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 appendix for the questionnaire used in IMT5251 Advanced Project Course (see p. 22). Updated Section 1.6: Planned contributions. Added content to the following Chapters: 4 and 7.
0.3	Added content to the following Chapters: 2.1, 2.2, 5 and 6. Changed the title of Section 2.1 and 2.4. Added Section 2.3. Updated Section 1.1 (removed some redundant sentences). Updated Section 1.6 and 4.2 (in regards to the developed prototype). Added Hidden Markov Model and Bayesian Networks to keywords. Changed the list of research questions from unordered to ordered list.
0.4	Added content to the following Sections: 3.1 and 3.2
0.5	Updated the following based on feedback from supervisor: Section 1.6: altered footnote in regards to Turing test. Section 2.4: Renamed to passing limited Turing test. Section 4.3.1 and 4.3.2: Updated subversions for May. Updated research question 6: Changed to limited Turing test. Added quick fix to print authors; added natbib and changed bibstyle temporary to unsrtnat (prints author-name).
0.6	Updated Section 2.2 based feedback from supervisor
0.7	Added content to the following Sections: 2.4 and 2.5 Started partially on the following Sections: 3.3, 3.4 and 3.5

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 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) themselves. 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 a limited 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, Hidden Markov Model, Bayesian Networks

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 Russell and Norvig [1, p. 2].

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 Burke et al. [2] and Bzz Crutzen et al. [3]), and not as a replacement for the teacher (e.g. CALMsystem Kerly et al. [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 Crutzen et al. [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. Knill et al. [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.

Kowalski et al. [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 Jia [7]) is a ChatBot developed for learning english, and was tested in Jia and Ruan [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]Crutzen et al. [3].

1.5 Research questions

1. How was the Chat Agent perceived by the user (e.g. using the Chat Agent vs. using a search engine)?
2. Did the Chat Agent help the user understand/learn more about programming?
3. By using A/B testing, is there a (statistical) improvement from the students using the Chat Agent, vs. those not using it?
4. In a given amount of executed queries, how many correct results were presented to the user?
5. In what way can the technology (i.e. the Chat Agent) be improved?
6. In what way can the conversation pattern (and algorithms used) be improved to pass a limited Turing test?
7. When retrieving question-answer(s) from e.g. 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 (if this is needed) and available resources from Open Edx.

The Master thesis will be an extension of a prototype that was developed in IMT5251 Advanced Project Course. This prototype just takes the users question and looks for matches on StackOverflow. The answer that is marked as correct (or the most top-voted one, if no answers are marked) will be returned and displayed to the user. Part of the contribution is to make this prototype more intelligent, by using AI and to do research on the implemented technology (e.g. trying to humanize the Chat Agent (limited Turing Test[7])). Part of the reason for wanting to humanize the Chat Agent is because the target group are students who are learning programming. It would be interesting to see if having this tool available can help them learn and understand more about programming. An attempt to confirm this will be by having an experiment with 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.

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

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

[7] For this thesis it would be a limited, domain specific Turing Test. Passing the true Turing test is out of the scope for this thesis, but it would be satisfactory if the users felt like they were talking to another human.

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 will be completed.

The final part of the planned contribution is to research and analyse the definition of a good question. The StackExchange community has a wide amount of sites, with a high level of professional expertise and strict rules for posting questions. Here, questions can be closed or put on hold, based on the questions asked. Some examples are duplicates, questions that are to broad, or marked off-topic Stackoverflow.com [9], CommunityWiki [10]. By analysing the posted questions and those asked by the students, could it be possible that the Chat Agent can help students ask better questions? An example scenario would be a programming lecture teaching incrementation. For a student wanting to learn more about this (presuming the student have no previous knowledge on the topic), a natural question would perhaps be to ask Q1) "What is incrementation?" or Q2) "How to increment?". If you input these questions into StackOverflow, you get approximately 31.600 results for both[8]. Furthermore, the first results may not even be relevant. The natural progression would perhaps be then to add to the question "... in programming". This reduces the returned results to 3.083 for Q1, and 2.715 for Q2. This example can be extended even further, e.g. by adding the programming language (e.g.: C++), the results are now halved, with an amount of 1.429 for Q1 and 1.610 for Q2. This simple example proves that knowing how to phrase a good question can have a real impact of the results you get, and as stated in Lucky [11], coming up with a problem can be the hardest part.

[8]This search was executed 29.11.2015, on StackOverflow.com.

