzen, R., Peters, G.-J. Y., Portugal, S. D., Fisser, E. M., & Grolleman, J. J. 2011. An artificially intelligent chat agent that answers adolescents' questions related to sex, drugs, and alcohol: an exploratory study. *Journal of Adolescent Health*, 48(5), 514–519.

[4] Kerly, A., Ellis, R., & Bull, S. 2008. Calmsystem: a conversational agent for learner modelling. *Knowledge-Based Systems*, 21(3), 238–246.

[5] Knill, O., Carlsson, J., Chi, A., & Lezama, M. 2004. An artificial intelligence experiment in college math education. *Preprint available at* `http://www.math.harvard.edu/~knill/preprints/sofia.pdf`. Last Accessed: 25.09.2015.

[6] Kowalski, S., Pavlovska, K., & Goldstein, M. 2013. Two case studies in using chatbots for security training. In *Information Assurance and Security Education and Training*, Dodge, RonaldC., J. & Futcher, L., eds, volume 406 of *IFIP Advances in Information and Communication Technology*, 265–272. Springer Berlin Heidelberg.

[7] Jia, J. 2009. Csiec: A computer assisted english learning chatbot based on textual knowledge and reasoning. *Knowledge-Based Systems*, 22(4), 249–255.

[8] Jia, J. & Ruan, M. 2008. Use chatbot csiec to facilitate the individual learning in english instruction: A case study. In *Intelligent Tutoring Systems*, Woolf, B., Aïmeur, E., Nkambou, R., & Lajoie, S., eds, volume 5091 of *Lecture Notes in Computer Science*, 706–708. Springer Berlin Heidelberg.