2 Related work

2.1 Comparison of Information Retrieval (IR) when using Chat Agents and Search engines

I could not find anything in relation to this topic on the following sites:

- `http://dblp.uni-trier.de/`
- `http://link.springer.com/`
- `http://ieeexplore.ieee.org/`
- `http://www.sciencedirect.com/`
- `http://dl.acm.org/`

The following is a list of the keyword searches that were made:

- 'chatbot vs search engine'
- 'chatbot and search engine'
- 'comparison of chatbot and search engine'
- 'evaluation of retrieval systems'

The reason for this might be related to the fact that search engines are able to retrieve all sorts of information from numerous documents and web-sites, whereas ChatBots are usually made for light conversation or for specific topics and purposes. The research question this is related to (Research question 1) is more on the qualitative side, ie. will the users continue to use the search engine, or would they switch to the Chat Agent?

Although it is not an evaluation, in Crutzen et al. [3] a comparison of the ChatBot Bzz (for Windows Live Messenger) is made against search engines and information lines. The goal is to see which is better at answering adolescents' questions related to sex, drugs, and alcohol. The comparison was done by giving the users a questionnaire with a 5-point Likert scale. The results showed that the users found the ChatBot to be faster and more anonymous, in addition to being easier to use. Information quantity was considered less then both information lines and search engines, and it performed better when it came to conciseness and information quality Crutzen et al. [3, p. 517-518].

If one compares this paper to the goal of the Chat Agent I plan to develop, one can see some similarities. Even though search engines can give you numerous results, the quality of the results may vary, and there may not be any correlation between what you are looking for and what you find. Whereas with my Chat Agent, the focus is only on the StackExchange community, specifically on programming and StackOverflow. Therefore, one could also argue that rather then comparing the Chat Agent against a search engine, perhaps it rather should be compared against StackExchange. E.g. comparing the results

based on the question asked in the Chat Agent vs. the question searched for on the given StackExchange site.

There is also a program called FAQ FINDER, which is documented in Burke et al. [2]. As with my Chat Agent, here users can phrase their questions as they would when asking another person, rather then use keywords (as they perhaps would had to when using a search engine[1]).

2.2 Chat Agents for Learning and Education

There are numerous scientific articles and reports on using ChatBots in education, which are mentioned in many studies [3, 4, 5, 6, 7, 12, 13, 14, 15, 16]. Even though the names and definitions varies e.g. ChatBot, Virtual Teacher (VT), Intelligent Agents (IA) and Intelligent Tutoring System (ITS), the main purpose is mostly related to either relieving the teacher of work or to aid the user/students to learn more and acquire new knowledge. In the papers by [4, 5, 14] they found that students also wanted the ability to do smalltalk and have off-topic conversations. This would be a useful thing to implement, since this can increase the chance that the students will use the Chat Agent, since it will not be restricted to just the curriculum (e.g. being able to ask about the weather or just random conversations). There can however also be issues with having too free conversations, since users can attempt to use offensive language, invalid input causing the application to hang, spelling/grammatical errors or abuse in some way way (Kerly et al. [4]). Issues can also come if the knowledge base used is outdated, or is based on resources where there is no proper control of who is adding the information (Knill et al. [5], Imran and Kowalski [13], Reed and Meiselwitz [15]).

All information available in the StackExchange community is based on knowledge from the users who posts their answers there. This means that answers can be both outdated and invalid. However, StackExchange consists mostly of professional sites, where both moderators and the members are actively following the all posts, be it questions, answers or comments. Answers can also be graded by giving votes, and the answer that solved the users problem can be marked as correct. This data can then be used to ensure that the solution the Chat Agent presents to the user is based on useful knowledge (e.g. by filtering out answers with votes below a set threshold). An additional filtering can also be added by looking at the users reputation and badges [17, 18, 19]. Badges are awarded based on your contribution to the community, whereas reputation represents how much the community trusts you.

The goal is not for the Chat Agent to function as a VT, but more of an aiding tool to help students with the more general problems and help them be better at phrasing their questions. Not only that, but it can also help the teachers to understand how students learn by looking at the questions they ask the Chat Agent (Knill et al. [5], Rossi et al. [16]). The papers does not list a direct scientific proof that there is a learning improvement by using ChatBots. However, this does not mean that the use of ChatBots cannot have a positive impact. As noted in Kowalski et al. [6], the quantitative analysis showed no difference between those using and not using a ChatBot, but qualitatively they found that the use of a ChatBot was well received, and were open for using it again in the future.

[1] It should also be noted that I am biased towards it being better to phrase questions to find answers, rather then having to enter a list of keywords.

2.3 What is the quality of the results when using Hidden Markov Model (HMM) and Bayesian network (Bayes Net)?

TODO: Write this section

2.4 Passing a limited Turing test

The Turing Test (or Imitation Game) is based on the paper by Turing [20]. In this paper, Turing discusses whether or not a machine can be defined as intelligent, and to what ends the intelligence can be measured as. The original Imitation game was based on a man, a woman and a judge, where the goal was for the judge to guess which gender belonged to which participant. Turings suggestion was to alter this test to then include a human, a machine and a judge, where the judge would decide whether or not he was talking to a human or machine. However, Harnad [21] argues that the Turing Test is outdated and that a machine easily can trick another human into passing the test. Harnad therefore defines five levels[2] for the Turing test, where the level starts with t1 ("toy" functionality) and goes up to T5 (Grand Unified Theory of Everything). Based on Harnads paper, for a ChatBot it would be sufficient to pass T2.

Today, the Loebner Contest has replaced the Turing test (Shieber [22], Zdenek [23]). In the Loebner Contest, the contestants are graded on a numeric scale, where those that gets the highest score are perceived as most human-like. The winner was the one with the highest average score. The question however is if these types of tests truly can judge intelligence, since most of the time, it is all about the illusion of intelligence (Livingstone [24], Shieber [22]). Furthermore, often when attempting to reply by using the users input as base, it can often produce weird sentences (Shieber [22, p. 6]). To account for the lack of intelligence in the beginning, the judges had a script they had to stick to, but today the conversations can easily go out of proportions (Zdenek [23, p. 13]).

In summary, to pass a limited Turing test (e.g. Loebner Contest) it is all about whether or not you can fool the judge. Not about the intelligence. However, intelligence is not a key element in the Chat Agent, as most of the "intelligence" will be directed towards giving the students answers that are relevant and meaningful. Considering the goal is only to present meaningful answers, the students may be more forgiving when it comes to the probability of being presented with weirdly constructed sentences.

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

Question can be defined in many ways (e.g. a subject-related, situational, research/thesis, etc) (Boyer et al. [25, p. 3]). The main focus is on the academical questions, e.g. valid questions when writing the problem statement for the Bachelor or Master thesis. Basically questions of a level to be expected when pursuing an academical degree. StackOverflow alone has a lot of pages of how questions should and should not be phrased [9, 10, 26, 27, 28]. Lezina and Kuznetsov [29] attempted to predict closed questions on StackOverflow by analysing a database dump[3], but they noted that it would be too time consuming to analyse everything. Slowiaczek et al. [30] researches hypothesis testing, and what defines a good question and answer when you are restricted to only yes and

[2]The levels are t1: "Toy" functionality, T2: Pen-pal function, T3: Sensors and motoric (e.g. robot), T4: Humanoid in both looks and appearance and T5: Grand Unified Theory of Everything (Harnad [21]).

[3]You can also download all data available on StackExchange through the BitTorrent link found here: `https://archive.org/details/stackexchange` (last accessed 17th December 2015).

no answers.

Ragonis and Shilo [31] analysed problem-solving question which were sorted into categories and keywords. This can be used to see if it will be possible to create a taxonomy for questions. Boyer et al. [25] analysed how to encourage problem-solving in students, where they should starting with thinking about whether or not they understand the problem, before they attempt to find a solution. They also introduce new ways for instructors to ask questions to help students understand whether or not they understand the current task they are given. This can be useful when analysing the questions students asked compared to questions teachers ask. If the student uses an existing question from edX (asked by the teacher), presuming that a valid answer is not given, in what way will the student then re-phrase their question in an attempt to find the answer?

By having the Chat Agent as an alternative to the teachers, the students can also improve their own question quality. Students may feel that the question they ask is wrong, too stupid, or fear that may be ridiculed when asking it. Through the anonymity of the Chat Agent, they can ask the question in whatever form they want. The answer they get will be based on the question they ask, so in time they may improve as they learn what type of question format gives them the answer they seek.

3 Choice of methods

3.1 Hyptotheses and variables

In this section, I will attempt to identify the hypotheses and variables relevant for my Master thesis. The following Hypotheses are based on the research questions, and is an attempt to identify the possible outcomes of my research.

- H0: The Chat Agent will have no effect on neither the students knowledge, or the students ability to phrase good questions.
- H1: The Chat Agent will have no effect on the students knowledge, but the students will be better at phrasing good questions.
- H2: The Chat Agent will have an effect on the students knowledge, but not on the students ability to phrase good questions.
- H3: The Chat Agent will improve both the knowledge and the students ability to to phrase good questions.

The threats to the causality in my thesis is mostly the students maturity. In the beginning they may have little to no knowledge, but as they are coming closer to the end of the Spring semester, they will have acquired more understanding and knowledge on the given subjects. Their improvement in asking questions can also be affected by their supervisors (e.g. through the iterative process of asking questions, they learn to be more specific when asking for help). Previous knowledge from before the Bachelor started can also have an impact. Students with background in programming, be it self-taught or through work may know the basics, but its first at the end of the course they can put all the pieces together (the third variable problem).

3.2 Survey and Interview

To identify such underlying issues, everyone in the classes that participate will be given surveys to fill out during the thesis. The goal of this survey is to try to find out if there is a correlation between the use of the Chat Agent and the knowledge acquired at the end of the semester. The survey will ask them questions related to their current knowledge level and what experience, if any, they have from before. It is also necessary to find out in what way the different students learn, e.g. by using Fleming's VARK questionnaire[1] (used in Kowalski et al. [6, p. 152] and Sarabdeen [32]). This way, I can also try to see if there is a correlation between those that are of the type read/write and their grades at the end.

It would also be necessary to conduct interviews with participants, to see if there are any issues or variables that the survey has not picked up.
TODO: write more here...

[1] `http://vark-learn.com/the-vark-questionnaire/?p=questionnaire`

3.3 A/B testing as an experimental method

The students will be dived randomly into two groups; an experimental and a control group. The control group will follow the course the same way the last years students did. The same goes for the experimental group, but in addition they will have access to the Chat Agent. This means that the experiment will be a semi-quasi experiment (Leedy and Ormrod [33, p. 226-248] and Ringdal [34, p. 114-115]).

3.4 Question-Answering (Q/A) model

The students ability to ask better questions can be analysed by looking at the questions they ask the Chat Agent. Learning to ask better questions is an iterative process, and by comparing the questions asked in the beginning and the end, it might be possible to see if there is a qualitative improvement in their questions. This can also be done quantitatively, by comparing the amount of questions asked before marking the received answer as correct.

Since answers will be of various length, long answers will be shortened with a "Read More?". When the user then clicks the "Read More?", the value that this answer was read will be stored in the database. This can then be used as a measurement to see out of all answers retrieved for a given question, how many were of interest to the user (in addition to looking at the answer marked as correct by the Chat Agent user). This can also be used to see how precise and accurate the Chat Agents AI is.

3.5 Quantitative comparison of the students results

How can we see if there is any notable difference in the students results? This can be done by comparing the results against both the students not participating, but also against the students from the previous year. This could be done by using either Analysis of Variation (ANOVA) or Dunnett's test[2]. Dunnett's test was used in Simon and Snowdon [35] to compare the results of the current students and those from the previous year.

[2] Dunnett's test is effective when the sample sizes are small, the separate populations are not normally distributed, and their variances are not equal (as determined by separate tests) (Simon and Snowdon [35, p. 3]).

4 Milestones, deliverables and resources

4.1 Table of Contents: Master thesis

Front page

Abstract

Acknowledgements

Table of contents

Glossary

Acronyms

1. Introduction
   - 1.1. Keywords
   - 1.2. Topic covered/Research area[1]
   - 1.3. Problem description
   - 1.4. Research questions
   - 1.5. Methodology to be used
   - 1.6. Justification, Motivation and Benefits
   - 1.7. Limitations
   - 1.8. Thesis contribution
   - 1.9. Thesis structure
2. State of the art
   - 2.1. Chat Agents vs. Search Engines
   - 2.2. Artificial Intelligence (AI) for Chat Agents
   - 2.3. Chat Agents for Learning and Education
   - 2.4. Question-Answering (Q/A): What defines a good question?
   - 2.5. Turing test: Humanizing the AI
3. Methodology
   - 3.1. Hidden Markov Model (HMM)
   - 3.2. Bayesian network (Bayes Net)
   - 3.3. A/B Testing
   - 3.4. Survey and Interview

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

4.2 Obtaining the desired knowledge

The most important key element in this thesis will be the development of the Chat Agent, since it is the focus of my thesis. Preliminary work has already been done in the course IMT5251 Advanced Project Work, where a prototype has been developed. The prototype runs in Open Edx as an XBlock. XBlock runs as an Fragment in Open Edx, allowing developers to add their own content which then can be re-used in multiple systems[3]. As previously mentioned, the prototype developed is a simplistic version, meaning there is no AI. The prototype simply takes the users input (the question) and copy/pastes it to search for matching questions on StackOverflow. It then returns the first result, where the answer the user sees is either the answer marked as correct, or the top-voted answer (if no answer is marked as correct by original poster). Students from the 1. and 2. year (who are learning to program) were invited to test the prototype, where I observed them. Afterwards they were asked to fill out a questionnaire (the questionnaire is shown in Appendix A.2).

This means that most of the development needed during the master thesis will be to

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

[3] For more on XBlock, see `http://edx.readthedocs.org/projects/xblock-tutorial/en/latest/overview/index.html`

implement the AI and for it to be able to use WordNET for semantics. This would probably take a whole calendar month, but it depends on the actual hours invested. If I work between 6-8 hours each day (40+ hours each week), then this should be at least operational at latest mid-February. The reason for this extended time is to ensure I have time to setup and test the AI properly before allowing students to test them to ensure the collected data is valid. When it comes to equipment, I already have most of what I need, since I am already working on the prototype (development is done in Arch Linux). I also have a USB stick with Arch Linux installed, so that I can work on my laptop in case something should happen to my Desktop. I have also acquired a student license for PyCharm Professional[4] which is valid for 1 year (until 25. November 2016). I am also aware of people that have the required knowledge who I can ask for help, such as my supervisor Simon McCallum, Sule Yildirim-Yayilgan, Mariusz Nowostawski and Rune Hjelsvold.

The user testing will of course require students to be willing to partake and test the Chat Agent. The plan is to use A/B testing, so that only part of the students have access to and can use this Chat Agent. One thing that may affect the end results are the knowledge level of the users, which means that even if there is an increase in the end results, it may not be because of the Chat Agent. The solution to account for this issue is by having the users also grade their own knowledge level, from being novice to expert (e.g. having programmed for years). Although there is not a set time limit for required use, if the Chat Agent is not in use, there will not be enough test data. However, the Chat Agent will be available from around February to May, and participation may not be that high if there is a weekly requirement for usage. A solution could be to require the participants to at least use the Chat Agent for at least 10-20 hours each calendar month. If there then are 20-30 participants, that means the Chat Agent will have a usage of 200-600 hours each month. Which in turn should provide a good amount of test data. There will be at least three surveys for the participants, the first when they start using the Chat Agent, the second when testing is halfway and the last at the end, to see if there is a correlation between usage, their results and the students own observations.

[4] PyCharm: `https://www.jetbrains.com/pycharm/`.

4.3 Produced deliverables

4.3.1 Hours needed by me

Product	Time (calendar)	Time ('man-hours')[5]	Version #	Notes
MSc thesis report	January - May	125-200 Hours	v0.1 - v0.5	Draft should be presented to supervisor monthly
1) Chat Agent (AI research)	January - February	50-75 Hours	v0.1 - v0.2	Should be available at latest Mid-February for course start-up
2) Chat Agent (AI development)	January - February	50-75 Hours	v0.1 - v0.2	Should be available at latest Mid-February for course start-up
Meeting w/ supervisor	January - May	10-20 Hours	v0.1 - v0.5	Once a week, estimated between 30-60 minutes
Analyse questions on StackExchange	February - May	50-100 Hours	v0.2 - v0.5	One way of doing this would be to use the Chat Agent to e.g. retrieve questions that are closed
Read various scientific papers	January - May	50-100 Hours	v0.1 - v0.5	Read up on the latest published papers to keep up-to-date
Process student surveys	Jan/Feb, March/April and April/May	10-30 Hours	v0.2 - v0.4	Surveys delivered by students participating in the A/B Testing
Process Student/ Chat Agent interactions	February - May	60-80 Hours	v0.2	Should at least try to keep a steady update on the current data on a weekly basis

4.3.2 Hours by others

Who?	Time (calendar)	Time ('man-hours')[6]	Version #	Notes
Supervisor	January-May	10-20 Hours	v0.1 - v0.5	Once a week, estimated between 30-60 minutes
Students	February-April/May	40-80 Hours	v0.2 - v0.5	Students interaction with the Chat Agent (time per person)
Students	Jan/Feb, March/April and April/May	1.5 - 3 hours	v0.2 - v0.4	Filling out survey/ questionnaire (per student)

[5]Total time spent during set period.
[6]Total time spent during set period.

5 Feasibility study

As mentioned in 4.2 the prototype has already been developed and tested by two 1. year Bachelor students (in addition I got a lot of useful feedback). Through the feedback I can ensure that the next version is more useful for the students, and the main focus of the development will be on the AI. Aside from having a personal interest in AI, I have also had two courses in Machine Learning at Gjøvik University College (GUC) and I have also taken a course in Content-based Indexing and Retrieval. One of the many things learned in Machine Learning was the development and use of both single and hybrid algorithms to solve different tasks. E.g. in the course IMT4641 Computational Forensics I developed a program that analysed Android SQLite databases by using Fuzzy rules and Decision Tree. Meaningful insight in retrieving information and how e.g. search engines work were learned through the course Content-based Indexing and Retrieval. Furthermore, I already know what type of hybrid algorithm I will use for the Chat Agents AI; HMM and Bayes Net. There is also a lot of research available on using these for language processing (and since HMM uses states, it can also remember the conversation history).

When it comes to the research for questions (and what defines a good question), there is already a lot of data available through the StackExchange community. This means that I do not have to rely solely on one of the community pages for my research. E.g. if StackOverflow is down for maintenance, I can just switch my focus to one of the other sites in the community. I also intend to add an editable backend for the teachers in the Chat Agent, so that they can themselves decide which of the StackExchange sites the Chat Agent should use. This way, it will not be locked to only one course (or one site), which also means that I can suggests for other professors at GUC (and potentially NTNU) that they should participate in testing the Chat Agent. Furthermore, this can also increase the amount of students using the Chat Agent, meaning I can get more data to analyse and compare to see if there is any benefit from using the Chat Agent.

6 Risk analysis

There are two points that can have a major negative impact on my Master thesis. The first is delay of the AI development (if it is not ready within time), and the second is not having enough users to test the Chat Agent. The greatest problem when developing and training an AI is the time it takes to make it work properly, ensuring it can handle invalid data and that it does not use too much time processing and presenting the result to the user. To account for this, the goal should not be to have a 100% perfect AI. The most important part in the beginning would be that the AI works. This way, the students can get started by testing the updated prototype, and then give feedback if something is not working as promised. However, if the AI is too stupid or give to many useless results, this can cause the students to get a negative view and stop using the Chat Agent.

During the prototype testing in IMT5251, although the students liked the concept, they did not find the prototype to be useful or preferable when compared to a search engine. This is also one of the arguments for why AI should be implemented before releasing it to the students. To avoid the students getting a too negative bias when starting to use it, they need to be properly informed that it is still just a prototype and given proper information about what functionality is expected to work. This can be then seen in relation to ending up not having enough participants testing it. As noted in Section 5, a way to get a lot of students to participate is by trying to get as many professors as possible interested in trying it out their courses. In a worst-case scenario where to many should stop using it, there will still be those who have not tried it, so replacements may be available. It is therefore important to ensure on a monthly (perhaps even weekly) basis that the participants are both using the Chat Agent, and they are willing to continue using it.

Since AI is a very large field, it can easily become complex and one can also get distracted and lose focus (or get hung up in what one could call "eye candy" features). Therefore, I think the best would be in the beginning to have weekly meetings with my supervisor. This way, the supervisor will not only know that work is being done and progress has been made, but he can also ensure that the work I have done and are planning do to is what I should focus on. It will also be helpful, because if I get issues which I cannot solve (and supervisor is not available at the time), I know that within the next week I have a scheduled meeting and can then focus on something else in the meantime.

7 Ethical and legal considerations

There are no legal considerations in this project. All user data be anonymized, and the only data logged by the Chat Agent is the questions that the students ask, and the answers they are presented with. The greatest ethical concern in this thesis is the invested amount of hours testing the Chat Agent. The students will have projects, deadlines and exams to relate to, so requirement of 10-20 hours of usage each month may be too much. On the other hand, if the Chat Agent is not used at all, there may not be enough test data to prove whether or not it had any effect on the students learning curve. If one were to set a timed requirement, it could increase the chance of stress and performance anxiety, ending with students declining to participate or withdrawing before the testing is completed. It is of course important for this thesis that the Chat Agent is used, but it is also important that the users want to use it. Another concern in relation to time is that there is no guarantee that the Chat Agent have any beneficial effects on learning. The results may show that it is faster or easier to use search engines, meaning students have lost time they could have invested elsewhere.

For the validity of the end results, there is also the question of whether or not the Chat Agent actually had any effect. Students past knowledge can effect the average outcome, for better or worse. During the thesis, the testing will be conducted by using A/B testing, meaning only a selected group will have access to the Chat Agent. If a large part of the group have a lot of knowledge from before, the end results may be mostly False-Positives. Because even though the data show improvement, they improved because they build on previous knowledge, and not by use of the Chat Agent. To catch this, one of the questions for each test user will be to grade their own knowledge level to scale the end results. There is also the issue that students who might benefit from the Chat Agent, will not be able to use it. Novice students, with little to none programming knowledge may find it unfair that they are not the primary selected group. However, by not selecting by type (e.g. knowledge) one can avoid affecting the data and the end results.

A Appendix

A.1 Acronyms

AI Artificial Intelligence. 1, 3, 10, 12–15, 17

Bayes Net Bayesian network. 3, 4, 15

GUC Gjøvik University College. 15

HMM Hidden Markov Model. 2–4, 15

IA Intelligent Agents. 6

ITS Intelligent Tutoring System. 6

LMS Learning Management System. i, 1

Q/A Question-Answering. 3

VT Virtual Teacher. 6

A.2 IMT5251 Advanced Project Course: Questionnaire

Questionnaire/survey for ChatBot testing

Are you a (cross the one correct for you):

___ 1. year student | ____ Bachelor
___ 2. year student | ____ Master
___ 3. year student | ____ Other? If so, what? _____

On a scale from 1 - 5 (*1 being novice and 5 being knowledgeable*), how much do you know about programming? _____

On a scale from 1 - 5 (*1 being novice and 5 being skilled*), how good are you at programming?
_____

Do you have any previous experience with programming (*before bachelor*)? __ Yes __ No
If yes, how many months/years? _____
Where did you have this experience (*self-taught, school, work*)? _____

When looking for answers to something you do not know or understand, grade the following from 1 - 5 (*1 being used the most, 5 being used the least*):

Ask the teacher	___	
Ask a classmate/friend	___	
Look for answers in books/library	___	
Use a search engine	___	Which one(s)? _____
Use an online forum/community	___	Which one(s)? _____

When using a search engine, on a scale from 1 - 5, how satisfied are you with the results you get (*1 very satisfied, 5 not very*)? _____

_____

_____

_____

Why are you so satisfied/dissatisfied with the search engine(s) your are using?

_____

_____

_____

_____

ChatBot questions:

Having now tried the ChatBot prototype, the following are some questions about your experience.

What are your thoughts on having this tool available during class (*e.g. would it be helpful to have it*)?

_____

_____

If you were to compare the ChatBot against a search engine, which one performs better and why?

_____

_____

When asking the ChatBot Questions, did you get the answers you wanted (*e.g. did it return meaningful answers*)?

_____

_____

How many times did you have to re-phrase your question to get an answer? _____
(leave space for multiple questions)

- Do you think that this is acceptable? _____
- How many re-phrases do you find acceptable? _____

What did you like when using this ChatBot?

_____

_____

What did you not like when using this ChatBot?

_____

_____

In what ways can this be improved? Ie. what can be done to increase the chance that you will use it?

_____

_____

Bibliography

[1] Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson Education, 3 edition, 2013. ISBN 978-1292024202.

[2] Robin D Burke, Kristian J Hammond, Vladimir Kulyukin, Steven L Lytinen, Noriko Tomuro, and Scott Schoenberg. Question answering from frequently asked question files: Experiences with the faq finder system. *AI magazine*, 18(2):57, 1997.

[3] Rik Crutzen, Gjalt-Jorn Y Peters, Sarah Dias Portugal, Erwin M Fisser, and Jorne J Grolleman. 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, 2011.

[4] Alice Kerly, Richard Ellis, and Susan Bull. Calmsystem: a conversational agent for learner modelling. *Knowledge-Based Systems*, 21(3):238–246, 2008.

[5] Oliver Knill, Johnny Carlsson, Andrew Chi, and Mark Lezama. An artificial intelligence experiment in college math education. *Preprint available at* `http://www.math.harvard.edu/~knill/preprints/sofia.pdf`, 2004. Last Accessed: 16.12.2015.

[6] Stewart Kowalski, Katarina Pavlovska, and Mikael Goldstein. Two case studies in using chatbots for security training. In Jr. Dodge, Ronald C. and Lynn Futcher, editors, *Information Assurance and Security Education and Training*, volume 406 of *IFIP Advances in Information and Communication Technology*, pages 265–272. Springer Berlin Heidelberg, 2013. ISBN 978-3-642-39376-1. doi: 10.1007/978-3-642-39377-8_31.

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

[8] Jiyou Jia and Meixian Ruan. Use chatbot CSIec to facilitate the individual learning in english instruction: A case study. In BeverleyP. Woolf, Esma Aïmeur, Roger Nkambou, and Susanne Lajoie, editors, *Intelligent Tutoring Systems*, volume 5091 of *Lecture Notes in Computer Science*, pages 706–708. Springer Berlin Heidelberg, 2008. ISBN 978-3-540-69130-3. doi: 10.1007/978-3-540-69132-7_84.

[9] Stackoverflow.com. How to ask. `http://stackoverflow.com/questions/ask/advice?`, 2015. URL `http://stackoverflow.com/questions/ask/advice?`

[10] CommunityWiki. What is a "closed" or "on hold" question? `http://meta.stackexchange.com/questions/10582/what-is-a-closed-or-on-hold-question`, 2015. URL `http://meta.stackexchange.com/questions/10582/what-is-a-closed-or-on-hold-question`.

[11] Robert Lucky. When the problem is the problem. http://spectrum.ieee.org/at-work/innovation/when-the-problem-is-the-problem, 2011. URL http://spectrum.ieee.org/at-work/innovation/when-the-problem-is-the-problem.

[12] Iwan Gulenko. Chatbot for it security training: Using motivational interviewing to improve security behaviour. http://ceur-ws.org/Vol-1197/paper2.pdf. URL http://ceur-ws.org/Vol-1197/paper2.pdf. Last Accessed: 16.12.2015.

[13] Ali Shariq Imran and Stewart James Kowalski. Hip - a technology-rich and interactive multimedia pedagogical platform. In Panayiotis Zaphiris and Andri Ioannou, editors, *Learning and Collaboration Technologies. Designing and Developing Novel Learning Experiences*, volume 8523 of *Lecture Notes in Computer Science*, pages 151–160. Springer International Publishing, 2014. ISBN 978-3-319-07481-8. doi: 10.1007/978-3-319-07482-5_15.

[14] Alice Kerly, Phil Hall, and Susan Bull. Bringing chatbots into education: Towards natural language negotiation of open learner models. *Knowledge-Based Systems*, 20 (2):177–185, 2007.

[15] Kevin Reed and Gabriele Meiselwitz. Teacher agents: The current state, future trends, and many roles of intelligent agents in education. In A.Ant Ozok and Panayiotis Zaphiris, editors, *Online Communities and Social Computing*, volume 6778 of *Lecture Notes in Computer Science*, pages 69–78. Springer Berlin Heidelberg, 2011. ISBN 978-3-642-21795-1. doi: 10.1007/978-3-642-21796-8_8.

[16] Pier Giuseppe Rossi, Simone Carletti, and Maria Antonietta Impedovo. Learning environments supported by software agents and chat-bot. http://www.iiis.org/cds2011/cd2011imc/iceti_2011/paperspdf/eb230hk.pdf, 2011. URL http://www.iiis.org/cds2011/cd2011imc/iceti_2011/paperspdf/eb230hk.pdf. Last Accessed: 16.12.2015.

[17] Stackoverflow.com. What is reputation? how do i earn (and lose) it? http://stackoverflow.com/help/whats-reputation, 2015. URL http://stackoverflow.com/help/whats-reputation.

[18] Stackoverflow.com. Badges. http://stackoverflow.com/help/badges, 2015. URL http://stackoverflow.com/help/badges.

[19] CommunityWiki. How does "reputation" work? http://meta.stackexchange.com/questions/7237/how-does-reputation-work, 2015. URL http://meta.stackexchange.com/questions/7237/how-does-reputation-work.

[20] A. M. Turing. Computing machinery and intelligence. http://cogprints.org/499/, 1998. URL http://cogprints.org/499/.

[21] S. Harnad. Minds, machines and turing. *Journal of Logic, Language and Information*, 9(4):425–445, 2000. ISSN 0925-8531. doi: 10.1023/A:1008315308862.

[22] Stuart M Shieber. Lessons from a restricted turing test. *arXiv preprint cmp-lg/9404002*, 1994.

[23] Sean Zdenek. Passing loebner's turing test: A case of conflicting discourse functions1. *Minds and Machines*, 11(1):53–76, 2001. ISSN 0924-6495. doi: 10.1023/A:1011214808628.

[24] Daniel Livingstone. Turing's test and believable AI in games. *Comput. Entertain.*, 4 (1), January 2006. ISSN 1544-3574. doi: 10.1145/1111293.1111303.

[25] Kristy Elizabeth Boyer, William Lahti, Robert Phillips, Michael D. Wallis, Mladen A. Vouk, and James C. Lester. Principles of asking effective questions during student problem solving. In *Proceedings of the 41st ACM Technical Symposium on Computer Science Education*, SIGCSE '10, pages 460–464, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0006-3. doi: 10.1145/1734263.1734417.

[26] Stackoverflow.com. What does it mean if a question is "closed" or "on hold"? `http://stackoverflow.com/help/closed-questions`, 2015. URL `http://stackoverflow.com/help/closed-questions`.

[27] Stackoverflow.com. What topics can i ask about here? `http://stackoverflow.com/help/on-topic`, 2015. URL `http://stackoverflow.com/help/on-topic`.

[28] Stackoverflow.com. What types of questions should i avoid asking? `http://stackoverflow.com/help/dont-ask`, 2015. URL `http://stackoverflow.com/help/dont-ask`.

[29] C Galina E. Lezina and Artem M. Kuznetsov. Predict closed questions on stackoverflow. `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.394.5678`, 2013. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.394.5678`. Last Accessed: 25.09.2015.

[30] Louisa M Slowiaczek, Joshua Klayman, Steven J Sherman, and Richard B Skov. Information selection and use in hypothesis testing: What is a good question, and what is a good answer? *Memory & Cognition*, 20(4):392–405, 1992.

[31] Noa Ragonis and Gila Shilo. What is it we are asking: Interpreting problem-solving questions in computer science and linguistics. In *Proceeding of the 44th ACM Technical Symposium on Computer Science Education*, SIGCSE '13, pages 189–194, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1868-6. doi: 10.1145/2445196.2445253.

[32] Jawahitha Sarabdeen. Learning styles and training methods. 2013.

[33] Paul D Leedy and Jeanne Ellis Ormrod. *Practical research*. Pearson Education, 2012.

[34] Kristen Ringdal. *Enhet og mangfold: Samfunnsvitenskapelig forskning og kvantitativ metode*. Fagbokforlaget, 2 edition, 2007. ISBN 9788245005691.

[35] Simon and Susan Snowdon. Explaining program code: Giving students the answer helps - but only just. In *Proceedings of the Seventh International Workshop on Computing Education Research*, ICER '11, pages 93–100, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0829-8. doi: 10.1145/2016911.2016931.